Developing Data-Driven Methods for Movement Pattern Analysis Using Geographic Context

by

Avipsa Roy

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved May 2021 by the
Graduate Supervisory Committee:

Trisalyn Nelson, Chair
WenWen Li
Peter Kedron

ARIZONA STATE UNIVERSITY

August 2021

©2021 Avipsa Roy

All Rights Reserved

ABSTRACT

The role of movement data is essential to understanding how geographic context influences movement patterns in urban areas. Owing to the growth in ubiquitous data collection platforms like smartphones, fitness trackers, and health monitoring apps, researchers are now able to collect movement data at increasingly fine spatial and temporal resolution. Despite the surge in volumes of fine-grained movement data, there is a gap in the availability of quantitative and analytical tools to extract actionable insights from such big datasets and tease out the role of context in movement pattern analysis. As cities aim to be safer and healthier, policymakers are in need of methods to generate efficient strategies for urban planning utilizing high-frequency movement data to make targeted decisions for infrastructure investments without compromising the safety of its residents. The objective of this PhD dissertation is to develop quantitative methods that combine big spatial-temporal data from crowdsourced platforms with geographic context to analyze movement patterns over space and time. Knowledge about the role of context can help in assessing why changes in movement patterns occur and how those changes are affected by the immediate natural and built environment. In this dissertation I contribute to the rapidly expanding body of quantitative movement pattern analysis research by 1) developing a bias-correction framework for improving the representativeness of crowdsourced movement data by modeling bias with training data and geographical variables, 2) understanding spatial-temporal changes in movement patterns at different periods and how context influences those changes by generating hourly and monthly change maps in bicycle ridership patterns, and 3) quantifying the variation in accuracy and generalizability of transportation mode detection models using GPS (Global Positioning Systems) data upon

adding geographic context. Using statistical models, supervised classification algorithms and functional data analysis approaches I develop modeling frameworks that address each of the research objectives. The results are presented as street-level maps, and predictive models which are reproducible in nature. The methods developed in this dissertation can serve as analytical tools by policymakers to plan infrastructure changes and facilitate data collection efforts that represent movement patterns for all ages and abilities.

DEDICATION

To my late grandparents for their unconditional love, blessings, and constant encouragement in my academic endeavors during their lifetime.

ACKNOWLEDGMENTS

The work presented in this thesis could not have been completed without the help, guidance, advice, and encouragement of my colleagues, friends, and family. It is to these many people that I owe my deepest gratitude.

First and foremost, I would like to thank my supervisor Trisalyn Nelson for her tremendous support and the constant guidance in terms of sharing ideas, thinking about a broader perspective where to situate my research, laughs, and disappointments; all of which has greatly contributed to my PhD experience and made me a better individual. Her knowledge and insight into every detail of my work are inspiring and it has been an honor to work with her over these past four years. She has always been patient with my questions and pushed me harder each day to excel in my academic endeavors and helped me connect the dots between computer science and geography.

I would also like to give a general thanks to all my friends, colleagues, and staff in the School of Geographical Sciences and Urban Planning. My mentors and fellow graduates and postdocs in the Spatial Analysis Research Center, for their generous advice, help, and friendship. Each one of them has made significant contributions to my graduate experience at ASU and has made my Ph.D. experience enjoyable and exciting. I am thankful to the Graduate College through the SGSUP at Arizona State University, the City of Phoenix, Los Alamos National Laboratory and Oak Ridge National Laboratory, and the National Science Foundation for their generous financial support to carry out my graduate work. A generous thanks also to the data providers Strava Metro, Team INTERACT,

Maricopa Association of Governments (MAG), City of Phoenix, and the City of Tempe, for supporting the work carried out in this dissertation.

I also am grateful to my committee members, WenWen Li and Peter Kedron, whose support in terms of helping me frame nuanced statistical questions and contextual using my research has changed my outlook on quantitative research. To amazing mentors, Emily Casleton and Geoffrey Fairchild, at LANL, and Bandana Kar at ORNL for their enthusiasm, support, and guidance during my internships. To my co-authors, Trisalyn Nelson, Stewart Fotheringham, Meghan Winters, Karen Laberee, Daniel Fuller, Colin Forster, Vanessa Brum-Bastos, Jaime Fischer, Kevin Stanley, Pavan Turaga, Peter Kedron, Bandana Kar, Emily Casleton, Geoffrey Fairchild, Ting Chen, Cathy Wilson, and Kurt Solander thank you for all your hard work and timely responses in the generation of manuscripts and research discussion.

My heartfelt gratitude goes to my parents, Avijit Roy and Bidisha Ray, for being by my side while I pursued my graduate education in a foreign country. Without their encouragement and innumerous sacrifices, I would not be where I am today. I also share my love for my brother, Avishuman Ray, who aspires to be a physicist someday, and hope I can encourage him to stay focused and persevere to achieve his academic goals. Finally, I would like to mention my wonderful and loving husband, Tirtha Banerjee, for always being by my side across three different continents – your care, patience, love encouragement, and constructive criticism have carried me through the difficult periods of my graduate studies. Last but not least I owe my gratitude to Tirtha's parents, Tarasankar Banerjee and Rina Banerjee, whose love, support, and encouragement have always helped me shape my career the way I aspired.

TABLE OF CONTENTS

Page

CHAPTER Page

CHAPTER Page

LIST OF TABLES

Table Page

Table Page

LIST OF FIGURES

Figure Page

Figure Page

LIST OF ABBREVIATIONS

AADB – Average Annual Daily Bicyclist

AUC – Area Under the Curve

AIC – Akaike Information Criterion

ACS – American Community Survey

ADA – Ada Boost algorithm

API – Application Programming Interface

CART – Classification and Regression Trees

FDA – Functional Data Analysis

GPS – Global Positioning Systems

GIS – Geographic Information System

LASSO – Least Absolute Shrinkage and Selection Operator

MAG – Maricopa Association of Governments

PANDAS – Python Data Analysis Library

POI – Point of Interest

RF – Random Forests

SMOTE – Synthetic Minority Sampling Technique

SRVF – Square Root Velocity Function

SVM – Support Vector Machines

TBAG – Tempe Bicycle Advocacy Group

VIF – Variance Inflation Factor

WHO – World Health Organization

XGB – Extreme Gradient Boosting algorithm

CHAPTER 1

INTRODUCTION

1.1 Problem Statement

The recent technological advances in acquiring high-quality movement data from global positioning systems (GPS) and other satellite tracking technologies (radiotelemetry, fitness apps, health monitoring devices, smartphones, accelerometers, etc.) have opened up a new avenue for researchers to study movement processes using data-driven analysis of movement patterns. Movement is defined as a continuous process that is represented discretely in space by a sequence of an object's locations (in the form of (X, Y) coordinates) captured synchronously in time. The abundance of temporally dense movement data has outpaced the available methods for analyzing such data. Analytical approaches to movement processes have evolved, however, methods to develop a standardized framework integrating movement data with geographic covariates to define a specific behavior is limited. There is a need within Geographic Information Science to contextualize readily available 'big' movement data generated by ubiquitous platforms like smartphones and fitness trackers using geographic covariates that quantify the immediate surroundings of users. Such knowledge from a geographic context can facilitate a better understanding of the underlying factors that govern human mobility patterns and can go a long way in planning more resilient cities in the face of natural hazards like floods and hurricanes or pandemics like COVID-19 (Roy and Kar, 2020). To strategically develop generalized analytical methods of spatial-temporal patterns an overall understanding of the

underlying structure of movement data is essential along with an overview of the methods already in place.

1.2 Background Literature and Research Gaps

Movement can be defined as a process that operates in both space and time. It is a continuous process where an object follows a unique trajectory between an initial and a final point in space and time. Movement data are used to represent the continuous process of movement for geographical analysis. Current and existing geospatial data collection techniques represent movement data most commonly as a collection of point objects with time stored as an additional attribute. A more formal definition of movement data is the collection $\{M_t\}$ of t = 1, …, n ordered records each comprising the triple <ID, S, T>, where ID is a unique object identifier, S are spatial (x,y) coordinates, and T a sequential (non-duplicated) timestamp (Hornsby and Egenhofer 2002). Each of the points specifies a unique trajectory of an object's motion. Additional attributes such as distance, speed, and azimuth (or relative turning angle) that help generate a trajectory can be derived from the raw spatial locations. The temporal component is usually captured in the form of a time series represented as a set of locations ordered in time for each trajectory. In this dissertation, I have used data from different crowdsourced platforms including Strava Metro (Strava Metro,2017) which is an anonymized bicycle trip data collection and storage platform, and applications like MTL Trajet (MTL Trajet, 2017) and Itinerarium (Patterson et al., 2019), which collects mobility data from GPS enabled smartphones, to represent movement data.

While movement data have been collected using a variety of techniques like GPS (Laube et al, 2007; Laube and Dennis, 2006; Laube and Purves, 2011; Dennis et al, 2010), RFID (Bleisch et al, 2014), radio telemetry (Stewart et al, 2013; Laberee et al, 2014; Calenge et al., 2009) the more current acquisition schemes include crowdsourced fitness apps (Jestico et al., 2016), LiDAR (Kirkeby et al, 2016), accelerometers (Roy et al, 2020) and data from smartphones (Roy and Pebesma, 2017). Local authorities have traditionally offered manual and automated counts through sensor tracking technologies to capture pedestrian movement and bicycle ridership data for enhanced transportation planning. Although these methods of movement data collection were at the coarser spatial and temporal resolution, with the evolution of crowdsourced data from platforms like Strava Metro, it is now possible to capture finer granularities associated with frequent sampling intervals that provide a detailed representation of movement. More recently, GPS-enabled rideshare vehicles such as Lime Bike, JUMP, GriD, and OFO that capture active modes of transportation along with motorized rideshares such as Uber and Lyft offer a new addition to the movement data paradigm.

Despite being an emerging area of research there are some limitations in bicycling ridership studies in the United States. Official data on active modes of transportation are usually sparse. Active transportation data are typically collected by traditional methods such as manual counting or tubes. When traditional methods are used questions about the representativeness of the spatial distribution of count locations and the spatial sampling scheme may arise (Roy et al., 2019; Nelson et al., 2021). More recently sampling efforts have shifted to use ecocounters, which are automated counters that sample a number of bicyclists crossing any single intersection or street segment continuously in time, are being

installed but the placement of counters are not spatially uniform (Roy et al., 2019). The authorities typically place counters in high ridership areas which leaves low ridership areas under- or unrepresented in the sample and hence not stratified enough to capture different sets of population groups. This gap in our data measuring active transportation is limiting the use of such count data for modeling ridership patterns that is representative of the entire population. New sources of data are now emerging to fill this gap in the data. One such source is the data collected by crowdsourced fitness apps like Strava. Strava data are a huge source of high-resolution data for planning and policymaking.

Several studies have confirmed that crowdsourced data from fitness apps can be used for modeling movement patterns for example – mapping bicycling ridership patterns (Jestico et al., 2019; Garber et al., 2019) from GPS data, detecting changes in bicycling volumes based on infrastructure changes (Boss et al., 2018) or studying route choice of individual bicyclists from smartphone apps (Pritchard et al., 2019). However, further research is necessary to understand how bias in crowdsourced data can be addressed before making reliable policy decisions from such emerging data sources. Failure to address bias in the data can lead to biased policy outcomes that can lead to an 'overinvestment' in infrastructure in high ridership places. However, to evaluate if such overinvestments are needed we must properly estimate riders in other areas as well. Additionally, if local authorities invest in high ridership areas, they are more likely to also invest in wealthy, white neighborhoods than socially vulnerable (i.e., low income, non-white population, less access to cars) areas which may further exacerbate social inequalities.

Movement data from fitness apps like Strava once corrected for sampling bias, can play an important role in monitoring trends over time which is critical to understand travel

behavior and plan interventions. These datasets are temporally dense and could be used as a mappable timeseries for planning purposes. However, sampling inaccuracies and temporal variability introduced during data collection may lead to misalignment issues in such timeseries representations. If the misalignment in the data is not accounted for they might distort features and distance measures when calculating change or modeling similarity from functional representations of such 'big' data resulting in an incorrect visual representation of change. For example, street segments with a decrease in bicycle ridership volume annually may turn out to show no change at all leading planners to ignore invest in better bicycle infrastructure in that area. It is therefore essential to account for such data misalignment issues when dealing with temporal analysis.

Additionally, combining movement data from multiple sources for spatial-temporal analysis is challenging too as different data modalities have varying granularity. The pre-processing techniques for one modality might not be appropriate for another data source, due to a lack of a uniform standardized framework for data preprocessing. For example, in the case of transportation research, bias correction in crowdsourced data often requires matching GPS data in terms of spatial and temporal resolution with official counts data (Jestico et al., 2016). This could be overcome through a generalized framework that accounts for variations in data resolutions and models the inputs from multiple sources in a uniform representation. In terms of predicting travel modes from crowdsourced data, the planners and policymakers are also limited by the process of extracting meaningful features from such data that can deliver actionable insights for decision-making (Roy et al., 2020). Moreover, geographic context plays an important role in movement pattern analysis and

there is a research gap in methods that combine context with crowdsourced movement data to understand travel mode choice or changes in transportation patterns (Boss et al., 2018).

Quantifying spatial-temporal patterns at different scales is also an important component of movement analysis to be considered by researchers, as it facilitates the comparison of movement trajectories between individuals across space and time (e.g., Fryxell et al., 2008; Graham and Stenhouse, 2014). For instance, changes in movement patterns occurring at one space-time scale may be masked if represented at another scale (Fleming et al., 2014; Gurarie et al., 2009; Schick et al., 2008). Existing methods for movement analysis can be used in an exploratory capacity to quantify changes in movement patterns(Sur et al., 2014), or in an explanatory capacity to link pattern changes back to changes in underlying processes (Barraquand & Benhamou, 2008). However, increasing model complexity with the growing volume of data necessitates models that can tackle such big data with an increase in technical capabilities and computational power (Morales et al, 2004; Nams, 2014). Differing assumptions amongst movement models could result in different characterizations of the same data (Gurarie et al., 2015; Schick et al., 2008), resulting in a mismatch between model assumptions and movement processes which might generate erroneous results (Nams, 2014). Therefore, to effectively apply model-based approaches, a well-grounded knowledge of the geographic context which captures underlying processes and immediate environment influencing movement is essential (Barraquand and Benhamou, 2008).

Existing approaches for modeling big movement data requires a great amount of preprocessing that involve data fusion from different platforms, trajectory segmentation from large GPS traces but they are challenging owing to the noise introduced in such data

by GPS devices that overrides the natural variation in movement patterns (Nams, 2014) as well as differences in scales. Methods like frequent pattern mining (Han and Yin, 2000; Roy and Pebesma, 2017), similarity analysis from time-series databases (Agarwal et al, 1993; Michelot et al., 2016), and spatial similarity using applying Euclidean distance between trajectories (Yanagisawa et al, 2003) have been used for capturing similar movement patterns. Methods like the Edit Distance (Chen et al, 2005), One-Way Distance (Lin & Su, 2008), Hausdorff distance (Goodrich et al, 1999), Fourier descriptors (Rafiei and Mendelson, 2002), Longest Common Subsequence (LCSS) (Vlachos et al, 2002) have been used for grouping trajectories into clusters from timeseries. However, further investigation is needed to contextualize why similarity in movement patterns emerge.

In summary, although crowdsourced data can overcome some of the existing challenges in movement pattern analysis by providing more fine-grained data, there is a gap in the literature in terms of analytical methods that can demonstrate their appropriate usage along with geographic context and identify the limitations in movement pattern analysis. Identifying these gaps I have identified specific research objectives defined in section 1.3 to move the research in movement pattern analysis forward.

1.3 Research Objectives

With the growing volume of movement data, addressing transportation planning from a data and computation-intensive perspective has become inevitable. Unfortunately, there are large gaps in the data resolution, coverage, and quality for transportation at the street segment level. Traditional data sources like manual counts travel diaries and questionnaires - data have poor spatial detail and/or limited temporal coverage. Crowdsourcing has,

therefore, emerged as a tool of interest for collecting data at a finer resolution at a higher sampling frequency.

Although crowdsourcing has facilitated the mechanism of data collection by reducing cost and time, such data are a biased sample of the entire population. The bias in crowdsourced data has limited its use in the real world by practitioners. Decision-makers are also in need of better and more efficient tools to convert crowdsourced data into actionable insights by combining geographic context along with movement metrics like speed and acceleration. An overall mechanism of correcting bias in crowdsourced data as well as integrating mobility data along with geographic correlates in space and time can ease the process of decision-making for practitioners.

In this dissertation, I have identified three primary research questions and developed quantitative methods to understand patterns emerging from movement data by adding geographic context. I have used statistical and machine learning approaches to develop modeling frameworks that address these questions in the context of transportation planning in urban areas.

a. *How can we correct bias in crowdsourced GPS data to map ridership of all bicyclists in a city?*

Crowdsourced data, although dense in spatial and temporal resolution, are biased towards users with access to a mobile device and willing to record trips. It is a subsample of true counts observed on the ground and hence needs to be checked for representativeness before using it to make decisions related to policymaking. Thus, there is a need to quantify and correct the inherent bias in crowdsourced data (Lieske et al., 2017) for a better

representation of the ridership patterns of all riders, across varying ages and abilities. A mechanism for identifying additional geographic covariates to adjust for the bias in Strava ridership is desirable to facilitate mainstream usage of crowdsourced fitness app data for public health and urban planning. To address this gap I have developed a bias correction framework using a machine learning based variable selection operator. This framework aims to identify significant variables that could account for the Strava ridership bias and help predict annual ridership volumes at the street segment level. The goals for bias-correction in crowdsourced data are —first, to quantify which geographical variables can help in correcting bias using a Least Absolute Shrinkage and Selection Operator (LASSO); and second, to predict overall bicycle ridership volumes using the LASSO selected bias-adjustment factors and generate maps representative of all bicyclists at a street-level spatial resolution.

b. How can we detect changes in movement patterns from big spatial-temporal data?

Monitoring change is an important aspect of understanding variations in spatial-temporal processes. Recently, big data on mobility, which are detailed across space and time, have become increasingly available from crowdsourced platforms. New methods are needed to best utilize the high spatial and temporal resolution of such data for monitoring purposes. These data can be considered mappable time series, but are challenging to use owing to varying sampling rates and issues of temporal misalignment. However, there is a gap in understanding the efficient use of high-resolution crowdsourced data for change detection as automated approaches of change detection from big datasets are not commonplace. These methods can open a new avenue for urban planners for decision-

making ahead of time and stay prepared for urgent scenarios. I introduce a novel functional data analysis framework in Chapter 3 for quantifying temporal change in mobility patterns from crowdsourced GPS data. The framework utilizes crowdsourced data from Strava and automates change detection employing a functional k-means clustering technique that calculates distance matrices based on the Fisher-Rao metric after aligning the functional curves using the square root velocity function. Hourly and monthly changes are classified into four categories and mapped along with exposure density. Using spatially and temporally continuous data our study advances the existing approaches to mobility analysis, by capturing data about the underlying processes, rather than monitoring change between discrete snapshots of time. This method is reproducible by practitioners for monitoring changes from crowdsourced ridership data and for making necessary infrastructure changes to assure the safety of bicyclists. The possibility of utilizing such fine-grained data for detecting temporal changes can help planners resolve controversies over new infrastructure (Nelson et al., 2020) and identify long term trends in changes in human mobility patterns during natural hazards (Han et al., 2019) like hurricanes or pandemics like COVID-19 (Roy et al, 2020).

c. *How can we classify transportation modes from movement patterns combining geographic context? What are the key challenges and outcomes in terms of the generalizability of results?*

The increasing availability of health monitoring devices and smartphones has created an opportunity for researchers to access high-resolution (spatial and temporal) mobility data for understanding travel behavior in cities. Although information from GPS data has

been used in several studies to detect transportation modes, there is a research gap in understanding the role of geographic context in transportation mode detection. predictive models lack generalizability. Integrating the geography in which mobility occurs, provides context clues that may allow models predicting transportation modes to be more generalizable. In Chapter 4, I developed a data-driven framework for transportation mode detection using GPS mobility data along with geographic context and second, to assess how model accuracy and generalizability varies upon adding geographic context. The method will account for GPS data of varied resolution and be able to combine it with nearby points of interest thereby adding geographic context which can help detect travel modes. The method can assess the change in predictive accuracy and generalizability of multiple machine learning algorithms upon the addition of contextual variables in terms of built/natural environment, land use types, and availability of transportation infrastructure. With additional research using travel surveys and user inputs, the study can be used for planning strategic data collection efforts by identifying which geographic factors contribute towards specific travel mode choice and how the built environment influences travel mode choices.

1.4 Dissertation Overview

The following chapters elaborate further on each of the research questions including the methods applied to answer these questions along with their results, limitations, and future work necessary to expand the research. In the first two chapters, bicycling data from the Strava fitness app is used as a case study for movement pattern analysis at a fine spatial

and temporal resolution that is relevant to understanding the role of geographic context and its role in active transportation.

Chapter 1 begins by motivating the problem followed by identifying the main research goals of this dissertation followed by an extensive review of the existing literature available on traditional and emerging sources of movement data, current approaches to movement pattern analysis, and identifying the research gaps and limitations of the methodologies in place. The following chapters highlight three different quantitative approaches to understand movement patterns using crowdsourced data.

In Chapter 2, I introduce a generalized bias correction approach across all spatial and temporal scales that is desirable to facilitate mainstream usage of crowdsourced fitness app data from platforms, such as Strava, for public health and urban planning. The bias-corrected bicycle ridership is used to generate maps representative of all bicyclists at a street-level spatial resolution for the city of Tempe.

Chapter 3, further elaborates on the functional data analysis framework for quantifying change in mobility patterns from Strava data across hourly and monthly scales. The framework utilizes crowdsourced data from Strava and automates change detection employing a functional k-means clustering technique that calculates distance matrices based on the Fisher-Rao metric after aligning the functional curves using the square root velocity function. The change clusters are used to generate change maps for hourly and monthly bicycling ridership in the city of Phoenix.

Chapter 4 showcases a data-driven framework for transportation mode detection using GPS mobility data along with geographic context and second, to assess how model accuracy and generalizability varies upon adding geographic context. Finally, chapter 5

elucidates some of the key findings, major contributions, existing limitations to each study, and concluding remarks highlighting future work revolving around each research objective.

CHAPTER 2

CORRECTING BIAS IN CROWDSOURCED MOVEMENT DATA

2.1 Abstract

Traditional methods of counting bicyclists are resource-intensive and generate data with sparse spatial and temporal detail. Previous research suggests big data from crowdsourced fitness apps offer a new source of bicycling data with high spatial and temporal resolution. However, crowdsourced bicycling data are biased as they oversample recreational riders. Our goals are to quantify geographical variables, which can help in correcting bias in crowdsourced, data and to develop a generalized method to correct bias in big crowdsourced data on bicycle ridership in different settings in order to generate maps for cities representative of all bicyclists at a street-level spatial resolution. We used street-level ridership data for 2016 from a crowdsourced fitness app (Strava), geographical covariate data, and official counts from 44 locations across Maricopa County, Arizona, USA (training data); and 60 locations from the city of Tempe, within Maricopa (test data). First, we quantified the relationship between Strava and official ridership data volumes. Second, we used a multi-step approach with variable selection using LASSO followed by Poisson regression to integrate geographical covariates, Strava, and training data to correct bias. Finally, we predicted bias-corrected average annual daily bicyclist counts for Tempe and evaluated the model's accuracy using the test data. We found a correlation between the annual ridership data from Strava and official counts ($R^2 = 0.76$) in Maricopa County for 2016. The significant variables for correcting bias were: The proportion of white population, median household income, traffic speed, distance to residential areas, and

distance to green spaces. The model could correct bias in crowdsourced data from Strava in Tempe with 86% of road segments being predicted within a margin of ±100 average annual bicyclists. Our results indicate that it is possible to map ridership for cities at the street level by correcting bias in crowdsourced bicycle ridership data, with access to adequate data from official count programs and geographical covariates at a comparable spatial and temporal resolution.

2.2 Introduction

Lack of physical activity is identified as one of the primary factors leading to increased risk of chronic diseases, including obesity, cardiovascular diseases (Sallis et al., 2012), and type 2 diabetes (Colberg et al., 2010) as well as cancer (Kushi et al., 2012). The World Health Organization recommends a minimum of 150 min of moderate physical activity per week (WHO, 2010). Active transportation modes (bicycling and walking) help to incorporate routine physical activity among adults with a sedentary lifestyle to reduce health risks. Consequently, public health and urban planning agencies are increasingly recognizing the importance of active transportation (Mansfield et al., 2016) in their pursuit of broader public health goals (Lyons et al., 2012), creating a demand for a better understanding of the influences on bicycle ridership. Previous studies (Larsen et al., 2013; Lovelace et al., 2017) have used empirical methods to inform policymakers about necessary infrastructure changes using origin-destination surveys to help increase physical activity levels among adults.

Unfortunately, there are large gaps in the data resolution, coverage, and quality for active transportation at the street segment level. Existing approaches to bicycle counting

result in data with poor spatial detail and/or limited temporal coverage (Ryus et al., 2014). The three most common ways to collect bicycle ridership data are manual counts (Griswold et al., 2011), temporary, and continuous counters (Ryus et al., 2014). Manual counts, often conducted by volunteers, typically enumerate the number of cyclists at major street intersections during peak commuting periods for a few days of the year (Nordback et al., 2011), and lack dense spatial coverage and temporal detail (Griffin et al., 2018). Temporary counts (i.e., tube counters set out for a week or two) provide a snapshot of ridership at a location over time, but, typically, the spatial coverage is limited. Automated counters (counting bicyclists crossing a specific street intersection continuously) (El Esaway et al., 2015) have great temporal detail but often lack spatial coverage.

Crowdsourcing has, therefore, emerged as a tool of interest for collecting data on bicycling ridership (Shen & Stopher, 2014; Griffin & Jiao, 2015; Heesch & Langdon, 2016), comfort mapping for bicyclists (Bil et al., 2015), understanding the effects of the built environment on ridership (Winters et al., 2010), and promoting safety among riders (Nelson et al., 2015). The emergence of crowdsourced data generated by fitness apps (e.g.,Strava.com) has provided a new source of ridership data with enhanced spatial and temporal resolutions (Jestico et al., 2016). With the proliferation of smartphones, fitness apps, such as Strava, have emerged as one of the most popular and rich sources of data for physical activity tracking; Strava records an average of 2.5 million GPS routes weekly by users across 125 cities all over the world (Strava Metro, 2018).

However, the primary concern with crowdsourced data is the bias towards recreational riders, who are frequent users of GPS-enabled fitness apps. Thus, there is a need to quantify and correct the inherent bias in crowdsourced data (Lieske et al., 2017) for a better

representation of the ridership patterns of all riders, across varying ages and abilities. A generalized bias correction approach across all spatial and temporal scales is desirable to facilitate mainstream usage of crowdsourced fitness app data from platforms, such as Strava, for public health and urban planning. Most studies on bias in crowdsourced data (Feick et al., 2013) focus on characterizing the nature of the bias (Solymosi et al., 2017; Ton et al., 2018). We hypothesize that crowdsourced data in urban settings can be used to map bicycling ridership (Jestico et al., 2016; Sun & Mobasheri, 2017). Here, we move the research forward by developing a generalized approach to bias correction that combines traditionally collected ridership data with crowdsourced data to fill gaps in the spatial and temporal detail.

Our goal is twofold—first, to quantify which geographical variables can help in correcting bias in crowdsourced data; and second, to develop a generalized method to correct bias in big, crowdsourced data on bicycle ridership in different settings to generate maps representative of all bicyclists at a street-level spatial resolution. Maps were created with enhanced spatial and temporal detail given the 'big data' provided by crowdsourced fitness apps. Bias correction was framed as using crowdsourced fitness app user counts along with additional geographic covariates to predict average annual daily bicyclist (AADB) counts on a street network. The result is a map that shows the ridership of bicyclists of all ages and abilities, even those that do not use the app.

2.3 Study Area

Our study area was Maricopa County in the state of Arizona, USA, and covers 9200 square miles (Figure 2.1). Maricopa County includes 27 cities anchored by Phoenix (MAG,

2016). With a population of over 3.3 million people, it is the fourth most populous county in the USA (US Census Bureau, 2012). The weather is mostly arid with summer temperatures ranging from 50 °F (10 °C) to 108 °F (42 °C) and winter temperatures between 35 °F (1 °C) and 90 °F (26 °C), with an average precipitation of 132 mm in summer and 236 mm in winter. The city of Tempe, within Maricopa County, specifically has more than 175 miles of bikeways and the highest percentage of residents commuting by means of bicycles at 4.2%, far higher than the Maricopa County average of 0.8% (City of Tempe, 2015).

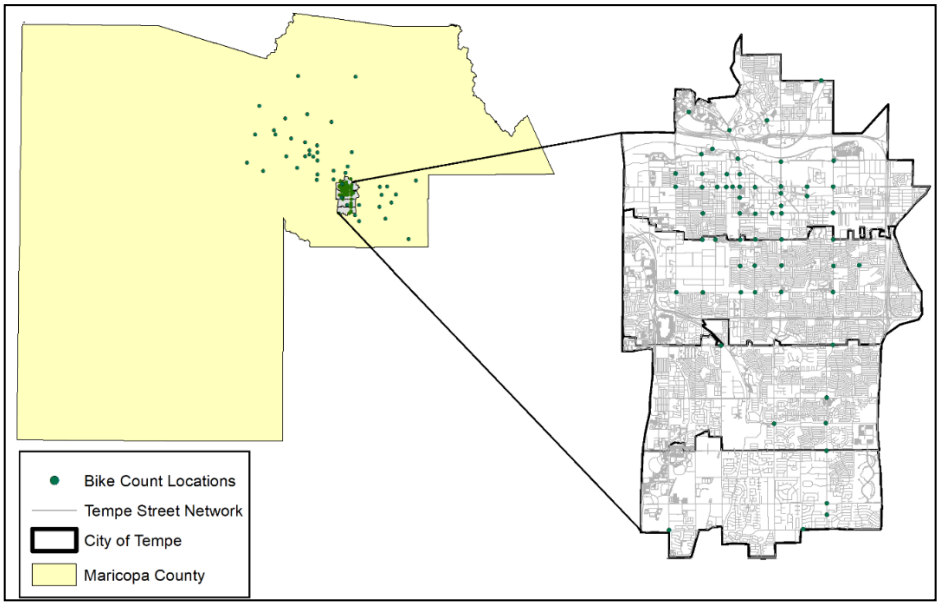


Figure 2.1: Map showing the geographic location of the study area within Maricopa County, AZ, USA along with the street network layout.

2.4 Data

Two official count data sets were used, the first to train the model and the second to test the model. To train the model, we used temporary, automated bicycle counts

completed by the Maricopa Association of Governments (MAG) at 44 locations in 2016 (Figure 2.2). We used the commonly reported time period, the annual average daily bicyclist (AADB) count, for the official counts as provided by the MAG. Bicyclists were counted by the MAG using automated counters with pneumatic tubes over a span of eight continuous two-week periods in the months of April, May, October, and November to understand and capture the variation in seasonal cycling volumes.

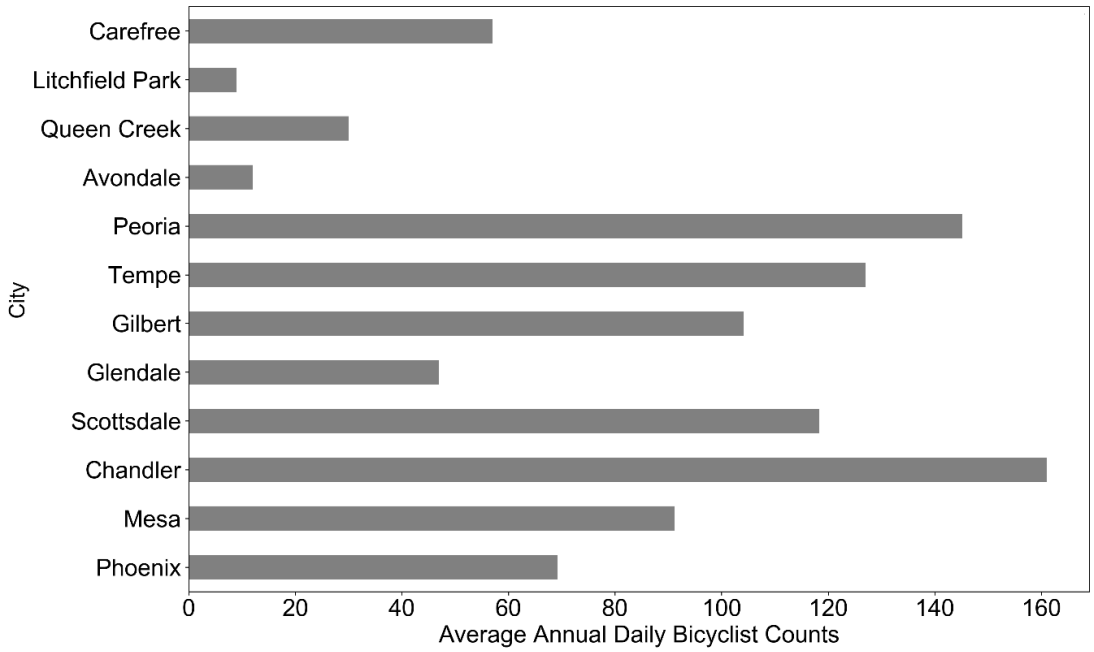


Figure 2.2: Average annual daily bicyclist counts in Maricopa County in 2016.

The count locations covered the most populated regions within Maricopa County and spanned 12 major cities, including Avondale, Carefree, Chandler, Gilbert, Glendale, Litchfield Park, Mesa, Peoria, Phoenix, Queen Creek, Scottsdale, and Tempe. Figure 3.2 shows the AADB counts in order of the population density of each city within Maricopa County. The counters were located across a range of locations, including freeways and arterials with and without bike facilities, as well as bike paths, such as near canals and

trails. The AADB counts were extrapolated based upon the 2-week period counts. Also, owing to the extreme weather conditions, overall ridership is generally lower in the study area compared to other North American cities.

We used an independent test dataset to evaluate the model prediction accuracy, from the city of Tempe, where manual bicyclist counts across 60 locations were available. These manual counts were conducted by a non-profit organization, the Tempe Bicycle Action Group (TBAG), at peak periods in the morning (0700–0900) and evening (1600–1800) on weekdays in the months of April to May and October to November in 2016, and 12,345 cyclists were recorded. The bicycle ridership data collected by the TBAG were used to evaluate the global model accuracy at a smaller spatial scale, just for the city of Tempe.

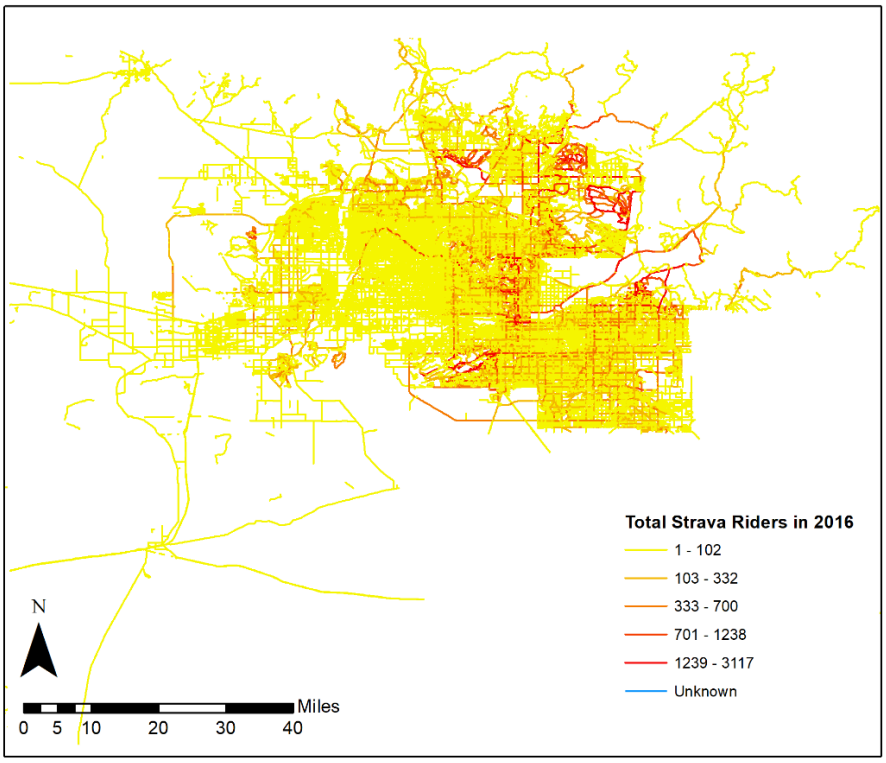


Figure 2.3: Distribution of Strava riders in Maricopa County for 2016.

The Maricopa Association of Governments distributed Strava bicycling data for 2016 for the entire Maricopa County. Strava data included street network shapefiles with anonymized bicyclist count information along with each street segment as well as at street intersections, at a one-minute temporal resolution. The high spatial and temporal coverage of the Strava data in Maricopa County allowed for counts to be obtained in the same locations and time periods as those collected through automated count stations. The total number of Strava riders throughout Maricopa County in 2016 is shown in Figure 2.3. Among all the Strava riders, nearly 76.5% of Strava riders in 2016 in Maricopa County were male, 17.6% were female, and 5.9% did not specify a gender, as shown in Figure 2.4, which indicates Strava riders were not fully representative of the entire population and there was an inherent bias in the ridership data, which requires correction.

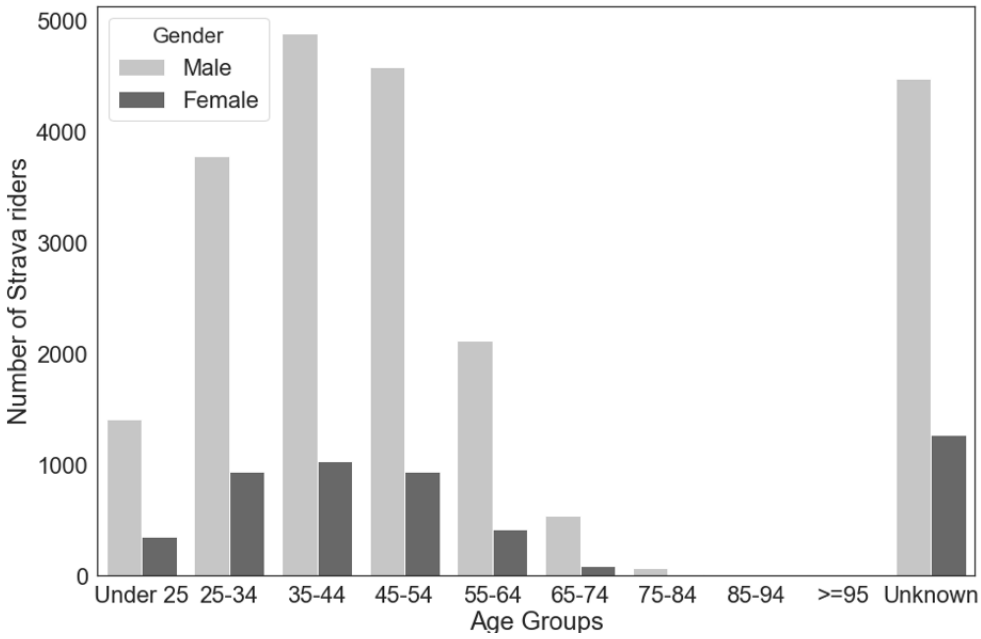


Figure 2.4: Age–gender distribution of Strava riders in Maricopa County for 2016.

In Table 2.1, we list the explanatory geographical covariates used in our model along with their potential relationship with bicycling. The geographical covariates were provided by the MAG for each census block group in Maricopa County. We identified those census block groups which were intersected by a unique street segment and assigned the mean of all the variables in the intersected polygons to the respective street segment. We also used the shortest distance technique to compute the proximity to green spaces, residential areas, and commercial areas for each individual street segment. The shortest distance is the Euclidean or straight-line distance from the nearest land-use polygon of a specific type (e.g., green space/residential area/commercial area) to the street segment. The MAG also provided the shapefiles on land-use classes, which were used to categorize green spaces, residential, and commercial areas.

Table 2.1: Geographical covariates influencing ridership in Maricopa County (2016).

Description	Measure	Source	Year	Resolution	Relevance
Crowdsourced Fitness App	Bicyclist count across street segments grouped by location and timestamp	Strava Metro	2016	Street Segment	Crowdsourced cycling data help predict categories of cycling volumes in urban environments [20,15].
Built Environment	(a) Average daily traffic volume (b) Average segment speed limit	(a) USDOT Federal Highways Administration (b) OpenStreetMap	2016	Street Segment	Built environment has a significant influence on active transportation choices [30,31,18,1]. Improving traffic promotes bicycling[33].
Demographics	(a) Population density (b) % white population (c) Median age (d) % veterans	Maricopa Association of Governments Open Data Portal	2010	AZ Census Block Group	Densely populated areas have higher number of cyclists [32,34]. Ethnicity variations affect bicycle ridership levels.[35].

	(e) % high school educated				
Land Use Mix	(a) Proximity to greenspace (b) Proximity to residential areas (c) Proximity to commercial areas	Maricopa Association of Governments Land Use Data	2016	Street Segment	Nearness to residential areas and green open spaces has shown positive associations with an increase in physical activity [36,1].
Socio-Economic	(a) Median household income	Maricopa Association of Governments Open Data Portal	2010	AZ Census Block Group	Areas with lower income levels tend to bike more [37,38,10].
Commute Patterns	(a) % of population who commute to work with bicycles	Maricopa Association of Governments Open Data Portal	2010	AZ Census Block Group	Frequent bicycle commuters are more likely to have a higher level of education [39].

2.5 Methods

2.5.1 Overall design of the bias-correction framework

We performed the following steps for correcting bias in the Strava data:

(i) The relationship between the Strava ridership data and official counts across 44 locations in Maricopa County (train data) was quantified using ordinary least squares regression.

(ii) Additional geographic data from multiple disparate sources (Table 2.1) were then aligned, controlling for variable multicollinearity, with ridership data from Strava, and a variable selection technique—LASSO—was used to identify the most significant geographical variables from all the listed variables in Table 2.1.

(iii) A generalized linear model with a Poisson distribution was fitted using the observed AADB counts as a dependent variable and the Strava ridership data along with the geographical covariates selected by LASSO, which were outcomes of step (ii), at comparable spatial and temporal scales as independent variables. Using this model, we corrected the bias in the crowdsourced bicycle ridership data by age and ability across Maricopa County using a 10-fold cross-validation across the 44 locations.

(iv) The coefficients of the model fitted in step (iii) were then used to explain the variation in the AADB counts and the bias-corrected predictions.

(v) The best-fitted model from step (iii) was cross-validated, which is a technique used to test the model fit by holding out 10% of the data and training the model with 90% of the data in multiple iterations, and the model with least cross-validation error was used to predict the observed AADB at unknown locations and to create a street-level map of bias-corrected AADB counts in Tempe.

(vi) Finally, the prediction accuracy of the model, shown in step (iii), was evaluated in Tempe across 60 locations where ground truth data for the AADB counts were available (test data).

Each of the steps is explained in further detail in the sections that follow. The exploratory analysis and data preprocessing were performed using Jupyter Notebooks (Kluyver et al., 2016). Spatial analyses were undertaken in ESRI® ArcGIS 9.3 and the model was partly built using both Python 3.5 (Python.org, 2019) and R 3.4 (R Statistical Software).

2.5.2 Quantifying representativeness of crowdsourced movement data

To quantify how the bicycle ridership of all riders is represented by sampling the crowdsourced app ridership, we compared the ridership counts from Strava with official counts from automated bike counter systems installed by the MAG (MAG, 2016) across 44 locations in Maricopa County for a two-week period in the months of April, May, October, and November. The Python package, PANDAS (Python data analysis library) (McKinney, 2012), was used to summarize, match, and extract crowdsourced data counts for each individual road segment in Maricopa County to account for ridership estimates.

Comparisons between the two datasets were made at daily, monthly, and annual levels. We used regression analysis to quantify how much of the variation in bicycle ridership was explained by the crowdsourced data. To do this, we matched counts from Strava, aggregated them into hourly intervals, and matched those to the time windows when official counts were conducted by the MAG. Once counts were matched temporally, we compared both datasets at daily, monthly, and annual levels. We obtained R^2 values using simple linear regression for each time period and retained the volumes with the highest R^2 for further analyses.

2.5.3 Selecting geographic covariates for bias correction using LASSO

In order to correct for the bias in the crowdsourced ridership data, we included the geographical covariates from Table 2.1. Variable multicollinearity, which is the state of high inter-correlations among independent variables, was limited by retaining only those variables which had a variance inflation factor (VIF) below 7.5 (Crawley, 2005). If the variance inflation factor of a predictor variable was 7.5, this meant that the standard error

for the coefficient of that predictor variable was 2.73 ($\sqrt{7.5}$) times as large as it would be if that predictor variable was uncorrelated with the other predictor variables. These covariates were hypothesized to influence bicycle ridership at a geographic scale comparable to that of the Strava data. Spatial joins from the Python library, Geopandas (Jordahl, 2019), were used to link bicycling count data with the geographical covariates.

We used an average of the geographical covariates for all the census block groups that a particular street segment intersected. The distance variables were calculated in ArcGIS using a simple Euclidean distance measured in miles. Since the number of independent variables for our analysis was 15, even after accounting for inter-correlations through VIF, we used a statistical method to select only those variables that explained most of the variance in the overall bicycle ridership. A variable selection technique using LASSO (least absolute shrinkage and selection operator) (Tibshirani, 1996) was applied to select covariates that best explained the bias in the Strava data while accounting for the bias-variance tradeoff (Tibshirani, 1996).

The purpose of LASSO is to apply a constraint on the sum of the absolute values of the model parameters with a fixed upper bound. To do so, the method applies a shrinkage process (also known as regularization), where it penalizes the coefficients of the independent variables, shrinking some of them to zero. The variables that still have a non-zero coefficient after the shrinkage was selected to be inputs of the final Poisson regression model. By using LASSO, we intended to minimize the prediction error of the final AADB counts. The LASSO can be thought of as an additional step, which can help transportation planners choose, from a large set of variables in a study area, only those which can in effect

help improve the prediction results and contribute significantly in explaining variation in the overall bicycle ridership.

Given the set of explanatory variables, x_1, x_2… x_p, and the outcome, y, the observed bike counts, LASSO fits the linear model:

$$\hat{y} = \beta_0 + \beta_1 . x_1 + \ldots\ldots + \beta_p . x_p \quad (1)$$

by minimizing the following criterion:

$$\sum_{i=1}^{n} (y - \hat{y})^2 + \lambda \sum_{j=1}^{p} |\beta_j| \quad (2)$$

In doing so, the non-contributing geographical covariates are shrunk to zero. We ran 200 iterations of the LASSO on our training data using a 10-fold cross-validation approach to obtain the optimal value for λ (tuning parameter), which yielded the minimum cross-validation error on the training dataset for all iterations. The variable selection was performed using the randomized LASSO scores provided by the scikit-learn Python library based on the stability function proposed by (Meinshausen & Buhlmann, 2010). For a cut-off, π_i, with $0 < \pi_i < 1$ and a set of regularization parameters, Λ, the set of stable variables is defined as:

$$\hat{S}^{stable} = \{k : \max(\widehat{\prod_k^{\lambda}}) \geq \pi_i\} \quad (3)$$

With the chosen λ, we retained the variables with a high selection probability and disregarded those with low selection probabilities using a score function, which provided the coefficient of determination, R^2, of the prediction ranging from 0 to 1. The coefficient, R^2, is defined as $(1 - u/v)$, where u is the residual sum of squares and v is the total sum of

squares of the variables retained with the chosen value of λ. The LASSO module from the Python machine learning library, scikit-learn, was used to perform variable selection.

2.5.4 Predicting counts using bias-corrected movement data

We fitted the geographical covariates selected by LASSO to a generalized linear model following a Poisson distribution to explore the relationships among the selected covariates, and the bicycle ridership counts in Maricopa County using Equation (3). We chose the Poisson model as it generates non-negative predictions, which are appropriate for modeling count data.

As shown in Figure 2.2, the geographical variables, as well as the official counts and counts from the crowdsourced app, were provided as inputs to the model. The LASSO variable selection algorithm determined the stable covariates that best replicate the bicyclist counts. Following variable selection, the Poisson model predicted the AADB counts along all street segments in Maricopa County. The regression model was specified as a Poisson distribution with a log-link function (Dobson & Barnett, 2008) as follows:

$$Y_i \sim Poisson\left(\mu_i\right), log\left(\mu_i\right) = \beta_i X_i \quad (4)$$

where:

- Y_i = the AADB counts at site *i*
- β_i = vector of parameters for count site *i*
- X_i = vector of the observed geographical covariates for count site *i*.

The AADB counts were generated by the model across the entire road network in Maricopa County, including paved streets with and without bike facilities. The segments from Strava that were matched in spatial and temporal resolution to the official MAG counts were used in fitting the Poisson model. Hence, we could compare the counts from both sources at only those locations, where both counts were available, which were then used to train our model. The remaining segments that only had Strava counts were used to test the predictive power of the model. The average annual counts from Strava along with the geographical covariates were the independent variables for the model. The significant variables were those with a p-value < 0.001. Since we assumed that our dependent variables (the MAG counts) follow a Poisson distribution with a mean that depends on some covariates, we used a generalized linear model that takes into account the heteroscedasticity in the data.

The Poisson model coefficients were used to predict ridership at all street segments in Tempe. A k-fold cross-validation technique (Kohavi, 1995) was used to determine the best fit for our training data using the Poisson model, and Akaike' s information criterion (AIC) was computed at each step to determine the best-fitting model for our training data. The bias-corrected ridership estimates were then classified using a histogram into five different categories—very low, low, medium, high, and very high.

2.5.5 Mapping predicted counts from bias-corrected movement data

For ease of visualization and to support our validation of the prediction accuracy with independent data, we generated a map for a smaller area and compared the bias-corrected map with the annual Strava ridership map for the city of Tempe, where ground truth data

were available from the TBAG. Results were visualized across the city of Tempe with a uniform color scheme representing each category with varying widths of street segments.

As the ultimate goal of the model was to predict bicycling volumes that were corrected for sampling bias, we applied the model to spatially continuous data from 60 locations across the city of Tempe and predicted annual bicycle ridership across all street segments. The bicycle counts provided by the TBAG were used to determine the prediction accuracy.

We verified our model using a 10-fold cross-validation approach in order to account for overfitting. We performed 100 iterations of the model, splitting the dataset into a train-test sample ranging from 15% to 85%, and chose the model with minimum cross-validation error as the best fit. We then calculated the differences between the predicted and observed AADB counts and analyzed the variation of the differences with the percentage of segments predicted.

2.6 Results

The crowdsourced data from Strava captured 642,298 trips for 28,571 unique bicyclists across Maricopa County for the entire year of 2016. A total of 24,917 riders were captured using automated counters in Maricopa County in 2016. The AADB counts ranged from 0 to 522 with the highest ridership in the city of Chandler and the lowest in Litchfield Park. The average number of daily Strava cyclists at the same locations ranged from 0 to 34 when compared with the Strava data. The manual counts from the TBAG comprised 60 locations within Tempe with a total of 12,151 riders. The ordinary least squares regression analysis between the AADB counts from the MAG and Strava accounted for 76% of the variation between the two datasets.

In Table 2.1, the geographical covariates along with the month and day of count used for determining the most significant variables to use as input for the Poisson model are shown. The tuning parameter, λ, was 1.85, based on the minimum cross-validation error of 0.014 on the training set. In Table 2.2, all input variables used by LASSO are listed. The most significant variables which were not shrunk to zero (λ = 1.85) and had a score above 0.65 were: Distance to residential areas, distance to green spaces, percentage of the white population, median household income, average segment speed limit, and average number of Strava riders. In Table 2.3, a list of the parameter estimates of the Poisson regression on the six variables chosen from Table 2.2 through LASSO is provided.

Table 2.2: Variable importance based on LASSO variable selection (λ = 1.85).

Covariates	**LASSO Scores**
Distance to residential areas	1.00
Distance to green spaces	1.00
% white population	1.00
Median household income	1.00
Average segment speed limit	0.98
Strava counts	0.96
Average daily traffic volume	0.59
% veterans population	0.43
Population density	0.4
% population who commute with bicycles	0.05
Distance to commercial areas	0.02
Median age	0
% Population with at least high school education	0
Count month	0
Count day	0

The model had an AIC of 1832.9 and yielded the lowest mean-squared error of 0.0045 after 100 iterations of cross-validation. The pseudo-R^2 of the fitted model was 0.59. In Table

2.3, the standard errors and 95% confidence intervals of the associated parameter estimates are also highlighted.

Table 2.3: Parameter estimates using Poisson regression.

Dependent Variable: AADB Counts from MAG					
Explanatory Variables (x_i)	**Estimate(log)** (β_i)	**Std. Error**	***p*-value**	**95% CI**	
				Lower	**Upper**
Strava counts	0.17	0.01	<0.001	0.15	0.18
Average segment speed limit	−0.09	0.01	<0.001	−0.11	−0.08
Distance to residential areas	−0.51	0.01	<0.001	−0.59	−0.43
Distance to green spaces	−0.74	0.07	<0.001	−0.88	−0.59
Median household income	−0.09	0.01	<0.001	−0.01	−0.08
% white residents	0.11	0.01	<0.001	0.09	0.14
Intercept	3.78	0.08	<0.001	3.63	3.92

The variables, distance to green spaces, distance to residential areas, median household income, and traffic speed, have an overall negative impact on ridership while the number of Strava riders and the percentage of white population have an overall positive influence on bicycle ridership. The in-sample fit for the entire Maricopa County (where ground truth was available from the MAG) using the Poisson model resulted in an R^2 of 0.64 between the observed and predicted counts (Figure 2.5).

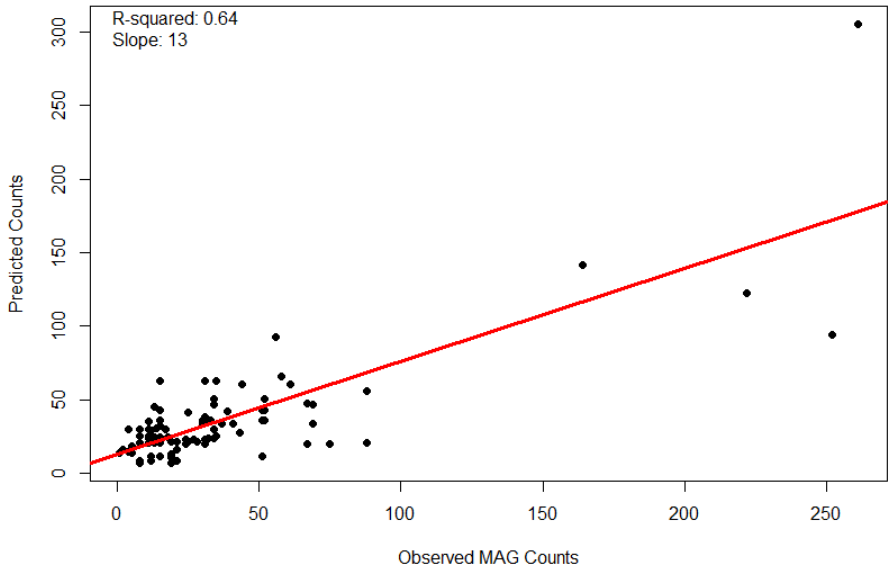


Figure 2.5: Poisson model predicted vs. actual AADB counts for Maricopa County (2016).

Changes in the predicted ridership volume relative to changes in each of the covariates in the model help to demonstrate the contribution of geographic covariates in correcting bias in the overall ridership volume estimates. To provide a baseline for comparison, the intercept of the Poisson model alone provides an estimate of the mean observed counts on a street segment, independent of all other covariates. Table 2.4 shows the factor by which each covariate influences overall ridership.

In this study, the mean AADB count on each street segment represents 43 bicyclists. The parameter estimate for the average number of Strava riders was 1.18 (Table 2.4), which indicates that if the average number of Strava riders on a particular street segment increases by 1, given that all other variables were held at their respective average values, the estimated ridership on that segment would increase to approximately 50.74. In other words, Strava counts account for 1 in every 50 bicyclists along a particular street segment.

Table 2.4: Variation in predicted AADB counts for each variable, with all other attributes, held constant, when the variable is changed by a factor, $e^{(\beta\_i)}$.

Variables (x_i)	**Scale (per unit)**	**Change Factor (e^{β_i})**	**Change in Observed Bicyclist Counts (y) (all other Variables Held Constant at Their Mean)**
Intercept	-	43	-
Strava riders	1 rider	1.18	18% increase
Distance to residential areas	1 mile	0.6	40% decrease
Distance to green spaces	1 mile	0.48	52% decrease
Average segment speed limit	10 mph	0.91	9% decrease
Median household income	$10,000	0.91	9% decrease
% white population	10%	1.12	12% increase

The proximity of a street segment to a residential neighborhood and green spaces was found to impact overall ridership significantly. With every 1 mile increase in the shortest

distance of a street segment from a residential neighborhood, the predicted number of bicyclists decreases by 40% (Table 2.4). Similarly, for every 1-mile increase in the shortest distance between a street segment and green space, the observed bicyclist counts decrease by 52%, ceteris paribus. Ethnicity is a weaker, but still significant, contributing factor to ridership volumes, with ridership counts being positively related to the percentage of the white population in the neighborhood of a street segment. The number of observed bicyclists on a street segment that is located in a neighborhood with a 60% white population will have 12% more observed bicyclists than if it is located in a neighborhood with a 50% white population, ceteris paribus.

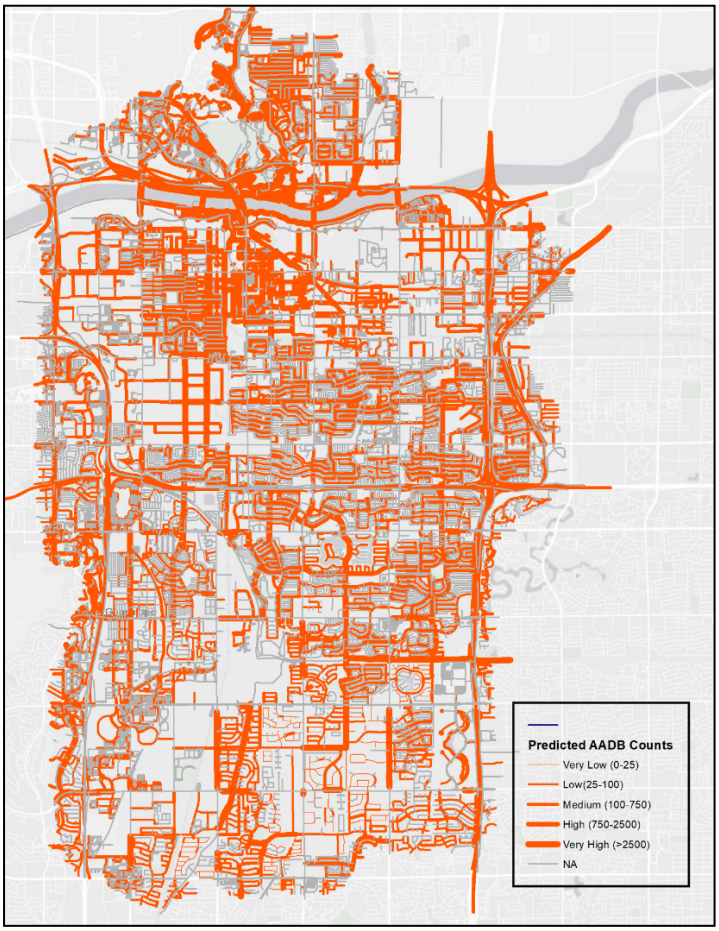


Figure 2.6: Predicted bicycle AADB counts for the entire street network of Tempe in 2016.

Additionally, high values of median household income and increased speed limits were found to be associated with low overall ridership. The parameter estimates show that the observed number of bicyclist counts decreases by 9% for every $10,000 increase in average income whereas for every 10 mph increase in the average speed limit on a particular street segment, the predicted number of bicyclists decreases by 9%, ceteris paribus.

Based on our model, we predicted bias-corrected ridership volumes across the city of Tempe, shown in Figure 2.6, classified using Jenks' classification into five categories: Very low (0–25), low (25–100), medium (100–750), high (750–2500), and very high (2500+).

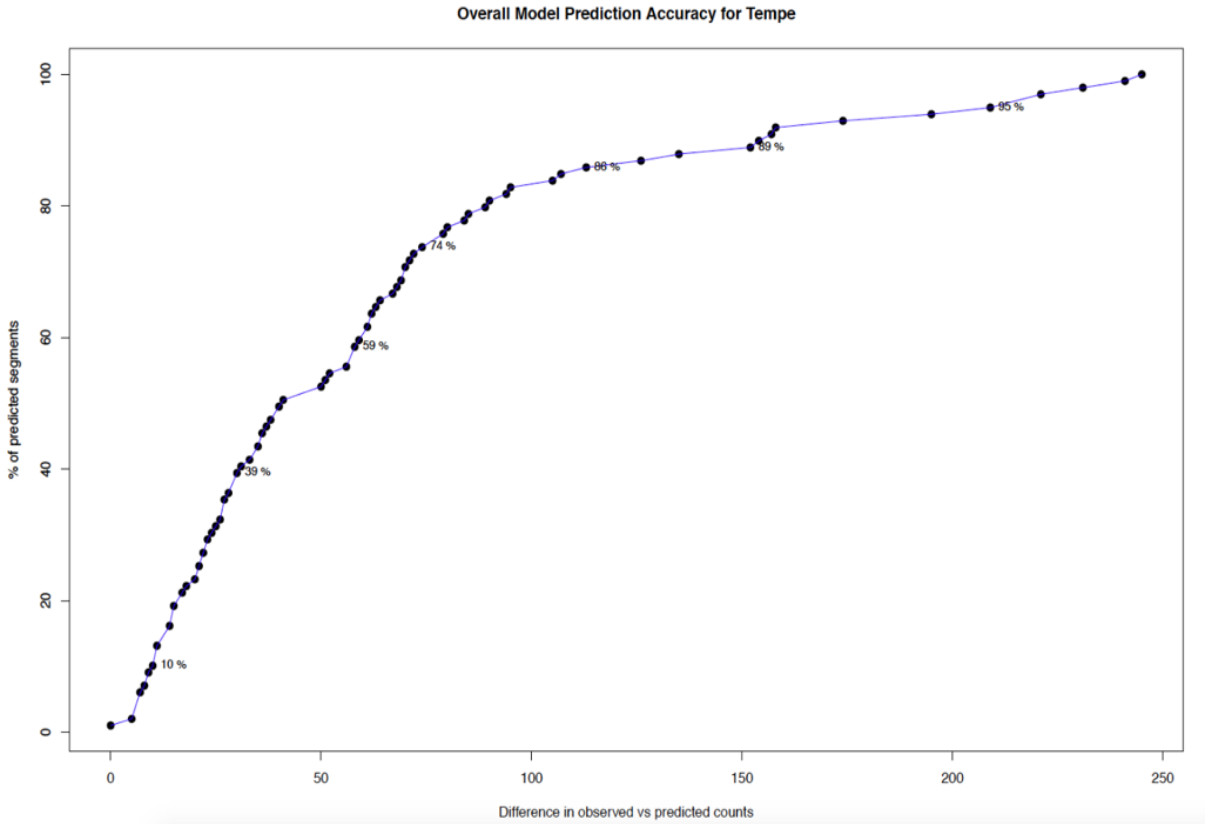


Figure 2.7: Model Prediction Accuracy for Tempe in 2016.

The thin lines indicate streets with a low volume of bicyclists while the thicker lines indicate streets with a high volume of bicyclists. In Figure 2.7, the result of the prediction accuracy as a function of the difference in the predicted AADB counts from the observed counts across 60 count locations in Tempe is given. Overall, for 59% of the segments, we predicted ridership volumes within ±50 AADB, 86% of the segments were within a ±100 AADB, and 95% within ±200 AADB.

2.7 Discussion

We were able to correct the bias in crowdsourced data from Strava using a set of covariates describing the street location and from this we were able to generate maps representative of all bicyclists at a street-level spatial resolution. The correlation between the Strava and official counts alone can explain nearly 52% of the variation in overall bicycle ridership, however, we were limited by the availability of ground truth data, which could help improve the R^2 further. Our goal in combining geographic covariates with Strava counts was to account for additional factors that influence bicycle ridership and may not be captured solely by crowdsourced sampling. These variables helped in making necessary adjustments for the estimation of observed counts of all bicyclists, including those not using the Strava app, thereby correcting bias in crowdsourced data from Strava.

The Poisson regression approach has frequently been used in bicycle crash analysis by (Griswold et al., 2011) and (Hankey et al., 2012) as well as for exposure measurement of accidents by (Hamann & Peek-Asa, 2013). A probabilistic joint analysis approach has been used for correcting sampling bias in species distribution models by (Fithian et al., 2015).

Using the mixed-model approach, this paper proposes a new technique in the bias correction of crowdsourced data for physical activity mapping from bicycle ridership.

The results of our study indicate that bias correction of crowdsourced data may prove to be a useful method for the estimation of bicycle ridership in North American cities. Our results for the city of Tempe (Figure 2.7) indicate that for 80.3% of road segments, where ground truth data were available, estimated bicycle counts were correct to within 25% of the observed counts (± 50 riders). Our findings are in alignment with recent research by (Jestico et al., 2016) and (Griffin & Jiao, 2015), who found strong relationships between Strava and all bicycling ridership in North American cities.

As expected, the proximity of a street segment to a residential neighborhood had a significant influence on the overall bicycle ridership. Most segments that are close to a residential area in Tempe have better road infrastructure, including paved sidewalks and dedicated bike lanes. This encourages inexperienced bicyclists to ride safely and also adds comfort to the overall bicycling experience in general for riders of all ages and abilities. Hence, transportation planners should pay more attention to the use of wider streets with dedicated bike lanes in residential areas to help increase active transportation among riders irrespective of their age and ability. Similarly, closer proximity to green space also had a positive influence on ridership. One reason for the relationship between bicycling and green space may be that green corridors, which connect the bicycle network within a city, facilitate increased overall bicycle ridership.

The positive coefficients for Strava counts and the white population percentage are in alignment with the fact that ethnicity influences ridership and that in urban areas, generally, a higher proportion of white residents ride bicycles than non-white residents. Previous

studies by (Winters et al., 2010) and (Huang et al., 2009) have shown that positive relationships exist between ethnicity (white, non-white) and physical activity. Our model results also show that ethnicity is an important factor to use when correcting bias Strava sampled bicycle ridership volumes.

Median household income was also significant in influencing overall bicycling ridership. It has been found in previous research (Reis et al., 2013) that people from lower economic backgrounds are less likely to adopt an active lifestyle and our results also indicate a similar trend. Bicycling should be made more cost-effective for daily use by commuters in order to promote active transportation.

High-speed limits are often correlated with roads having greater concentrations of larger-sized vehicles and more traffic, both of which are major deterrents to bicycling in general (Winters et al., 2010; Piatkowski & Marshall 2015) and often result in crashes (Chen & Shen, 2016). These results suggest policy directions for the safety of bicyclists by means of the reduction of speed limits on busy traffic corridors, and the provision of dedicated bike lanes or green zones on major streets connecting areas of interest (schools, business centers, parks, shopping complex, etc.) to attract riders of all ages and abilities.

The model framework proposed in this study can be used for correcting bias in Strava riders from other cities or bicyclist counts comparable in space and time obtained from other bicycling apps. However, our study has a few limitations. The official counts from the MAG were collected at 44 locations scattered across the whole of Maricopa County. As the segments containing available ground truth data were mostly within the city limits, data from open spaces on the outskirts of cities were sparse. The choice of geographic covariates was specific to the study area and might vary for different geographic settings,

depending upon their relevance. The model could have been improved in terms of prediction accuracy if more ground truth data were available across diverse locations to train the model. Street conditions with a low prediction accuracy can be targeted for future sampling and organized data collection efforts can be proposed for better quality data, which could help in the improvement of the model. With the availability of sufficient data, further studies could examine the spatial heterogeneity of bias-corrected ridership across varying geographies using localized regression on more realistic conditions across larger spatial scales.

2.8 Conclusion

Big, crowdsourced data from fitness apps, like Strava, on bicycling volumes, can be used to make informed decisions on factors that influence ridership in urban areas on a much finer spatial scale. We introduced a new method for correcting bias in crowdsourced data with the help of a three-step mixed-model approach by quantifying crowdsourced data and official counts in a specific geographic region, using LASSO to choose the most significant geographic variables that could correct bias, and finally, fitting the covariates along with the crowdsourced data by means of Poisson regression to predict and map overall ridership in the region. The method developed in this study is broadly applicable for correcting bias in crowdsourced bicycling data when official counts and geographical data are available at comparable spatial and temporal resolution. Based on the results of this paper, in the future, it is suggested that local transportation authorities should work closely with researchers to improve the coverage of official count data, helping them to identify locations to place counters so that a denser spatial coverage, as well as more ground

truth data, are obtained to improve the model's performance. The proposed bias correction model, with detailed data that is continuous through space and collected repeatedly in time, can help transportation planners in making informed decisions related to bicycle infrastructure planning to promote healthier lifestyles among urban residents of all ages and abilities. Detailed maps of bicycling ridership are critical to professionals in making decisions regarding infrastructure investment and policy changes that support active transportation. The framework developed in this paper can be used as a generalized risk assessment and exposure modeling tool to benefit accident prevention among bicyclists, with sufficient availability of accident data from crowdsourced platforms, like Bikemaps.org (Nelson et al., 2016) and provide an estimate of bias-corrected bicyclist volumes for infrastructure planning to enhance comfort among bicyclists and promote active modes of transportation for healthier lifestyles among wider demographics.

CHAPTER 3

DETECTING CHANGES IN MOVEMENT PATTERNS FROM CROWDSOURCED DATA

3.1 Abstract

Monitoring change is an important aspect of understanding variations in spatial-temporal processes. Recently, big data on mobility, which are detailed across space and time, have become increasingly available from crowdsourced platforms. New methods are needed to best utilize the high spatial and temporal resolution of such data for monitoring purposes. These data can be considered mappable time series but are challenging to use owing to varying sampling rates and issues of temporal misalignment. We present a functional data analysis technique for change detection from spatial-temporal data by analyzing big, crowdsourced data captured continuously in time while addressing non-elastic rate variations in the underlying spatial-temporal processes. Using data from the Strava fitness app, captured every minute, we quantified ridership changes in Phoenix between 2017 and 2018 at the street-segment level. Hourly and monthly changes were classified into four categories and mapped along with exposure density. Using spatially and temporally continuous data our study advances the existing approaches to mobility analysis, by capturing data about the underlying processes, rather than monitoring change between discrete snapshots of time. Our method is reproducible by practitioners for monitoring changes from crowdsourced ridership data and for making necessary infrastructure changes to assure the safety of bicyclists.

3.2 Introduction

Monitoring change from continuous time-series data is critical for cities to understand travel behavior and make targeted decisions related to transportation infrastructure changes that can improve the overall safety of their residents. Through monitoring, policymakers are more prepared to meet rising infrastructure demands (Miranda-Moreno et al. 2013) and ensure accessibility to existing infrastructure more expeditiously (Boss et al. 2018). Change detection is essential to characterize the impacts of sudden fluctuations on overall spatio-temporal processes (Alaya et al. 2020), particularly where we observe changes in the frequency and/or in the intensity across multiple scales. Detection of changes is thus an essential step before performing any descriptive or predictive analysis.

Growth in the availability of crowdsourced GPS data from smartphones has created an alternative source of high-resolution spatial-temporal data to enable researchers to understand mobility patterns. Although these datasets are biased towards a specific demographic (males between 25-45 years of age), they can be used as an indicator of ridership once the bias is accounted for by including geographic covariates (Roy et al. 2019). Crowdsourced fitness apps like Strava (Strava Metro 2016) have been collecting anonymized bicycling trip data at the minute level, which can be used for monitoring change. These fine-grained bicycling trips can be represented using a functional form, as they are collected continuously over time. In the context of bicycling ridership, the major problems are that data are captured from different bicyclists who record their trips which are then aggregated in the background by Strava. We hypothesize that this data aggregation from millions of users at different sampling rates introduces errors that need to be

accounted for. The mismatched sampling rates can go a long way in classifying a street that shows an increase in ridership as a decrease or no change or vice versa.

Initial research (Boss et al. 2016, Nelson & Boots 2008) has shown that it is possible to detect changes between two time periods – and that the changes are representative of actual changes in infrastructure. Such studies quantify the spatial variations of change in ridership using two snapshots in time. Some studies (Yang et al. 2018) have looked at spatial change detection from GPS trajectories however, they ignored the temporal component, whereas other studies (Kang et al. 2019) have addressed the scale issue in land-cover change and activity zone detection from social-media platforms (Liu et al. 2019).

To utilize large volumes of raw timeseries data, we must identify analytical methods that also account for patterns in changes across multiple scales, as the underlying data generating processes vary both in space and time. There is a lack of a well-defined framework for extracting actionable insights from such big data for change monitoring across scales. To bridge this gap we propose a functional data analysis (FDA) (Ramsay & Silverman 2005) technique for mapping changes in bicycle ridership patterns. FDA is a statistical technique that is used to analyze high-frequency data represented as curves varying over a continuum like space or time. The fundamental structure of an FDA framework is functions representing the underlying data. FDA has found applications in several areas of research including ecology (Bourbonnais et al. 2017, Gurarie et al. 20), epidemiology (Aston and Kirc, 2012), remote sensing (Bourbonnais et al. 2017), physical activity recognition (Choi et al. 2018)) and traffic volume forecasting (Wagner-Muns et al. 2017, Kim, 2019).

We hypothesize that complex phenomena like bicycle ridership patterns, which involve several underlying factors involving route choice, built environment characteristics, infrastructure availability, and safety, are best viewed as observations from a dynamical process. The observed time-series results from a complex entanglement of several variables which are difficult to single out. Hence, small changes in certain conditions in the data generating process can cause non-elastic effects on the observed time-series of such big data, without altering the essential trends of the time series. Failure to account for these effects, increases the observation variance, thereby leading to erroneous results for variance-based analysis (like clustering). By removing the effect of elastic time-warping, we make the results of subsequent cluster-based analysis more stable and less prone to small misalignments in time. The novelty of our approach lies in applying FDA techniques to timeseries data gathered via crowdsourced platforms like Strava and generating a visual representation of those changes through a map for planning purposes.

In this work, our goal is to demonstrate a method for detecting and mapping change in data collected continuously through time. To meet this goal we (1) quantify the change in bicycle ridership using a special case of elastic FDA known as the square-root velocity function (SRVF) representation and (2) visualize the temporal changes across hourly and monthly scales. We generate street-segment level maps for different time scales that enable practitioners to make targeted decisions regarding bicycle infrastructure planning.

3.3 Study Area

Our study area is the City of Phoenix (Figure 3.1), which lies within the state of Arizona in the USA. Phoenix is the largest metropolitan city in the county of Maricopa in

Arizona with a population of 1,563,001 (US ACS 2015, City of Phoenix). Approximately 1.12% of the population who commute to their workplace use bicycles as their preferred mode to work with the highest weekday ridership exceeding 270 bicyclists per day (City of Phoenix, Bicycle Master Plan report, 2015). Bias-corrected Strava ridership in Phoenix is representative of nearly 76% of overall bicycling activities with bicyclist safety along with income and gender being the strongest indicators of overall ridership (Nelson et al. 2020). The city has an entire street network stretching to nearly 8,000 km, with about 1,140 km of total bicycle lanes that include 960 km of on-street bicycle facility and 190 km of off-street bicycle paths, 42 bicycle and pedestrian bridges/tunnels spanning the entire city (City of Phoenix, Bicycle Master Plan, 2015).

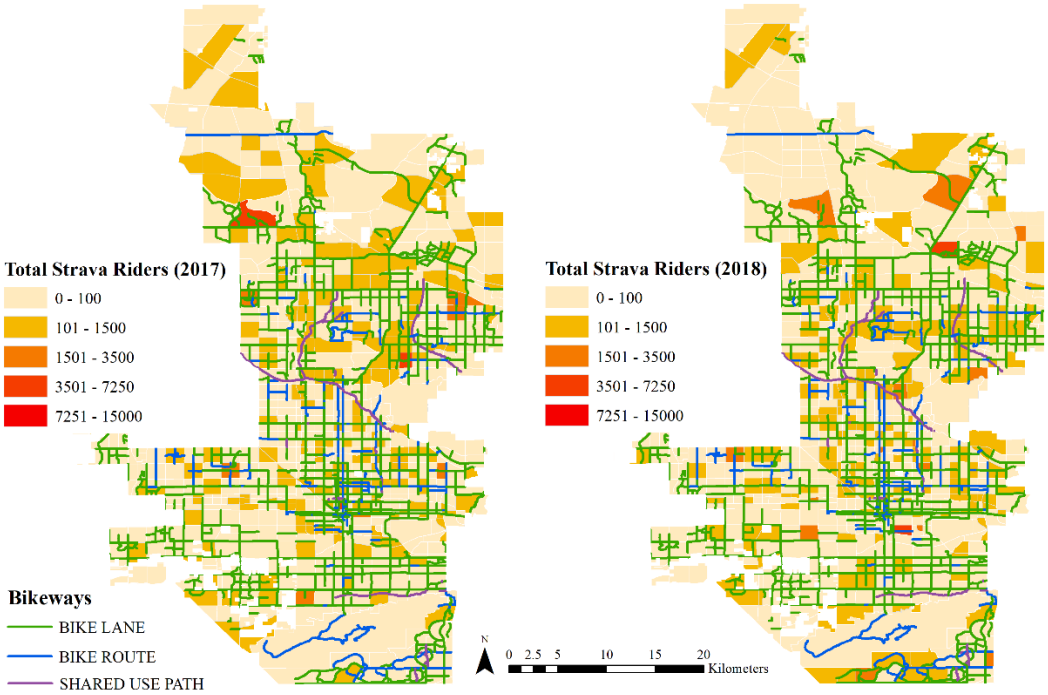


Figure 3.1: Map showing the spatial distribution of Strava riders across all traffic analysis zones along with bikeways in the City of Phoenix (2017-2018).

3.4 Data

The total number of bicycle trips in Phoenix increased from 52,976 to 74,191 between 2017 and 2018 (Strava Metro 2019). The City of Phoenix has gathered bicycle ridership from Strava Metro as part of a data acquisition effort by the Maricopa Association of Governments for estimating ridership estimates annually. Strava Metro provides information about anonymized bicycle trips recorded through the Strava fitness app. The data consists of activity counts (i.e. bicycle trips) per segment of transportation infrastructure in the Phoenix region, recorded every minute of the day.

We chose the period between 2017 and 2018 as several minor/major changes took place in the bicycling infrastructure during this period, which enables monitoring how ridership patterns varied before and after the changes were put into effect. Table 3.1 shows the trip information and the number of total activities recorded in each year.

Table 3.1: Summary of Strava ridership in Phoenix from 2017 to 2018.

Year	No. of commute trips	Total no. of activities	No. of street segments	% Male Strava riders	%Female Strava riders
2017	131,081	1.74 million	78,174	76.9%	18.4%
2018	138,714	1.78 million	74,191	76.5%	19.7%

Strava is commonly used by recreational bicyclists which introduces a bias in the overall sampling of bicycle counts, which can be adjusted using additional geographical covariates (Roy et al., 2019) but in dense urban areas correlates with all bicyclists (Boss et al. 2018). The demographics of the Strava users in Phoenix are not representative of the general bicycling population, there are differences in both gender (Table 3.1) and age. The

percentage of male Strava users (76%-77%) is higher than the percentage of female cyclists in the Phoenix region (17%-19%). The trends in the Strava data used in this study are similar to the age and gender trends of crowdsourced data used in other bicycling studies (Griffin and Jiao 2015, Romanillos et al. 2016). We also use additional data from the City of Phoenix showing the bicycle crashes representative of the time period of study.

3.5 Methods

Our study can be broken into three main objectives – first, we convert Strava ridership volumes to time-series representing it as functional data, second, we then use a temporal alignment technique using square-root-velocity-function (SRVF) (Srivastava et al. 2011) to account for temporal variability and quantify change. Finally, a functional K-means clustering of the change in ridership is used to group street segments into different clusters based on the functional means of the change clusters.

3.5.1 Generating functional curves from crowdsourced movement data

Before detecting changes, it is essential to estimate the underlying spatio-temporal processes first, and the subsequent analysis and inference are performed on the estimated continuous processes, which are referred to as the fitted functional data. Therefore, to highlight changes in ridership we first pre-processed the Strava Metro data and generated functional curves for individual street segments throughout the year aggregated to two different periods – each hour of the day to assess daily trends and each month of the year to observe annual trends. The typical annual ridership trends (Figure 3.1) appear similar

across both years, but as we focus on a finer temporal resolution like hourly/weekly/monthly noticeable changes tend to become more prominent.

Table 3.2: Features for functional data analysis on Strava ridership between 2017 and 2018.

Name	Operationalization	Time Period	Relevance
Mean ridership	The average number of bicycle trips at each temporal unit	Daily, Monthly	Understand hourly variability in ridership volumes (Brum-Bastos et al, 2019)
Mean Weekday Ridership	The average number of bicycle trips on weekdays at each temporal unit	Daily, Monthly	Weekday peak-period ridership helps identify commute patterns among riders & scale Strava data (Dadashova & Griffin, 2020)
Mean Weekend Ridership	The average number of bicycle trips on weekdays at each temporal unit	Daily, Monthly	Weekend peak-period ridership represents higher proportions of recreational riders. (Jestico et al., 2016)
Normalized Total Ridership	The ratio of the total number of bicycle trips in the individual temporal unit and the sum of riders across all temporal units	Daily, Monthly	Represents the proportion of all activity counts that occurred within that period on each segment. (Boss et al, 2018)

From the functional curves of bicycle ridership, we computed the average hourly and monthly activity count for bicycling on weekdays, weekends, and during the entire year for 33,101 segments using the variables listed in Table 3.2. Next, we normalized the mean of all the ridership variables to represent proportions of ridership in each time unit ranging from 0 to 1 for both daily and monthly periods. The scaled ridership data were finally used to represent the temporal profiles of Strava ridership in the selected periods.

3.5.2 Temporal alignment of functional curves

In the first step of our analysis, we removed a substantial amount of noise from the Strava ridership data and temporally aligned the hourly ridership counts as a function of time for similarity analysis. For temporal alignment, we adopted functional data analysis techniques based on the square root velocity function (SRVF) representation of the normalized hourly and monthly Strava ridership counts that would rectify temporal misalignment in the ridership data (which is now considered a signal) by separating its phase and amplitude components. The SRVF method allows the development of proper Riemannian metrics (Srivastava et al. 2011) over time series. It overcomes some limitations of Dynamic Time Warping such as the 'pinching' effect (Marron et al. 2015) which aligns completely different signals to each other by applying a warping function even though their phase and amplitude are not completely in synchronization.

We first computed hourly ridership volumes (i.e. the average number of bicycling trips along each street segment) from raw Strava Metro data. For ease of analysis, we defined these hourly ridership volumes across a street segment as our function x. Then, x was converted into a corresponding SRVF representation to compute the Fréchet mean (Srivastava et al., 2011) for each street segment. For each street segment, the original ridership function x was aligned by composition with the estimated warping functions as shown in Equation (1). The detailed warping functions are based on the Fischer-Rao metric and Frechet means discussed in the paper by Srivastava et al. (2011).

$$[\widetilde{x}] = [x] \circ \gamma \quad (1)$$

Consequently, a new data set was created from which features could be extracted in a sliding window procedure (non-overlapping) with varying window lengths. We chose

the warping function that resulted in the best fit. The warping function was computed by solving Equation (2)

$$\gamma^* = \underset{\gamma \in \Gamma}{\operatorname{argmin}} \left\| u - (q \circ \gamma)\sqrt{\dot{\gamma}} \right\| \quad (2)$$

The γ is the warping function which is solved for using dynamic programming to get the optimal alignment of the curves (Srivastava et al., 2011), u is the Frechet mean (Srivastava et al., 2011) obtained from the training phase, Γ is referred to as the warping group, and q is the SRVF representations of given functions defined as q(t) = $sign(f\dot{(}t))\sqrt{|f\dot{(}t)|}$, where f is the original timeseries function. The warped function $\widetilde{q_t}$ is given by equation (3).

$$\widetilde{q_t} = \frac{\frac{d}{dt}(f \circ \gamma)(t)}{\sqrt{\left|\frac{d}{dt}(f \circ \gamma)(t)\right|}} = (q \circ \gamma)(t)\sqrt{\dot{\gamma}(t)} = (q, \gamma) \quad (3)$$

We use the nonparametric form of the Fisher-Rao metric (Srivastava et al., 2011) for analyzing SRVFs. In order to align the functions, we define an elastic distance d between two curves representing the bicycle ridership for a street segment on the functional space S given by equation (4). The solution to the optimization over Γ can be solved using dynamic programming.

$$d([q_1], [q_2]) = inf_{\gamma \in \Gamma} \|q_1 - (q_2, \gamma)\| \quad (4)$$

The functional curves were aligned in order to remove unnecessary noise from the raw ridership data so the difference between the curves and the respective mean curve in each year could be compared without altering the phase and amplitude of the functional representations.

3.5.3 Calculating changes from aligned functional curves

Once the alignment for both the years was completed using the Frechet means of each curve, we generated a mean signature from the aligned data corresponding to the overall hourly and monthly ridership across all street segments for 2017. Next, we quantified the functional change in ridership patterns in consecutive years by calculating the difference of the aligned function of each street segment in a specific year from the functional mean curve for all street segments in the previous year as shown in Equation (3).

$$C_i = \gamma_i^* - \mu_{i-1} \quad (3)$$

where C_i is the functional change in ridership for year *'i'*, γ_i^* is the temporally aligned functional ridership in the year i and μ_{i-1} is the mean functional curve for temporally aligned ridership in the previous year $(i-1)$. The process was repeated for all street segments in the study area to generate an $N \times M$ matrix where N represents each hour of the day and M represents the number of street segments. We calculated the difference between the mean curve of the previous year (2017) with individual curves in the following year (2018) as the mean curve computed through SRVF essentially helps in reducing the noise in the raw data. We did not compute the difference between individual curves in both years due to the high computational cost associated with the process. Since the mean curve is the average representation of the ridership trend in a single year we use it as a standard signature and subtract individual curves from that mean curve to calculate the difference.

3.5.4 Clustering functional changes to generate change classes

With the change (C_i) computed for each street segment, we ran a functional K-means clustering to group similar and dissimilar streets into 'k' groups. We determined the optimal 'k' for grouping the street segments using the gap statistic method (Tibshirani, Walther & Hastie 2001), based on the within-cluster sum of squares that measures the variability of the observations in each cluster. Once the desired number of 'k' clusters was determined, we visualized these groupings to identify and categorize ridership patterns.

The functional K-Means is a distance-based clustering technique that defines clusters so that the total intra-cluster variation (known as the total within-cluster variation) is minimized. Given a set of 'n' SRVF-aligned hourly ridership corresponding to 'n' street segments in Phoenix, we partitioned these street segments into 'k' groups which were the pre-defined number of clusters we wanted to extract. K-Means groups street segments in a manner such that the change in hourly & monthly ridership within the same cluster are as similar as possible, whereas street segments from different clusters are as dissimilar as possible. The similarity is determined using a similarity index '$\boldsymbol{\rho}$' between two curves c_1 and c_2 is measured using equation (4) proposed by Sangalli et al. (2009).

$$\boldsymbol{\rho}(c_1, c_2) = \frac{1}{d}\sum_{p=1}^{d} \frac{\int_R^{\infty} c'_{1p}(s)c'_{2p}(s)ds}{\sqrt{\int_R^{\infty} c'_{1p}(s)^2 ds}\sqrt{\int_R^{\infty} c'_{2p}(s)^2 ds}} \quad (4)$$

Here $c_{ip}(s)$: $\mathbb{R} \rightarrow \mathbb{R}^d$ indicates the *pth* curve representing the SRVF-aligned hourly ridership of street *p* from a set of *d* curves given by, $c_i = (c_{i1},.....,c_{id})$ aligned using a function h(s): $\mathbb{R} \rightarrow \mathbb{R}$, derived using Equation (2). The similarity index (c_1,c_2) geometrically represents the average of the cosines of the angles between the derivatives of homologous components of c_1 and c_2. The two curves are similar when the value of $\boldsymbol{\rho}$ is 1, which

happens when both c_1 and c_2 are identical except for shifts and dilations in the phase and amplitude components. For a set of N curves $\{c1,....,cn\}$ aligned with a set of k functions $\varphi = \{\varphi 1,....,\varphi k\}$ obtained from equation (2), we assign a curve ci to cluster *'j'* which is defined as λ(φ,c) in equation (5)

$$\lambda(\varphi,c) = \min\{r: c \in \Delta r\ (\varphi))\} \qquad (5)$$

where Δr is the similarity operator selecting the function φ with which the curve c achieves the highest similarity index. In effect, this means that if, $\lambda(\varphi,c) = j$ then the similarity index of curve c when aligned to function φ is at least as large as the similarity index obtained by aligning c to any other function φ_r, where $r \neq j$.

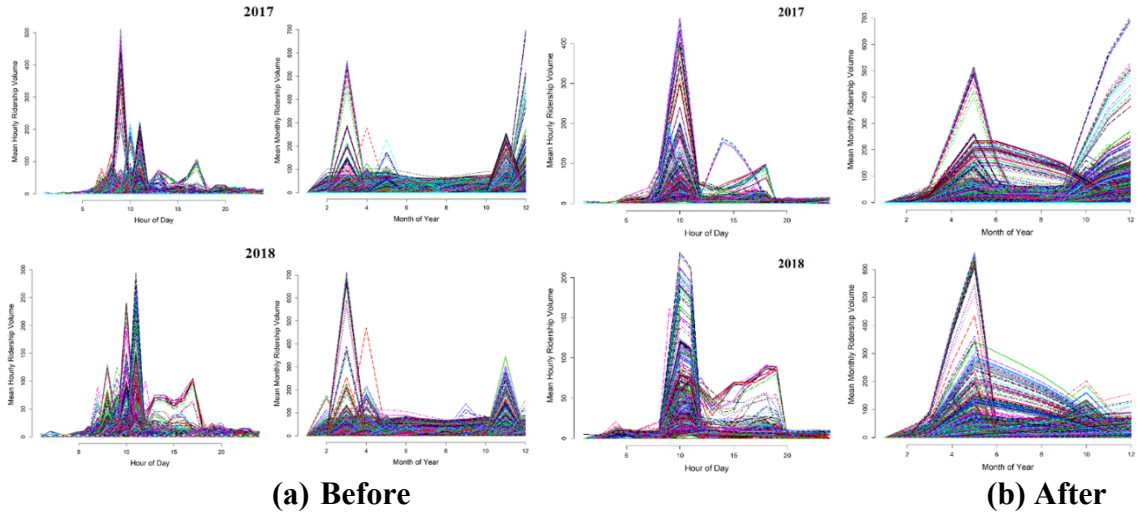
3.5.5 Mapping change classes to visualize changes in movement patterns

The mean functional change was finally used to categorize street segments into 'k' groups. We calculated the mean and coefficient of variation of hourly, weekday, weekend, and total ridership along with the root mean squared of the functional change of the streets within each cluster. We then generated named categories based on the summarized cluster statistics and visualize the results of the K-Means clustering by color-coding each street segment by a unique color scheme corresponding to the category to which it belongs.

Finally, we created a map for the entire city of Phoenix highlighting changes between 2017 and 2018 both at the hourly and monthly scale. To identify the potential causes for the change in ridership in each consecutive year we also incorporated an additional map layer indicating bicycle crash density in the City of Phoenix and infer the reason for changes by overlaying the results.

3.6 Results

The functional curves of the hourly and monthly patterns of the raw ridership volumes are shown in Figure 3.2. These curves were then aligned using SRVF to remove inconsistencies and mismatches in phase and amplitude of the functions. Figure 3.2 shows the original temporal profiles of ridership for all street segments in our study area for 2017-2018 along with their aligned temporal profiles. Similar profiles have been generated from single bicycling counters (Miranda-Moreno et al. 2013) but using Strava allows much higher spatial resolution (every single segment within the city's street network) along with the temporal richness (every hour of a day during the entire year).

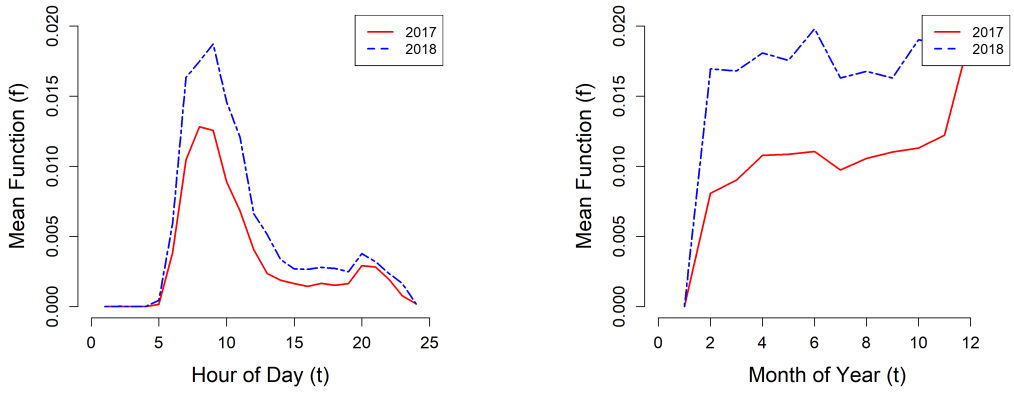


(a) Before **(b) After**

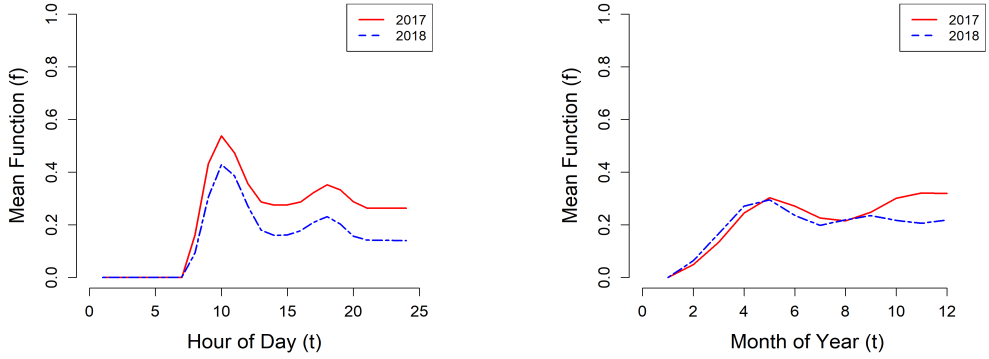
Figure 3.2: Functional curves of actual Strava ridership in 2017 and 2018 at the hourly and monthly scales before (a) and after (b) alignment.

While Strava provides data based on only a sample of riders and there are demographic biases in the app users, research has shown the spatial patterns in this ridership data correlate with bicycle ridership volumes, especially in dense urban areas (Jestico et al. 2016, Boss et al. 2018).

Post-alignment we also calculated the mean signature of the hourly ridership patterns in 2017 and 2018 using the Fischer-Rao metric and found changes in mean ridership behavior at the hourly level across both years as shown in Figure 3.3. It clearly shows how the overall distribution is bimodal with peak periods around 6 am - 9 am and 6 pm - 8 pm. There are some outlier peaks throughout the day but these two peaks are quite prominent.



(a) Before alignment



(b) After alignment

Figure 3.3: Functional curves of normalized mean Strava ridership for 2017 and 2018 before and after temporal alignment.**Note that the bimodal trends revealed post-alignment are more interpretable in general, as also indicated by the scale on the vertical axis. Before alignment, the trends are not as clear and the scale on the y-axis shows an order of magnitude lower values, which suggests that averaging the unaligned data results in loss of structure.

The normalized differences of functional curves for 2018 ridership from the mean signature of the 2017 ridership (Figure 3.3) were used to generate clusters. We varied the number of clusters as shown in Figure 3.4 to choose 4 as the optimal value for 'k' based on the gap-statistic.

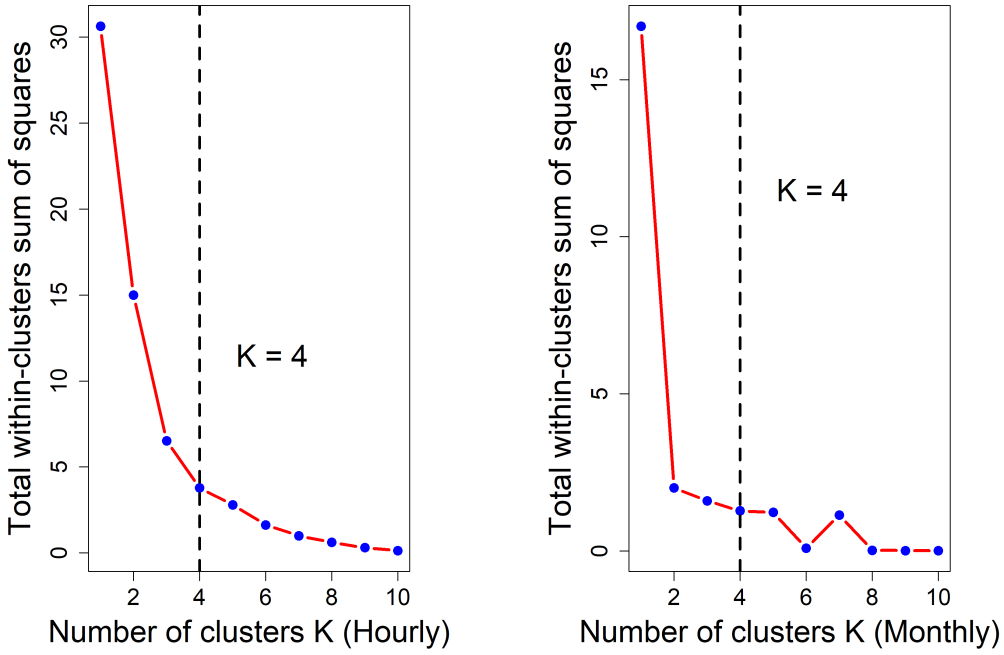


Figure 3.4: Determining the optimal number of clusters using different values of 'k'.

Figure 3.5 represents the individual change in ridership functions associated with each street segment grouped into four different clusters. The grey lines indicate the functional of the aligned hourly ridership in 2017 and the colored line indicates the overall cluster center of the functional curves in that cluster.

The summary statistics of each cluster shown in Figure 3.5 are listed in Table 3, which identifies the percentage of streets (n) in each cluster (k) and the highest, lowest and

average daily ridership along with the mean change of ridership (c_k) in each cluster 'k' calculated as the root mean square across all hours of the day and months of the year in each group. We also calculated the bicyclist exposure per cluster as the ratio of the number of crashes that occurred in the streets network and the total length of streets per cluster.

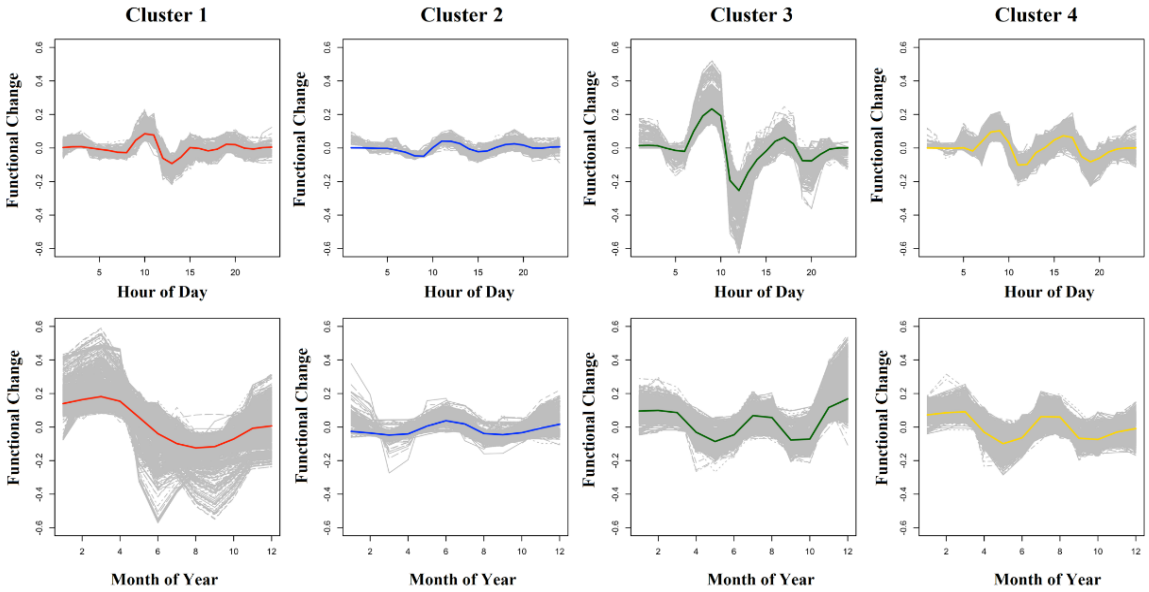


Figure 3.5: Clusters showing streets grouped by the functional change in ridership for hourly and monthly changes.

We show visualizations of the spatial distribution of the clusters based on the functional change in ridership at both hourly and monthly scales in Figure 3.6. The different categories of ridership changes listed in Figure 3.6 indicate higher changes during peak periods at the hourly scale were more prominent near the downtown area.

Table 3.3: Summary of Strava ridership in each of the 4 clusters shown in Figure 6 based on the functional change in Strava ridership from 2017 to 2018

Time Period	Cluster	% of Segments	Mean Functional Change	Weekday ridership		Weekend ridership		Daily/Annual Ridership		Bicyclist Exposure (No. of crashes/ Road length)	Category
				Mean	C.V.	Mean	C.V.	Mean	C.V.		
Hourly	1	31.15	0.04	3.59	1.13	5.48	0.82	42.16	2.24	0.51	High off-peak
	2	33.64	0.02	2.20	1.22	3.35	0.75	8.64	1.14	0.59	Low off-peak
	3	6.61	0.11	7.21	1.76	9.16	0.79	105.23	1.75	0.60	High peak-period
	4	28.60	0.05	4.66	0.92	5.67	0.77	82.95	1.80	0.35	Low peak-period
Monthly	1	11.33	0.08	5.01	0.94	7.63	0.78	123.61	0.97	0.41	High Winter
	2	42.53	0.02	4.32	1.12	6.30	1.12	67.90	2.98	0.60	Low Summer
	3	19.87	0.06	5.03	1.03	7.65	0.77	105.43	2.02	0.40	High Summer
	4	26.27	0.05	4.88	0.99	7.56	0.85	118.60	2.11	0.49	Low Winter

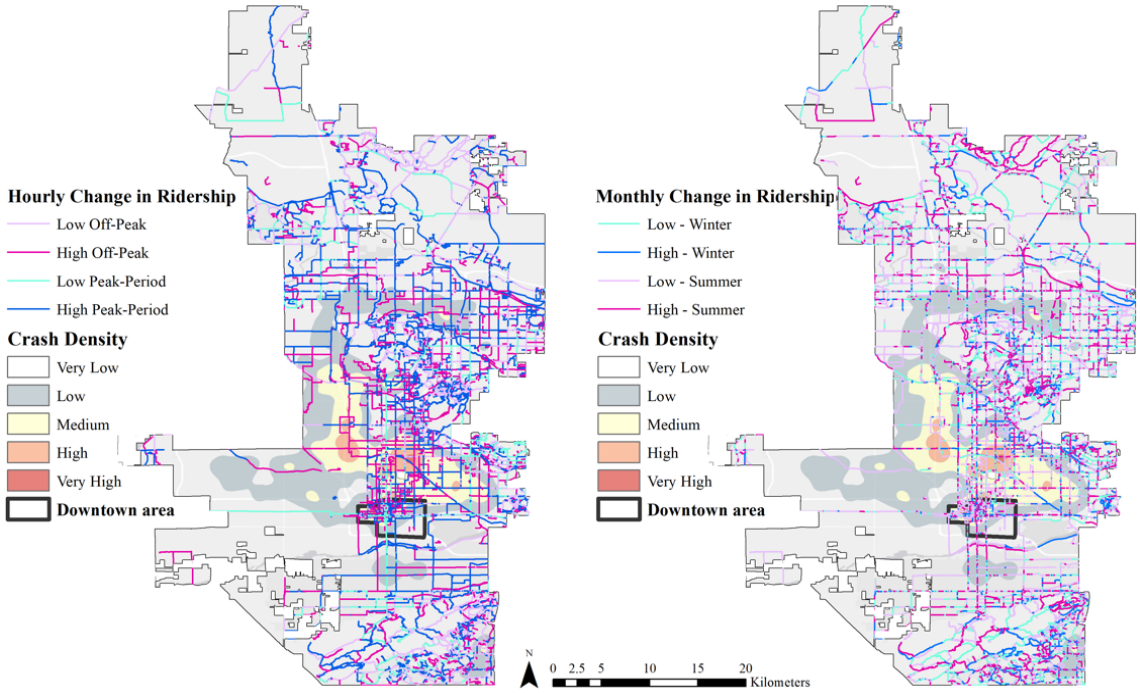


Figure 3.6: Maps showing different clusters of the change in hourly and monthly ridership between 2017 and 2018 along with bicyclist crash density.

On the monthly scale, more changes were observed in the summer months compared to the winter months (Figure 3.6). In Figure 3.6, average crash density is overlain on maps to demonstrate variations of ridership change in comparison with bicyclist exposure. The high crash density areas overlap with the high off-peak ridership changes at the hourly scale and high summer ridership changes at the monthly scale (Figure 3.6).

3.7 Discussion

The increase in popularity of health and fitness apps, such as Strava, has provided a novel source of high-resolution spatio-temporal big data. Strava data have been used to

examine where cyclists ride (Griffin and Jiao 2015), and several studies have examined the use of Strava data as a proxy for ridership volumes (Griffin and Jiao 2015, Jestico et al. 2016). Heesch and Langdon (2016) used heatmaps and counts of cyclists from Strava data to assess the impact of infrastructure change on bicycling behavior and Boss et al. (2018) use spatial autocorrelation techniques to monitor annual changes in spatial patterns of ridership. In the current study, we advance change detection approaches by operationalizing a generalized functional data analysis approach using Strava data to detect temporal changes in city-wide ridership patterns across hourly and monthly scales.

There is a growing interest among researchers in applying functional data analysis methods to study spatial-temporal processes across scale in multiple fields including ecology (Embling et al. 2012), environmental monitoring (Lee et al. 2015), and climate science (Ballari et al. 2018). Our results show that the use of a functional data analysis approach can help detect changes in fine-grained spatial-temporal data across multiple time periods.

The study also highlights the use of temporal alignment before detecting changes to account for elastic variations in the functional data (Srivastava et al. 2011). The SRVF technique used in this study can overcome the challenges faced by other methods such as wavelet analysis (Antoniadis et al. 2013) which have not considered real-world scenarios and alignment issues.

We mapped the changes in bicycle ridership for all street segments in the city of Phoenix post-alignment of the functional curves into 4 categories for both hourly and monthly scales (Table 3.2). Our results indicate that nearly 32% of the street segments in Phoenix show a high hourly change in ridership during the peak period between 8 am and

10 am (Table 3.2). There are also 6.6% of segments that account for high off peak-hour change in ridership (Table 3.2). These patterns indicated that improved infrastructure between 2017 and 2018 has led to a major increase in peak-period ridership as commuters feel safer riding their bikes to work. These results are consistent with previous studies (Akar et al. 2009) which show that bicyclists tend to ride more in areas with a high density of bicycling infrastructure as they feel safe biking and have a higher sense of comfort (Teixeira et al. 2020) bicycling in these areas.

The average number of bicyclists varies from 82 to 105 during the high and low peak periods (Table 3.3) and from 4 to 42 bicyclists during off-peak hours (Table 3.3). The changes during peak-period hourly ridership occur mostly around downtown Phoenix. The reason being commuters use bike lanes and bike paths for their regular commutes around this area the most. The high change areas also overlap with regions of high crash density as more incidents occur owing to exposure to a high volume of motorized traffic interspersed with bicycle trips during peak periods in these areas which is consistent with the results of the study by Fournier et al. (2019) and Saha et al. (2018) which highlight that traffic volumes have a positive correlation with bicycle crashes.

The exposure for high change during peak periods is 0.60 and for off-peak hours is 0.51 (Table 3.3) indicating a sharp increase in hourly ridership with lack of suitable infrastructure might lead to more crashes and affect bicyclist safety. Previous studies (Pucher and Buhler 2016, Vanparijs et al. 2020) have shown that North-American cities like Portland, Washington DC, and New York have already improved bicycling safety and increasing bicycling levels by greatly expanding their bicycling infrastructure. Therefore,

to reduce exposure in areas with a high change in ridership authorities need to provide more bicycling infrastructure.

On the monthly scale, most of the changes that occur during the winter months which consist of 37.6% of the street segments (Table 3.3) (including high and low changes in winter) located mostly on the outskirts of the city with recreational riders making more trips along trails and parks (Figure 3.6). However, the changes in summer across the street segments in and around the downtown area (Figure 3.6) are comparatively low as those areas are mostly used by commuters that have ridership patterns that are not as impacted by weather, a trend that is consistent with previous studies (Brandenburg et al. 2007, Miranda-Moreno et al. 2011). The remainder of the streets which are used for recreational trips experienced a sharp dip in commutes owing to high temperatures in the summer season (Brandenburg et al., 2007). The high and low changes during summer overlap with high crash density areas as the crashes occur more frequently in and around the high traffic zones specifically near the city center.

Surprisingly, 42.5% of street segments with low change during summer months (Table 3.3) have a high exposure of 0.60 (Table 3.3) indicating that the overall risk of crashes in streets with bicycle commuters in and around the city center is typically high (Loidl et al. 2016) throughout the year. Hence, local authorities should invest more in introducing bicycle-friendly infrastructure that reduces exposure in those areas even with a low change in monthly ridership.

The change maps shown in Figure 3.6 are categorized based on the continuous temporal changes derived from our functional change (Table 3.3) technique that captures fine-grained changes during all hours of a day and each month of the year. Our results also

highlight the importance of considering multi-scale temporal changes in bicycling when infrastructure changes in a city.

Hourly changes can be useful to detect commute patterns (e.g., Heinen et al. 2011) throughout the day whereas monthly changes give a summary of seasonal ridership patterns (e.g., Jestico et al. 2016) in the city. Practitioners can use maps at both scales (Figure 3.6) to identify regions that need improved infrastructure for tackling daily bicycle traffic as well as make long-term plans for future interventions that assure bicyclist safety within a region to quantify the cumulative impact over time.

Previous studies that have evaluated the mapped change in ridership focused on the comparison of two discrete snapshots in time (Boss et al. 2018). Snapshot approaches remove much of the temporal detail that could be valuable for understanding more nuanced changes. As well, the snapshot selected for evaluating change is often subjective. Our study overcomes the challenge of possible information loss by discrete snapshot approaches by combining the changes at hourly and monthly scales from continuous time-series data.

Functional data analysis approaches data have been used for monitoring precipitation changes (Suhaila et al. 2011), watershed modeling (Bourbonnais et al. 2019) as well as traffic flow estimation (Guardiola et al. 2014; Wagner-Muns et al. 2017). Aue et al. (2009) developed a mathematical formulation for change detection from the mean function of functional data, however, the method was not tested using real-world data. The increasing availability of big data from urban sensing technologies such as Strava has enabled monitoring change from spatial-temporal processes such as bicycle ridership continuously across multiple scales utilizing the functional data analysis framework.

We have developed the technique as a way for policymakers to visually represent change through maps and identify the infrastructure needs of a city. Often planners and policymakers face challenges in extracting actionable insights from raw big data that can inform decision-making. Our method is a step forward towards easing the process of detecting changes from big data by policymakers using visual approaches using an FDA framework.

The results from this study would be a good starting point for planners to make informed decisions on investments for modifying existing infrastructure or installing new infrastructure such as paved bike lanes, adding a new bike path, increasing the width of lanes, reducing the number of motor vehicle lanes, etc. . Our methods will be an effective tool for planners to make such targeted decisions in a more nuanced fashion from a data-driven approach. Although our study demonstrates a specific case study using Strava data for monitoring changes in bicycle ridership in Phoenix, the framework described in this study can be used for detecting changes in continuous time-series data obtained from big spatial-temporal data while accounting effectively accounts for elastic variations. For example, our method can be used to model changes in mobility patterns during an extreme event such as a tornado, hurricanes, or floods. It can also be used for environmental monitoring of air quality indicators over time in a city, studying temperature trends owing to global climate change as well as study variations in movement patterns in ecology.

3.8 Conclusion

Big, crowdsourced data pose numerous challenges ranging from the extraction of actionable information (Yang et al. 2017) to temporal misalignment (Choi et al. 2018) and

bias on app usage (Roy et al. 2019). The need for more accurate and reliable understanding and predictions requires improvements to algorithms that can recognize data inaccuracies, sampling errors. Efficiently integrating big data from different spatial-temporal scales is critical for earth system sciences (Hu et al. 2018).

Our research opens a new avenue for using functional approaches to data preprocessing and analysis across multiple scales from big spatio-temporal data. Functional approaches help in identifying the latent spatial-temporal patterns, which cannot be observed directly, through a data-driven perspective. Inferring such pattern changes from a raw noisy stream of individual trips is a rather non-trivial task and an ongoing area of GIScience research. Developing generalized techniques as outlined in our study, to automatically detect pattern changes from individual-level longitudinal spatial-temporal data, is therefore critical to developing behavior models that are adaptive over time.

While using big spatio-temporal data it is essential to account for nonlinear warpings for proper alignment and co-registration of functional curves. Our method highlights the use of square-root velocity functions to overcome such challenges and detect changes in hourly and monthly scales from functional data. From a broader perspective, this paper contributes to debates in time geography based on the theoretical foundation on how time and space constitute social life from the scale of individuals (Hägerstrand 1985). Considering previous research (Kwan 2002, Miller 2005, Long and Nelson 2013, Kwan and Neutens 2014) that highlight the role of underlying time and scale issues in geography, this paper builds a framework for analyzing change from real-world data at fine-grained

scales and contributes to the field of urban analytics from a methodological perspective which can help policymakers.

CHAPTER 4

CLASSIFYING TRANSPORTATION MODES COMBINING MOVEMENT DATA AND GEOGRAPHIC CONTEXT

4.1 Abstract

The increasing availability of health monitoring devices and smartphones has created an opportunity for researchers to access high-resolution (spatial and temporal) mobility data for understanding travel behavior in cities. Although information from GPS data has been used in several studies to detect transportation modes, there is a research gap in understanding the role of geographic context in transportation mode detection. Integrating the geography in which mobility occurs, provides context clues that may allow models predicting transportation modes to be more generalizable. Our goals are first, to develop a data-driven framework for transportation mode detection using GPS mobility data along with geographic context, and second, to assess how model accuracy and generalizability vary upon adding geographic context. To this extent we extracted features from raw GPS mobility data (speed, altitude, turning angle, and net displacement) and integrated geographic context in the form of geographic covariates to classify active (walk/bike), public (subway, bus, train, rideshare), and private (car, taxi) transportation modes in three different Canadian cities - Montreal, St. Johns, and Vancouver. To assess the role of geographic context in model generalizability & accuracy, we compared results from Random Forests, Extreme Gradient Boost, Decision Trees, and Ada Boost classifiers. Our results indicate that the accuracy of the models improved up to 4.2% on adding geographic context with Random Forests

achieving the highest accuracy of 83.8% upon adding contextual variables. Among the contextual variables that contributed to transportation mode detection distance to subways, distance to bicycle infrastructure, distance to bus stops, and distance to open green spaces were the most significant. We also found that the generalizability (i.e how accurately the model predicts modes from new unseen data) reduces upon adding contextual variables versus using GPS data alone. Our study highlights how policymakers can combine GPS data with geographic context to predict transportation modes and assess the generalizability of models in different geographic settings.

4.2 Introduction

Understanding the modes of transportation people use to travel within cities is key to planning safer, healthier, and more inclusive (Boulange et al., 2017) environments. Detailed information about mobility patterns and transportation mode usage can help planners and policymakers in making targeted decisions about investing in safe and equitable infrastructure (Nelson et al., 2021; Roy et al., 2019). The growing availability of health monitoring devices and smartphones has facilitated the process of collecting high-resolution (spatial and temporal) mobility data for cities which can avoid problems associated with traditional methods (Forrest & Pearson, 2005; Murakami et al., 2004). Such 'big' mobility data provides an opportunity to get a deeper understanding of transportation mode choices (Feng & Timmermans, 2013) highlighting the city's travel behavior (Bohte & Maat, 2009; Chen et al., 2016) and the need for improved infrastructure in terms of better accessibility (Ford et al., 2015; Cui et al., 2020) and comfort (Ferster et al., 2021) of its residents.

Information from mobility data has been used in transportation research (Zheng et al. 2008; Auld et al. 2009; Schuessler and Axhausen 2009; Stenneth et al. 2011; Hemminki et al., 2013) for understanding travel behavior by predicting modes of transportation from GPS features alone (Feng and Timmermans 2013; Carlson et al. 2015). However, there is a research gap in mode detection literature to comprehensively incorporate measures of geographic context and assess how the inclusion of such measures improves prediction accuracy. By extension, it is unclear if the inclusion of such measures in the mode detection process leads to more accurate predictions at the expense of model generalizability. It is therefore essential to understand whether geographic factors like the built and natural environment as well as land-use types of individuals could influence travel mode choices people make (Wang et al., 2017; Ewing and Cervero 2010).

While the geographic context may provide clues on modes of travel, geographic data has been used less frequently as a feature for classifying GPS data into travel modes. Traditionally, the context has been gathered using data from surveys and questionnaires (Van Vugt et al. 1996, Rodriguez and Joo 2004, Schwanen and Mokhtarian 2005, Wener and Evans 2007). In terms of existing methodologies for mode detection, traditionally, rule-based classifiers have been used more often. Although rule-based classifiers are known to have relatively rigorous boundaries on a relatively small number of features (Bohte and Maat, 2009, Chen et al., 2010, Gong et al., 2012, Sauerländer-Biebl et al., 2017, Schuessler and Axhausen, 2009, Stopher et al., 2008, Marra et al., 2019). Tree-based classifiers are less restrictive and hence more of an improvement over rule-based classifiers. Previous research (Wang et al., 2017; Cheng et al., 2019; Kim et al., 2021) has shown tree-based algorithms achieve higher accuracy for transportation mode

classification purposes. Tree-based classifiers like Random Forests (Wang et al., 2017; Cheng et al., 2019; Nguen & Armoogum, 2020), Gradient Boosting (Wang et al., 2018) as well as Decision Trees (Shah et al., 2014; Feng & Timmermans, 2016) have been used in previous studies and have proven to be the more appropriate algorithmic approach for mode detection while using GIS information.

Random Forests have been found to have high precision and recall accuracy (Stenneth et al., 2011) in classifying motorized and non-motorized transportation modes. More specifically, in situations where there are a huge number of features that might affect the classification of different modes, decision trees (Reddy et al., 2010), and random forests (Ellis et al., 2014, Mäenpää et al., 2017) are the most popular choices. They have the advantage of performing quite well while being relatively easy to implement. Extreme Gradient Boost algorithms, an ensemble of decision trees that learns by fitting negative gradients, are an improvement over decision trees as they (Friedman, 2001) have shown considerable success in a wide range of practical applications and could be used to avoid overfitting. Ada Boost has gained much popularity among researchers as it can achieve higher prediction accuracy with a relatively lesser number of iterations compared to Gradient Boost or Decision Trees (Wyner et al., 2017) and simultaneously continues to maintain low generalizability errors (Schapire et al., 1998) as they work employing interpolation and uses a self-averaging property (Wyner et al., 2017).

Although a number of methods exist for classifying transportation modes from GPS data, most of the existing methods are limited in terms of assessing the role of geographic context on predictive accuracy and how they can translate into policies that could improve urban life. The inclusion of measures of geographic context in the mode detection process

may lead to more accurate predictions needed for effective policymaking, but we have yet to test this hypothesis. An important caveat to the inclusion of geographic context is that prediction improvement may possibly come at the expense of model generalizability. Hence, we also need to know whether approaches that incorporate geographic context lead to overfitting that limits their use across urban areas. Generalizability is impacted both by the cost and effort of obtaining diverse datasets for training models for transportation mode detection. Some studies (Dexter et al., 2020) have shown how generalizability can be assessed in medical applications. However, there is little methodological knowledge on the causes of weak generalizability in the transportation mode detection paradigm as well as the value of leveraging varied geographic covariates from multiple sources for better generalizability of such models. Better assessment of weak generalizability combining datasets from multiple cities could improve our understanding of the generalization challenge.

To address these gaps, we have identified our research goals to examine whether combining GPS and contextual features can improve the prediction accuracy of transportation mode detection and to assess how generalizability varies when adding geographic context to transportation mode detection. To this extent, we first extract meaningful features from the GPS traces and combine these features with contextual information from geographic features guided by existing literature. Then, we train different supervised classification models to predict travel modes in three different Canadian cities and finally validate and assess the generalizability and accuracy of the trained models using just GPS features alone versus combining GPS and contextual features.

We aim to highlight the role contextual variables play in improving the prediction accuracy of transportation mode detection algorithms and if there is a trade-off between accuracy and generalizability. Our study can be used by practitioners as a guideline to choose appropriate contextual variables for accurately predicting transportation modes as well as testing the generalizability of prediction by combining those variables with new unseen mobility datasets.

4.3 Study Area

We performed the study across three Canadian cities: Montreal in Quebec, St. John's in Newfoundland, and Vancouver in British Columbia. Each of the cities used in the study has highlighted the role of equitable transportation infrastructure and therefore invested data collection at different times to understand travel mode choice using different GPS-based platforms. We chose these three cities owing to the availability of data with comparable spatial and temporal resolution to assess our hypothesis. In addition, the cities selected also have very different transportation systems, geographies, and population sizes which is an additional benefit for this study where we are assessing the role of geographic context on transportation mode detection. Table 4.1 shows the overall geographic setting of the study spanning the three Canadian cities. The weather conditions, population, and mode share of commuters for different sustainable and active transportation modes gathered from statistics Canada are listed in Table 4.1.

Montreal is the cultural and economic hub of the province, with the second largest population in Canada. It is a port city and is surrounded by St. Lawrence and Ottawa rivers. It is a walkable city and is interspersed with bike lanes and bike paths. The city is also well-

connected by different public transit modes like subways, buses, and trains connecting the city to the entire province.

Table 4.1: Description of the weather, population, and transportation mode share for each city

City	Annual Temperature		Population	Mode share of commuters			
	Min	Max		Walk	Bike	Public Transit	Carpool
Montreal	-4.7°C	19.0°C	4,247,446	5.2%	2.0%	22.3%	8.6%
St. John's	-3.8°C	16.4°C	108,860	4.6%	0.2%	3.1%	17.8%
Vancouver	3.1°C	17.9°C	2,463,431	6.7%	2.3%	20.4%	11.2%

*Source: Mode share was collected from data provided by Statistics Canada, Commuters using sustainable transportation, https://www12.statcan.gc.ca/census-recensement/2016/as-sa/98-200-x/2016029/98-200-x2016029-eng.cfm. The temperature data was provided by Environment Canada from highest and lowest temperatures averaged from 2013-2020 https://climate.weather.gc.ca/climate_data/almanac_selection_e.html

St. John's is a harbor city with a downtown of steep hills and winding streets. The City of St. John's maintains a road network of over 1,400 km, as well as a network of sidewalks for pedestrians and parking infrastructure throughout the city. The Metrobus transit is a popular public transit service in the city and alongside this, the city also maintains a road network of over 1,400 km, as well as a network of sidewalks for pedestrians and parking infrastructure throughout the city.

Vancouver boasts an accessible and convenient public transit system with several modes including bus, SkyTrain, ferries as well as bicycles. As the city is surrounded by water on three sides, it has several bridges to the north and south. Although similar to most other cities in that the automobile serves as the primary mode of transportation, it has

alternatives such as the SkyTrain system, which is the longest fully automated light metro system in the world, and an extensive network of bicycle paths.

Vancouver is much warmer than Montreal and St. John's which reflects higher use of active transportation modes including walking and cycling. Montreal has a well-connected transit network with nearly 22.3% of its commuters using public transit modes as their mode of choice. The overall population of Montreal is nearly 4.3 million with 38.1% (Table 1) of the population using either active, public, or shared mode of transit for commute purposes. On the other hand, St. John's is much smaller in terms of the total population and with much harsher climatic conditions which typically lead people to use private vehicles with only about 25.8% people (Table 1) availing active, public, or shared transit modes.

4.4 Data

We used GPS-enabled mobile applications to collect a total of 3,226,659 unique user-defined trips from Montreal, St John's, and Vancouver between August and December 2017. Data for St. Johns's and Vancouver were collected through a smartphone application Itinerum (Patterson et al. 2019) which collected GPS data at 1-minute temporal resolution.

The data for Montreal were collected using the MTL Trajet mobile application (MTL Trajet, 2017) which collected GPS trajectories of user movements from the origin and destinations by truncating to the nearest intersection. The data collection mechanism was similar to that in St. John's and Vancouver as the MTL Trajet uses the same underlying technology as Itinerum devices. All trips were for both St. John's and Vancouver were labeled by participants and for Montreal were inferred by the mobile app using a trip detection algorithm (MTL Trajet, City of Montreal,2020). The transport modes for each

trip were labeled by the GPS platforms into several different travel modes (i.e. bike, walk, subway, tram, carpool, bus, car, taxi) from all three cities (Table 4.2). To be consistent across all three cities we manually reclassified the transportation modes into three primary modes – active (walk/bike), private (car/taxi), and public (bus/train/subways). Table 4.2 summarizes the characteristics of the GPS trajectories obtained from all 3 cities and the percentage of trips grouped into different transportation modes. The total number of trips includes all trips made in the entire city. The percentages include what percent of each trip mode were categorized and labeled as 'Active', 'Public', and 'Private' transportation in the original dataset.

Table 4.2: Trip characteristics collected from GPS devices for multiple cities

City	Total no. of trips	Trips by Mode (%)		
		Active (Walk/Bicycle)	Public (Bus/Subway/Carpool)	Private (Car/Taxi)
Montreal	3,226,147	29.8	12.1	58.2
St. John's	12,861	59.7	11.8	28.5
Vancouver	5,732	74.1	10.6	15.3

Active modes comprised 29.6% of the total trips, followed by public transit modes which consisted of 11.9% of the trips and the rest 58.5% of the trips were private modes. In order to account for a comparable temporal and spatial resolution, we were limited in terms of the balanced number of trips collected by the different GPS platforms in each city.

4.5 Methods

We used a multi-step approach for mode detection from raw mobility datasets. As a preliminary step, we first preprocessed the data by eliminating noisy data points and those that are not classified as trips by using a trip detection algorithm. First, we extracted

meaningful features from the raw GPS data and then collected contextual information from the spatial surroundings of the GPS trajectories using GIS tools. Second, we generated two sets of input feature matrices to train the classification models- one using just the GPS features alone and the other combining both the GPS and GIS features. Third, we split the input feature matrices into training and test sets using an 80:20 split. Fourth, we trained different supervised classifiers based on existing literature to predict transportation modes using those feature set combinations using the training data and determined their cross-validation accuracy scores. Finally, each of the classifiers was tested with the remaining test data to determine their classification accuracies (using precision, recall, AUC, and F1 scores) and their generalizability scores (using G-score). A detailed description of each step is listed in the subsections that follow. An overall workflow is shown in Appendix A.

We used programming languages R 3.4 to extract the GPS features and Python 3.6.3 for developing the travel mode classification framework and used the ArcGIS 10.3 geographic information systems suite for extracting the GIS covariates for our analysis. The classification algorithms were built in Python using the scikit-learn library (Pedregosa et al., 2011), and feature selection and manipulation were performed using the pandas library (McKinney, 2011).

4.5.1 Extracting trip features from movement data and constructing measures of geographic context

The GPS records in all datasets were recorded as latitude and longitude and were converted to UTM (Universal Transverse Mercator) coordinates (easting, northing) using pyproj 1.9.5.1. Features were calculated for all GPS records available throughout each

study for each participant. The primary unit of analysis is the participants' GPS trajectories over the entire study period converted into trips. Every second, the GPS device registered the position coordinates (i.e., latitude, longitude, and elevation) of a participant, which was converted into trajectories using the R package "adehabitatLT" (Calenge, 2015). From the trajectories, we extracted the mean of the speed, net displacement, altitude, relative and absolute turning angles for each user along a single trip. These features (Table 4.3) were combined into a single feature set which we identify as 'GPS' in the following subsections.

To extract contextual information about the surrounding environment through which individuals move, we extracted proximity measures as Euclidean distances to the nearest points of interest around a GPS trajectory of each user.

Table 4.3: List of features extracted from raw mobility data captured by GPS platforms

Features	Type	Operationalization	Relevance	References
Speed	GPS	Speed calculated from the consecutive points of the trajectory	Variability in speed can highlight the difference between motorized and non-motorized transport modes.	Stenneth et al., (2011); Zheng et al. (2010); Bohte & Maat (2009), Reddy et al. (2010); Shen & Stopher (2014); Xiao et al.(2015); Roy et al (2020)
Altitude	GPS	The average altitude throughout the trip	The height can indicate whether the user travels in underground subways versus on foot or larger vehicles like buses etc.	Wang et al. (2017); Feng and Timmermans (2013); Roy et al. (2020)
Displacement	GPS	The net displacement between consecutive locations along the trajectory	The net displacement can distinguish between motorized and non-motorized transport modes with longer trips taken on public or private modes versus shorter ones are made using active	Zheng et al. (2010); Xiao et al (2015); Feng and Timmermans (2013); Roy et al. (2020)

			modes.	
Turning Angle	GPS	The relative and absolute turning angles of between consecutive points of a trajectory	The orientation can help distinguish a motorized vehicle that can only drive on roads and may not usually turn or change to a new lane unless necessary	Wang et al. (2017); Roy et al. (2020)

The points of interest were extracted using a data mining approach from Overpass API using Python package "geopandas_osm", which allows choosing all points of interest (POIs) around a user's location which were then imported into ArcGIS to calculate Euclidean distances in kilometers using the NEAR functionality. The POIs were categorized into land-use types such as residential areas, commercial areas, green spaces, and transportation hubs like bus stops, subway stations, bike lanes, and topographic characteristics like distance to the shoreline and comfort level of streets. Attributes like speed (Stenneth et al., 2011; Zheng et al., 2010; Bohte & Maat,2009, Reddy et al., 2010; Shen & Stopher, 2014; Xiao et al., 2015), acceleration (Stenneth et al., 2011; Roy et al., 2020), proximity to bus stops (Gong et al., 2012; Ngyuen & Armoogum, 2020) and rail lines (Stenneth et al., 2011) have been used several times in previous studies, however, proximity to different land-use types and infrastructure specific to active modes of transportation within the context of mode detection have been newly introduced in this research. Additional temporal features like time of day, week of the day, and month of the year were also included to capture the temporal context that influenced various transportation modes. We refer to the contextual variables shown in Table 4.4 as a separate feature set and call it 'GIS'.

Table 4.4: List of features extracted from the geographic context

Features	Type	Operationalization	Relevance	References
Distance to open space	GIS	Mean Euclidean distance to the nearest open space or green space from the points along the trip trajectory	People using active modes tend to use less traffic-prone areas and closer to open green areas like parks etc.	Roy et al. (2019); Semanjski et al. (2017); Böcker et al. (2015)
Distance to residential areas	GIS	Mean Euclidean distance to the nearest residential area from the points along the trip trajectory	Trips that have a longer duration and are closer to residential areas can be made using Public/Private modes.	Roy et al. (2019); Semanjski et al. (2017)
Distance to commercial centers	GIS	Mean Euclidean distance to the nearest commercial area from the points along the trip trajectory	Typical short trips in and around commercial areas can be used to distinguish public transit modes	Roy et al. (2019); Semanjski et al. (2017)
Distance to subway stations	GIS	Mean Euclidean distance to the nearest subway station from the points along the trip trajectory	Trajectories that are closer to subway stations can be used to identify public transport modes	Gong et al. (2012); Stenneth et al. (2011)
Distance to bus stops	GIS	Mean Euclidean distance to the nearest bus stop from the points along the trip trajectory	Trajectories that are closer to bus stops can be used to identify public transport modes	Gong et al. (2012); Stenneth et al. (2011); Ngyuen & Armoogum (2020)
Distance to bike infrastructure	GIS	Mean Euclidean distance to the nearest bike lane/bikeway/bike path from the points along the trip trajectory	Trajectories that are closer to bike infrastructure can be used to identify active transport modes	Roy et al (2019); Jestico et al. (2016); Semanjski et al. (2017)
Distance to shoreline	GIS	Mean Euclidean distance to the nearest shoreline from the points along the trip trajectory	Trajectories that are closer to the shoreline and have lower speeds can be used to identify active transport modes like biking or walking with an additional aspect of scenic effect for the comfort of pedestrians/bicyclists. It could also highlight	Nelson et al. (2021)

			the use of private modes if the trips tend to have higher speeds.	
Time of day	GPS	Morning peak (4 am – 10 am), afternoon (11 am – 4 pm), evening peak (5 pm – 9 pm), and night (10 pm – 3 am)	Commute trips are typically made during peak hours of the day & when combined with speed and height can be used to classify three different modes.	Jestico et al. (2016)
Season of year	GPS	Spring, Fall, Winter, Summer seasons for each city	More outdoor activities especially using active modes are carried out during pleasant weather conditions and they can be used to distinguish between motorized and non-motorized modes.	Jestico et al. (2016); Böcker et al. (2015)
Day of week	GPS	Day of the week when the trip was made	Recreational trips are made on weekends versus weekdays and using speed and altitude/orientation can be used to classify different modes	Böcker et al. (2015)
Comfort level of streets	GIS	Street comfort level classification for bicyclists	Active modes are associated with the comfort level of streets and can be separated from public/private transport modes	Ferster et al. (2021)

4.5.2 Training supervised classifiers to predict transportation modes from extracted features

Both Tables 4.3 and 4.4 highlights all the features extracted from the raw GPS data and contextual information surrounding the trajectories of various trips. All features were normalized using a min-max function and used as inputs to supervised classification algorithms. We constructed two different feature sets to train the supervised classifiers – one with just the 'GPS' features capturing the raw mobility metrics and the second with 'GPS + GIS' features capturing the geographic context along with the mobility metrics.

We used four supervised classification algorithms – Random Forests (RF), Extreme Gradient Boost (XGB), Ada Boost (ADA), and Decision Trees (CART) to classify the transportation modes. We used 80% data for training the classifiers with 10-fold cross-validation and held out the remaining 20% data for testing their predictive accuracy. We fitted both the feature set combinations for each classifier. First, we trained the classifiers with only GPS features, and then combined both GPS and GIS features to retrain the classifiers for predicting the labels – 'Active', 'Public' and 'Private' for three different transportation modes. To account for the imbalance in trip distribution across all three modes we used a resampling technique called Synthetic Minority Sampling Technique (SMOTE) (Chawla et al. 2002).

4.5.3 Comparing accuracy across and assessing generalizability

We assessed the classification accuracy of the classifiers first using just the GPS features and then again after combining the GIS features. A cross-validation approach was used to evaluate the accuracy of the classifiers in the training phase. The 10-fold cross-validation split the 80% training data into 10 subsamples, and in each validation step, the classifiers were trained with 9 subsamples and predicted using the remaining one subsample. Each fold generated an accuracy score for the classifiers and finally, a mean cross-validation accuracy score was reported from all 10 folds.

Once the classifiers were trained we used the 20% test data, which the classifiers were not trained with, to predict the transportation modes. These predicted modes were used to calculate classification metrics – precision, recall, AUC score, and F1-score. Precision (Equation 1) is a measure of the relevance of the results while recall (Equation 2) is a

measure of how many truly relevant results are returned by the models. A high precision score signifies low false-positive rates, and a high recall indicates low false-negative rates. The FI-score (Equation 3) is the harmonic mean of the precision and recall rates which measure the classification accuracy of the models based on true and predicted labels.

The overall predictive accuracy of all models was summarized using an Area-under-the-curve (AUC) score (Hand and Till, 2001) that was calculated using equation (4), where *k* denotes the total number of classes (k=3; active/public/private), *tp* stands for true positives, *tn* for true negatives, *fp* for false positives and *fn* for false negatives. The average across all classes gives the final AUC score for each model which determined the predictive accuracy of the models across all three cities.

$$P = \frac{T_p}{T_p + F_p} \quad (1)$$

$$R = \frac{T_p}{T_p + F_n} \quad (2)$$

$$F1 = 2 * \frac{P * R}{P + R} \quad (3)$$

$$AUC = \frac{\sum_{i=1}^{k} \frac{tp_i + tn_i}{tp_i + tn_i + fp_i + fn_i}}{k} \quad (4)$$

In addition to the classification accuracy, we also estimated the overall generalizability of all four classifiers. We calculated a G-score for all the model and feature set combinations to ascertain the generalizability of the models when used to predict modes from unseen data across completely new study areas. The G-score can be thought of as an L1 norm of the F1 scores for training and test sets on each classifier.

$$G = \|F1_{train} - F1_{test}\| \quad (5)$$

The G-score (Equation 5) is calculated as an absolute difference between the F1-scores of the train and test datasets on the same classifier. The approach is somewhat similar to the one first introduced by Barbiero et al. (Barbiero et al., 2020) who use the concept of a convex hull to assume training data points fall within the convex hull and test data points outside of it. The more generalizable the decision boundary of the classifier is, the lesser the deviation between the train and test F1 scores will be. Hence, the lower the G-score in our case, the more generalizable the classifier is based on the input features. The low G-score assures that the classifier can accurately predict transportation modes from new unseen datasets. Whereas, if the G-score is high it indicates the classifier will fall prey to a higher error when predicting transportation modes from trips that were not used to train the classifier. However, there may be a trade-off between high classification accuracy and high generalizability.

Based on the ultimate purpose of classifying trips, practitioners may either choose a highly accurate model or a highly generalizable model. The highly accurate model may produce correct transportation mode labels but would depend on a greater number of available training data and would perform well in a single study area. The highly generalizable model compromises a little on very high accuracy but will ensure the model will perform optimally well in multiple study areas with varying geographic context and will not be entirely skewed towards any single city or a high amount of readily available correctly labeled trips.

4.6 Results

The correlation among the variables listed in Tables 4.3 and 4.4 is shown in Figure 4.1 which highlights the Pearson's correlation coefficient among all independent features used to fit the classifiers. Most of them have a Pearson's correlation coefficient below 0.3 and above, 0.1 indicating none of these features suffer from multicollinearity and were used as input features to the classifiers listed in Table 4.4. All variables in Figure 4.1 were used to prepare two different feature sets – only GPS features comprised speed, distance, turning angle, and height whereas the distance variables (Figure 4.1) highlighted the features used to account for geographic context. Distance to residential areas, commercial areas, and open spaces were found relevant to account for trips closer to home or for running errands whereas the distance to bike infrastructure, subway stations, and bus stops accounted for access to public transit modes and active transportation mode choices.

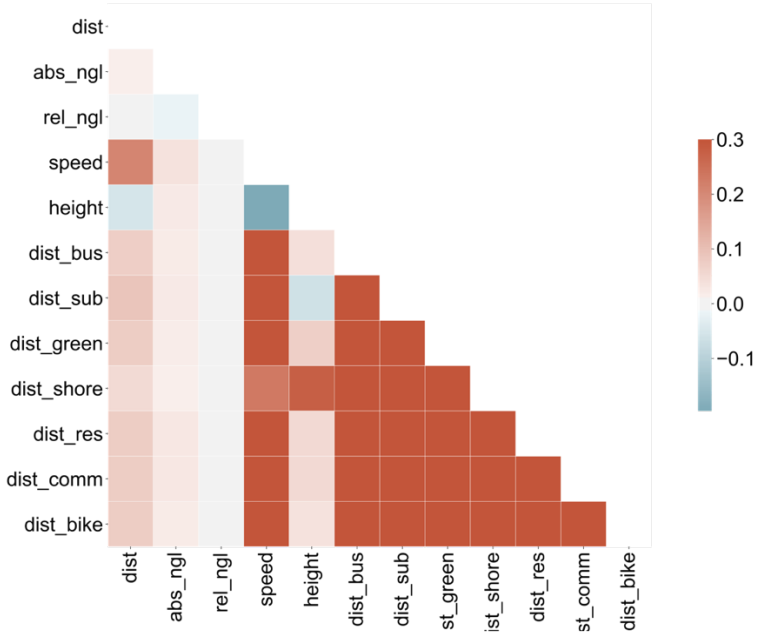


Figure 4.1: Correlation matrix of the numeric variables derived from GPS data and geographic context

We grouped the trips and mode used to extract the mean and variance of the variables (Figure 4.1) to train four different supervised classifiers with each of the two feature set combinations with a total of 28 features. Table 4.5 shows the accuracy of 10-fold cross-validation for all six models using the mean and standard deviation of the cross-validation accuracy metric calculated from 80% training data. All four classifiers typically have a higher accuracy when both GPS and GIS features were combined to train the models. Among these Random Forests have the highest mean cross-validation accuracy of 83.8 % using all features gathered from Montreal, St. John's, and Vancouver. We must remember that based on separate geographic settings, the model accuracy is assumed to a generalizable across all three cities.

Table 4.5: Model Accuracy of different supervised classifiers fitted to raw GPS and GIS data

Model hyperparameters: 'k' = 10-fold; balancing = 'SMOTE'; max depth = 5, n-trees = 50					
Model	**Feature Set**	**Cross-validation Accuracy**			**Change in max accuracy upon adding geographic context**
		Mean	**S.D.**	**Max**	
Random Forest	GPS	0.783	0.016	0.796	+ 0.042
	GPS + GIS	0.796	0.022	0.838	
Extreme Gradient Boost	GPS	0.773	0.016	0.796	+ 0.037
	GPS + GIS	0.783	0.029	0.833	
Decision Trees	GPS	0.684	0.015	0.709	+ 0.035
	GPS + GIS	0.695	0.048	0.744	
Ada Boost	GPS	0.749	0.019	0.778	+ 0.034
	GPS + GIS	0.751	0.026	0.812	

*The accuracy metrics are derived from training the models with 80% of the data (n = 2,138,273) with 10-fold cross-validation.

Overall, the results indicated that the accuracy of all four models improved from 3.4% up to 4.2% after GIS features were added to the models along with the GPS features (Table 5). The highest improvement in accuracy was achieved by Random Forests with an overall increase of 4.2% in maximum accuracy. Although a small increase, this confirms our hypothesis that contextual variables do matter in improving the accuracy of mode detection models. The variables used in our study indicate that they can contribute to more accurate prediction of active, private, and public modes of transportation compared to using mobility metrics.

Our results also indicated that among the models with GPS and GIS features combined - Random Forests achieved the highest accuracy among all four models (Table 4.5) with a maximum accuracy of 83.8% and a mean accuracy of 79.6% when both GIS and GPS features were combined. Decision trees had the lowest maximum accuracy of 74.4% and lowest mean accuracy of 69.5%. Although Extreme Gradient Boost (mean = 78.3%, max = 83.3%)) and Ada Boost (mean = 75.1%, max = 81.2%) had high accuracies (Table 4.5), the variability of the results were higher compared to Random Forests which had the lowest deviation from the mean accuracy (S.D = 0.022) across all 10-folds of cross-validation (Table 4.5) compared to Extreme Gradient Boost (S.D. = 0.029) (Table 4.5) and Ada Boost (S.D. = 0.026) (Table 4.5). The low variability in cross-validation accuracy of Random Forests indicates less overfitting compared to the other classifiers.

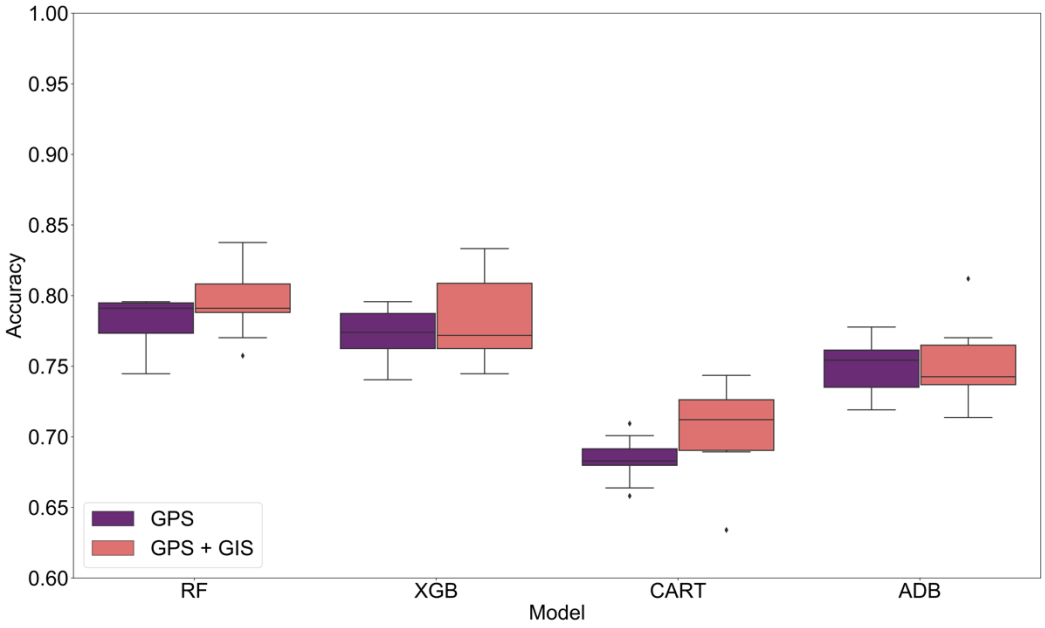


Figure 4.2: Boxplots showing variability in the predictive accuracy of different supervised classifiers using 10-fold cross-validation

Although the mean accuracy is reported in Table 4.5, the models do reach higher accuracies in some of the folds during the cross-validation process. The variability of the accuracy of each fold of cross-validation is shown in Figure 4.2 to highlight the uncertainty in prediction accuracy based on the training set. The model accuracy is highest for all four models when the mobility metrics from GPS data and geographic context from GIS data are combined to train the models. However, the Random Forest model reaches the highest accuracy in both cases. In comparison to Random Forests, the Extreme Gradient Boost is the second-best fit but has lesser variability in model accuracy (Figure 4.2). For ease of usage, we choose the Random Forest model as it proves to have the highest mean accuracy among all models.

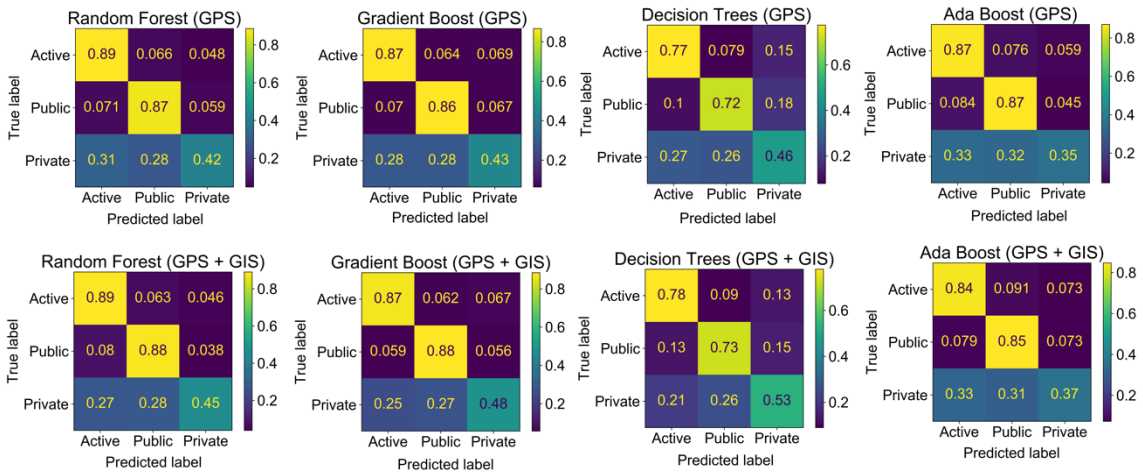


Figure 4.3: Confusion matrix for all eight models combining GPS features and geographic context

We used the trained classifiers to generate confusion matrices using test data as shown in Figure 4.4. The highest accuracy is achieved when both GPS and contextual data are combined. 89% of trips were accurately classified as active modes using Random Forests (Figure 4.4), 87% were classified as public modes and 45% were accurately classified as private modes. Other than ADA Boost all models showed an improvement in mode-specific accuracies when contextual features were added to the model. Decision Trees showed the highest accuracy for private (car/taxi) modes of transportation whereas Random Forests classified the active (bike/walk) and public (subways/buses/trains) modes most accurately.

Based upon the inclusion of geographic context, the variables that accounted for proximity to seashores, bus stops, green spaces, and subways proved useful to achieve higher precision for active and public modes of transportation.

Table 4.6: Comparison of model accuracy, classification metrics, and model generalizability

Model	Features	AUC	Precision	Recall	F1-train	F1-test	G-score	Change in G-score upon adding geographic context
Random Forest	GPS	0.905	0.752	0.742	0.760	0.753	0.007	-0.025
	GPS + GIS	0.898	0.762	0.747	0.780	0.710	0.032	
Extreme Gradient Boost	GPS	0.892	0.728	0.721	0.756	0.751	0.070	+0.011
	GPS + GIS	0.894	0.753	0.744	0.781	0.724	0.059	
Decision Trees	GPS	0.773	0.646	0.652	0.684	0.652	0.005	-0.060
	GPS + GIS	0.766	0.677	0.679	0.706	0.646	0.065	
Ada Boost	GPS	0.843	0.713	0.697	0.734	0.668	0.057	+0.022
	GPS + GIS	0.842	0.689	0.683	0.723	0.688	0.035	

However, there was a high misclassification between public and private modes as more contextual variables like speed limits, the number of lanes, and traffic volume are needed to further classify the mode choices between these two.

Given the cross-validation accuracy of each model in Table 4.5, we also tested the AUC scores to determine the classification accuracy and G-scores to determine the generalizability of these models in Table 4.6. The Precision and Recall values for all models are higher when GIS features are combined with GPS features (Table 4.6) indicating the models can classify the different transportation modes into active, private, and public more accurately from unknown data compared to when the models use just GPS features.

For Random Forests, precision is higher when contextual variables are added so actual modes (active/private/public) are predicted correctly in 75.2% cases (Table 4.6)

when just GPS features are used versus 76.2% cases when both GPS and GIS features are used. Similarly, for Extreme Gradient Boost (GPS: 89.2%, GPS + GIS: 89.4%), Decision Trees (GPS: 77.3%, GPS+GIS: 76.6%). However, for the AdaBoost model the precision deteriorates (GPS: 71.3%, GPS + GIS: 68.9%) when contextual variables are added to it (Table 6). A similar pattern is observed for recall values, which indicates the percentage of times on an average a trip classified as active/private/public among all trips, with Random Forest having the highest recall (GPS: 74.2%, GPS +GIS: 74.7%) and AdaBoost having the lowest recall which decreases with the addition of contextual variables (GPS: 69.7%, GPS + GIS: 68.3%).

In terms of the generalizability of the models, the G-scores are typically higher for models when GIS features are combined with GPS features (Table 4.6) than just using GPS features alone. This means that as the models predict modes from unseen trip data they tend to show a reduction in the overall accuracy of mode classification when contextual variables were added. If we consider training points to be within a convex hull whose boundary is defined by the classification model, then those models whose difference in distance between train & points from the boundary are lower (lower G-score shown in Table 4.6) are more generalizable versus those whose distance is greater. That would be the case when a model overfits the addition of contextual variables and are more tied to the geographic setting of the city it is trained with and becomes less applicable to other cities.

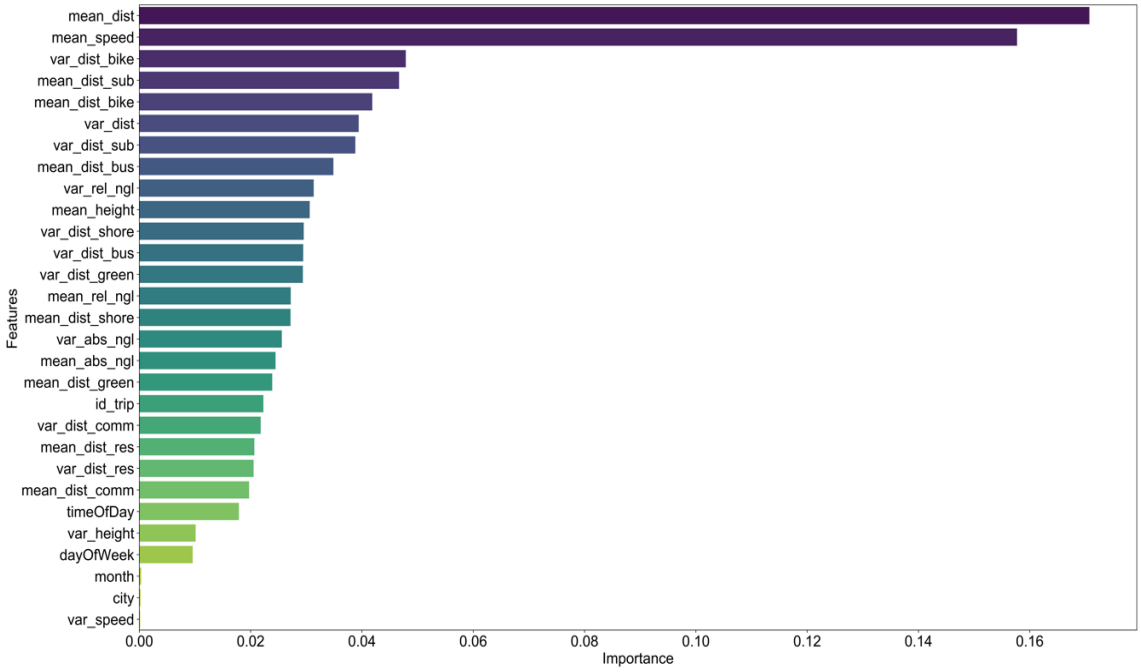


Figure 4.4: A bar plot showing the variable importance of GPS and GIS features used in the Random Forest model

The Ada Boost model is an outlier as its G-score lowers further when contextual variables (0.035) are added to it compared to just GPS features (0.057). The decrease in G-score for AdaBoost highlights its potential for higher generalizability when using geographic context along with GPS features to predict transportation modes from unseen trips in other cities, the trade-off being its lower accuracy compared to Random Forests, Extreme Gradient Boost, or Decision Trees. These three models are however more generalizable when trained with just GPS features alone as the geographic context does seem to overfit them making these three models context-specific to some extent.

We identified the feature importance using the Random Forest model (Figure 4.4) and found that the top five most important features were – average speed, average distance, the variance of distance to bike infrastructure, the average distance to subway

stations, and average distance to bike infrastructure. These variables can be used to build similar models across multiple other cities based on the availability of data.

4.7 Discussion

We developed a data-driven framework to classify transportation modes from raw GPS data by combining additional geographic context across multiple cities. Our results identified the key variables (distance to bike infrastructure, distance to subways, distance to shoreline, distance to open spaces) necessary to summarize geographic context across multiple cities which can be used to classify transport modes using a similar methodology in several different cities based upon the availability of data. We used Euclidean distances to subway stations, bus stops, and bike lanes that can determine a particular transport mode choice for commuters (Lunke, 2020) who might want to optimize their commute times by staying closer to areas with better availability of transportation infrastructure. The natural and built environment of a city also plays a key role in the transport mode choice (Winters et al., 2010) of its residents as earlier studies have found people using bicycles tend to ride near residential areas and (Roy et al., 2019; Semanjski et al., 2017), running an errand in an around city centers in either car or on bicycles can also be captured by distance to commercial areas (Semanjski et al., 2017) and distance to green spaces or seashores (Semanjski et al., 2017; Böcker et al., 2015) are often important for leisure trips made on foot, bicycles or in private vehicles who can spare time to interact more with their spatial surroundings (Páez and Whalen, 2010) as well as feel safe and comfortable (Ferster et al., 2021). Aesthetics like the visibility of the sea-shore (Nelson et al, 2021) or proximity to open green spaces (Roy et al., 2019) that have lower traffic volumes and safer speed limits

(Roy et al., 2019) could be a determining factor for people who use active modes of transportation.

Time of day, day of the week, and season were also considered in terms of capturing temporal context which has proven effective in identifying active modes of transportation (Jestico et al., 2016). The time of day and day of the week was more significant among other variables used in our models (Figure 5). Possible explanations for the relevance of temporal context could be related to the time of day barriers to active transportation such as reduced daylight, colder nighttime temperatures (Böcker et al, 2016), and concerns about safety (Dill and Carr, 2003; Aziz et al., 2018) which may lead to more selection of private modes of travel at night compared to public or active modes during the day. Harsh weather conditions based on different months can also be a deterrent for active modes (Böcker et al., 2016) like walking or biking but could be useful in classifying trips when people might prefer to use public transit modes.

The raw GPS data were converted into meaningful features that were combined with the contextual variables to fit supervised classification models and the accuracy of each model improved after adding contextual variables. Such improvement in accuracy establishes the importance of considering contextual variables in determining transportation modes in addition to mobility patterns derived from GPS data. The Random Forest model achieved the highest accuracy of 83.8% with the least variability across 10 levels of cross-validation (Figure 3), which is in alignment with previous research (Wang et al., 2017; Cheng et al., 2019; Kim et al., 2021), out as the best model with high recall, high precision as well as a high AUC score but the lower generalizability score while using

just GPS features is a trade-off that highlights accuracy and generalizability may not go hand in hand in certain scenarios.

Since GPS features are not tied to the local geographic setting, using just those features alone with any geographic covariates might seem to make the models more robust. Random Forests achieve the lowest G-score of 0.005 (Table 5) making them most robust among all other models followed by Gradient Boost with a G-score of 0.007 (Table 5) using just GPS features alone. However, when the predictive accuracy of Random Forests need is prioritized, we see that the highest AUC score of 0.898 (Table 5) with comparably high precision and recall of 0.762 and 0.747 help to classify trips more accurately (Table 5) as active, public or private modes using both GPS features and geographic covariates. AdaBoost, however, is more generalizable as it lowers the G-score without compromising much on the AUC scores- which assures high classification accuracy in different geographic settings (Table 4.5).

Our results indicated that although the cross-validation prediction accuracy of Random Forests is the highest when both GPS and GIS features are combined, the classification accuracy deteriorates when GIS features are added. This could mean adding contextual variables lowered the individual class-specific accuracy as there was a higher probability of misclassification between active and private modes as can be seen in the confusion matrix (Figure 4). Some of the contextual variables (i.e. distance to residential areas or commercial areas) that are used to train the model might not be sufficient to separate active modes from private modes as people tend to use their cars for running errands and might not do it on foot or a bike.

The accuracy and generalizability of the models showed a trade-off with models achieving a low G-score need not be the most accurate models. An interesting finding was higher generalizability was achieved using just GPS features alone. Adding additional data sources such as accelerometers can further improve the accuracy of the predictive models (Roy et al., 2020), however, the data preprocessing would need to ensure noise removal and matching up temporal signatures of varying datasets (Roy et al., 2020).

4.8 Conclusion

Overall, our results can inform policymakers to better understand how context influences travel behavior in cities. The models are reproducible and can be used to predict transportation modes from GPS data and contextual information in other cities depending upon the availability of data. However, the results might vary and policymakers need to prioritize their goals of higher accuracy versus high generalizability to choose an optimal model that suits their needs.

The results generated in this paper could provide a guideline to policymakers on which additional factors to consider for predicting transportation modes beyond the traditional instrumental factors like distance, speed, time, and cost. These insights could help policymakers to better understand how and why the travel demand for different transport modes fluctuates with the dynamics of space, time, and place. The results can be utilized in helping them design more equitable infrastructure that could enhance the overall livability and usage of outdoor environments for one and all.

We present a general framework that can be used in different geographic settings to assess the classification accuracy as well as the generalizability of different transportation

modes from big, crowdsourced datasets utilizing an interdisciplinary perspective combining relations between mobility patterns, natural and built environment composition of the city as well as the comfort of its residents. Future research could elaborate these themes into several directions and utilize the classification framework to categorize transport modes for infrastructure management and assessing the exposure of different modes in other cities based on the availability of data.

CHAPTER 5

CONCLUSION

5.1 Summary of research

Movement data captured by crowdsourced fitness apps pose challenges in terms of extracting actionable information (Yang et al. 2017) for policymaking owing to temporal misalignment (Choi et al. 2018) or bias on app usage (Roy et al. 2019). The need for a more accurate and reliable understanding of movement patterns from such 'big' data requires improvements to algorithms that can recognize data inaccuracies and sampling errors. Often movement pattern analysis techniques involving data captured by crowdsourced platforms fail to integrate geographic context and the models developed are not sufficient to capture the geographic variations that influence processes that generate such data. The research carried out in this dissertation opens a new avenue for movement pattern analysis by means of developing novel methods for correcting bias, removing temporal misalignment to detect changes, and adding geographic context to movement data captured by crowdsourced platforms for transportation planning purposes.

5.2 Major contributions

The major academic contributions of this dissertation are in the context of developing novel methods combining geographic context and big, crowdsourced movement data to make it usable for practitioners and policymakers to facilitate urban planning.

5.2.1 Methods development for contextualizing movement data

Throughout this dissertation, I have made significant contributions to the current state of quantitative movement pattern analysis tied predominantly to three overarching themes – advancing the use of real-world big spatial-temporal data from crowdsourced platforms and its application to statistical and machine learning approaches, developing methods for integrating geographic covariates with raw movement datasets to analyze and assess the influence of context on movement patterns and applying the methods developed to real-world transportation research scenarios for demonstrating their usage in transportation planning and policymaking.

In Chapter 2, I provide a bias-correction framework for improving the representativeness of big movement datasets from crowdsourced fitness platforms. The method developed in this study is broadly applicable for correcting bias in crowdsourced bicycling data when official counts and geographical data are available at comparable spatial and temporal resolution. In this framework, I have identified the key contextual variables (i.e., proportion of white population, median household income, traffic speed, distance to residential areas, and distance to green spaces) which can account for the bias in bicycle ridership patterns, a special case of movement patterns, and predict the annual average bicycle ridership volumes of all residents in an entire city at the street-segment level. I also quantified the uncertainty in prediction margins of overall ridership from bias-corrected data with 86% of street segments being predicted within a margin of +/- 100 bicyclists.

In Chapter 3, I expanded the research on movement pattern analysis across different time scales by converting movement data into mappable time series represented as functional curves. I used a square root velocity function to correct temporal misalignment among the functional curves and then calculated the difference of the aligned function of each functional curve in a specific year from the mean curve in the previous year. The changes were categorized by k-means clustering and change maps were generated in the context of bicycle ridership volumes at hourly and monthly scales.

Finally, in Chapter 4, I integrated spatial context using relevant geographic covariates (distance to several built/natural environment classes, land use types) to classify active (walk/bike), public (subway, bus, train, rideshare), and private (car, taxi) transportation modes. The key variables identified for summarizing spatial context were distance to subways, distance to bicycle infrastructure, distance to bus stops, and distance to open green spaces. The prediction accuracy of Random Forest classifiers increased when contextual variables were added compared to using features extracted from raw movement datasets. The generalizability of the models was assessed by differentiating between train and test F1-scores to indicate how the models performed with new movement datasets in varied geographic settings.

Chapter 2 has been successfully published (Roy et al., 2019) as a research paper in the Urban Science journal, which is a peer-reviewed interdisciplinary journal for urban planners, computer scientists, and geographers. The work has already been used to inform planners about generalized approaches to mapping bicycle ridership data across multiple cities (Nelson et al., 2021), highlighting the role of bias-corrected crowdsourced data in monitoring the safety of commuters (Ferster et al., 2021) and highlighting the optimal

location to improve spatial coverage of crowdsourced data (Brum-Bastos et al., 2019). The choice of Urban Science as a venue for this work was motivated not simply by its impact on GIScience, but also due to its broad, multi-disciplinary readership by both researchers as well as practitioners. Since the findings of this paper have direct policy implications, the venue was appropriate for disseminating the results. Chapter 3 has been revised and submitted to the International Journal of Geographic Information Science owing to its novel methodological framework for temporal change detection from movement patterns. The manuscript is currently under review. Chapter 4 has been prepared for submission to the Computers, Environment and Urban Systems journal owing to its broad applicability in urban planning and public policy as well as computational complexity demonstrated in the development of the mode detection framework and generalizability assessment.

5.2.2 Reproducibility of code development for movement pattern analysis

While the development of methods is valuable in and of itself, it is also of much importance to make the methods generalizable and reproducible for use by fellow researchers and practitioners/policymakers. I have ensured the use of open-source software development frameworks including R and Python throughout my dissertation to summarize the implementation of the methods discussed in the previous chapters.

Through my research, I have developed open-source code available via GitHub using R and Python to replicate the model in new study areas contingent upon the availability of movement data. My methods are generalizable in the sense that very minor modifications need to be made to existing code for applying movement pattern analysis in different cities from crowdsourced data. In Chapter 2, I have developed functionality for

bias correction in R using generalized linear regression along with a variable selection approach for contextual variables in Python. I have also simultaneously developed the SRVF functionality to realign functional curves obtained from raw movement data of varying spatial and temporal resolution and a functional K-means algorithm to categorize changes in movement patterns over time in R used in Chapter 3. For Chapter 4, I developed a Python-based machine learning framework for transportation mode detection as well as generalizability assessment using supervised classification algorithms like Random Forests, Extreme Gradient Boost, Decision Trees, and AdaBoost.

5.2.3 Generating policy-ready outputs from context-driven movement pattern analysis

A broader implication of the research carried out throughout this dissertation is the direct policy implications that the results bear in the context of transportation planning. The methods described in Chapter 2 have already been utilized by local authorities in the City of Phoenix to predict bicycle ridership in 2020. The results in Chapter 3 will be disseminated among transportation agencies in the city of Phoenix and Tempe to track changes in active transportation models like cycling and walking over long periods. The work is inspired by Boss et al., 2018 but is an improvement over the methodology discussed in their study in terms of advocating change detection using temporal signatures from movement data. The results of Chapter 4 are of significance to planners and policymakers who wish to understand travel behavior in their respective cities but are limited in the knowledge of methods used to extract actionable insights from big movement datasets. The work has been performed as part of a bigger research endeavor called INTERACT

(INTERACT) that involves researchers, city planners, health professionals as well as computer scientists who aim to make cities safer and healthier.

5.3 Key Findings

Based on the methods and research I have developed through my Ph.D. I have identified several key findings related to movement pattern analysis from movement data and geographic context which are listed below.

- The significant variables for correcting bias in crowdsourced data for bicycle ridership were: The proportion of the white population, median household income, traffic speed, distance to residential areas, and distance to green spaces. Combining these geographic covariates with Strava counts accounted for additional factors that influence bicycle ridership and may not be captured solely by crowdsourced sampling.

- The model developed in Chapter 2 was used in the city of Tempe and our results showed that for 80.3% of road segments, where ground truth data were available, estimated bicycle counts were correct to within 25% of the observed counts (± 50 riders). The results of our study indicate that bias correction of crowdsourced data may prove to be a useful method for the estimation of bicycle ridership in North American cities.

- Using data from the Strava fitness app, captured every minute, we quantified ridership changes in Phoenix between 2017 and 2018 at the street segment level in Chapter 3. Hourly and monthly changes were classified into four categories – high peak, low peak, high off-peak, and low off-peak for hourly scales and high winter, low winter, high summer, and low summer for monthly scales and mapped along with exposure density.

- 32% of the street segments in Phoenix show a high hourly change in ridership during the peak period between 8 am and 10 am which accounts mostly for commute trips. These patterns indicated that improved bicycle-friendly infrastructure between 2017 and 2018 led to a major increase in peak-period ridership as commuters felt safe to bike to their workplace.

- On the monthly scale, most of the changes occur during the winter months which consist of 37.6% of the street segments (Table 3) (including high and low changes in winter) located mostly on the outskirts of the city with recreational bicyclists making more trips along trails and parks. The changes in summer overlapped with high crash density areas within the city center that have higher traffic speed limits.

- Geographic context captured by variables like distance to commercial and distance to green spaces or seashores were found to improve travel mode prediction accuracy in Chapter 4. These contextual variables are often important for leisure trips made on foot, bicycles, or in private vehicles who can spare time to interact more with their spatial surroundings as well as feel safe and comfortable. They also

seem to be a proxy for routes that have lower traffic volumes and safer speed limits which could be used as a distinguishing factor for active and private modes of transportation.

- The Random Forest model achieved the highest accuracy of 83.8% when contextual variables were combined with features extracted from GPS data indicating geographic context plays a significant role in transportation mode choice. The accuracy and generalizability of the models showed a trade-off with the addition of contextual variables. Since these variables are more closely tied to the immediate surroundings of a users' route, the model tends to overfit in some cases, however, AdaBoost was found to be more generalizable than the Random Forests, Gradient Boost, and Decision Trees indicating it could be applied to different study areas without compromising the accuracy.

5.4 Challenges and Limitations

The research conducted in each chapter developed a new method for spatial and temporal analysis of 'big' movement data from crowdsourced platforms. Since the methods were developed using a specific study area there are few limitations both in terms of data and methodology that must be acknowledged and accounted for before applying the methods in a broader context, especially for policymaking purposes.

In Chapter 1, the bias correction framework developed was based on ground truth data collected from Maricopa County's permanent counters spread randomly throughout the entire county. The placement of the counters was done before conducting our bias-

correction research, hence, the sampling scheme is not spatially distributed in a way that is representative of the entire population. These gaps in data originating as a result of non-uniform sampling may trickle down as errors generated during the modeling process thereby increasing variance in the predicted estimates of overall bicycling ridership.

Additionally, the model was trained using county-level data but tested with city-level count data, therefore reapplying the global model to a local setting, which also added to the variance in the predictions. It highlights that with different geographic settings the same model may not be the best fit for re-estimating the bias-adjusted bicycle ridership volumes as to the same set of variables selected may not adjust for Strava bias in all study areas. A possible reason could be that the variables that are used for accounting bias vary spatially at local scales with the difference in bicycling cultures in different regions or cities. A more nuanced or city-specific choice of variables will be needed to improve the accuracy of the bias-correction model based on local knowledge about the city and what factors might influence ridership in that area.

Additionally, the point locations of permanent counters were compared with street segment level Strava counts and we ignored the directionality of the bicycling trips to get an aggregated estimate of total trips that occurred at the same time period for which official counts were collected. The choice of comparing two different spatial representations of point-based permanent counts with line-based counts from Strava was performed in order to generate street-segment level maps for the entire city of Tempe which would be otherwise impossible given the data organization and structure of the network level count aggregation framework applied by the Strava Metro platform to anonymize the data to preserve user's privacy.

A major challenge was the preprocessing of nearly 1.5 million Strava trips to assimilate and organize the data in a manner that was spatially and temporally comparable to the official counts as well as socioeconomic, demographic, and land-use data collected from disparate sources and at different spatial resolutions. The two sets of ground truth data one of training from the MAG and the other from TBAG used for testing the predictions were also sparse overall – given there were only 65 locations to compare the ground truth of predicted AADB counts. Future work will require tailoring data collection efforts towards the strategic placement of counters as well as planning to get geographic data in a single data platform for ease of researchers in terms of building statistical models for planning purposes. The statistical significance of the models can be hugely improved with help from well-planned spatial coverage of official permanent counters.

Chapter 3 highlights the use of FDA techniques for change monitoring purposes. Although the data resolution for this study was fine-grained, there was a limitation in terms of contextualizing the results of the study to understand why the changes happened. Additional data from Bikemaps.org had to be used to overlay the change maps in order to visualize which areas were prone to a high change and whether more bicycle-related incidents were recorded by the crowdsourcing platform Bikempas.org in those areas. We created a visual method to represent the ridership changes via a change map. However, developing a technique to determine the statistical significance of the functional change values could be useful to quantify the confidence intervals of the results. It will also be beneficial for researchers to get infrastructure-related data as an additional attribute along with the number of trips from the Strava data itself which can help in drawing inferences about why the change occurred in the first place. Further research is also needed to develop

a statistical significance measure for changes identified for each change cluster and quantify the uncertainties in change detected. In addition, to make the change maps more readable and usable by planners an interactive tool for visualizing changes over the entire street network will be beneficial.

In terms of the mode detection study, a major limitation was the availability of GPS data with labeled trips from a fourth city which could be used as a test bed for the generalizability of the model. Although the paper hypothesizes the generation of a G-score-based generalizability metric, it is challenging to test its validity without being able to apply the method to trips from another city which was not included in the training of the models. Additionally, the predictive accuracy of the Random Forest model could be further improved by having a more balanced dataset meaning an equal number of trips for each mode – active, public, and private or by introducing a weighting technique to account for the imbalance in the labeled trip data. As an introductory step to analyzing the generalizability of transport mode detection models using contextual data, the study highlighted a set of geographic covariates but additional research is needed to understand which factors are important predictors and could provide consistent data based on availability across multiple cities. This approach is a pilot study and would need further validation in different geographic settings to make conclusive remarks about the efficiency and applicability of the G-score approach. A major challenge in terms of the trip data acquired from three different cities was also that the number of trips varied greatly across these cities. Montreal collected the majority of the trip data with over 80% of the trips coming from Montreal whereas the remaining trips which came from Vancouver and St. John's were a much smaller subset of the entire data. As a measure to account for the skew

in the data volume we combined trip data from all three cities and split them into validation and test sets – but the accuracy could be greatly improved if the trip data are comparable across all cities.

5.5 Future Work

Utilizing the methods developed throughout my dissertation policymakers can bridge the gap between computational modeling and targeted decision-making by evaluating existing infrastructure as well as planning for future investments in new infrastructure to build safer and healthier cities. The findings of my dissertation are key in understanding the role of geographic context in movement pattern analysis in transportation planning. More specifically, it highlights the use of bias-corrected crowdsourced data in examining bicycle ridership patterns for an entire city, understanding changes in hourly and monthly bicycle ridership patterns and how safety and infrastructure relate to those changes, and the role of geographic context in transportation mode detection from movement data.

The method for correcting bias in crowdsourced data with the help of a three-step mixed-model approach in Chapter 2 is broadly applicable for correcting bias in crowdsourced bicycling data when official counts and geographical data are available at comparable spatial and temporal resolution. Based on those results, in the future, it is suggested that local transportation authorities should work closely with researchers to improve the coverage of official count data, helping them to identify locations to place counters so that a denser spatial coverage, as well as more ground truth data, are obtained to improve the model's performance. The proposed bias correction model, with detailed

data that is continuous through space and collected repeatedly in time, can help transportation planners in making informed decisions related to bicycle infrastructure planning to promote healthier lifestyles among urban residents of all ages and abilities. Detailed maps of bicycling ridership are critical to professionals in making decisions regarding infrastructure investment and policy changes that support active transportation. The statistical approach developed in this Chapter can be used to help stratify bicycling count programmes (Brum-Bastos et al., 2020) by strategic placement of temporary or permanent counters and generate a hypothesis on why variation in prediction varies by applying it to other cities as an effort to understand the best approaches for modeling bicycling volumes from crowdsourced data platforms like Strava Metro (Nelson et al., 2021). It is suggested that local transportation authorities should work closely with researchers to improve the coverage of official count data, helping them to identify locations to place counters so that a denser spatial coverage, as well as more ground truth data, are obtained to improve the model's performance. The proposed bias correction model, with detailed data that is continuous through space and collected repeatedly in time, can help transportation planners in making informed decisions related to bicycle infrastructure planning to promote healthier lifestyles among urban residents of all ages and abilities. Detailed maps of bicycling ridership are critical to professionals in making decisions regarding infrastructure investment and policy changes that support active transportation. The framework developed in this paper can be used as a generalized approach to include bias-corrected crowdsourced data for planning purposes to generate street-segment level maps that can assess crime patterns or model exposure by including and testing different sets of geographic covariates that influence mobility patterns. The

model also needs to be tested in different cities to quantify the uncertainties in prediction and assess its variability at different spatial and temporal scales.

Chapter 3 contributes to debates in time geography based on the theoretical foundation on how time and space constitute social life from the scale of individuals (Hägerstrand 1985). Considering previous research (Kwan 2002, Miller 2005, Long and Nelson 2013, Kwan and Neutens 2014) that highlight the role of underlying time and scale issues in geography, this paper builds a framework for analyzing change from real-world data at fine-grained scales and contributes to the field of urban analytics from a methodological perspective which can help policymakers. In the future, the work described here can be expanded to capture change monitoring for transportation networks in cities before and after natural hazards as well as assess the impacts of climate change on land use and land cover changes over time.

Finally, the method and results highlighted in Chapter 4 can inform policymakers to better understand how context influences travel behavior in cities. The models are reproducible and can be used to predict transportation modes from GPS data and contextual information in other cities depending upon the availability of data. The results generated in this paper could provide a guideline to policymakers on which additional factors to consider for predicting transportation modes beyond the traditional instrumental factors like distance, speed, time, and cost. These insights could help policymakers to better understand how and why the travel demand for different transport modes fluctuates with the dynamics of space, time, and place. The results can be utilized in helping them design more equitable infrastructure that could enhance the overall livability and usage of outdoor environments for one and all. Future research will employ techniques to improve the

accuracy of the models integrating spatial context by predicting the transportation modes in different cities using movement data that were not used to train the models previously. The results can also be used to motivate streamlined efforts in data collection with comparable spatial and temporal coverage across cities from which transportation modes could be predicted.

REFERENCES

Agrawal, R., Faloutsos, C., and Swami, A. N. 1993. "*Efficient Similarity Search In Sequence Databases.*" In Proceedings of the 4th international Conference on Foundations of Data Organization and Algorithms. D. B. Lomet, Ed. Lecture Notes In Computer Science, Springer-Verlag, London, 730, 69-84.

Akar, Gulsah, and Kelly J. Clifton. "Influence of individual perceptions and bicycle infrastructure on decision to bike." Transportation research record 2140, no. 1 (2009): 165-172.

Alaya, Mohamed Ali Ben, Camille Ternynck, Sophie Dabo-Niang, Fateh Chebana, and Taha BMJ Ouarda. "Change point detection of flood events using a functional data framework." Advances in Water Resources 137 (2020): 103522.

Andrienko, Gennady, Natalia Andrienko, and Stefan Wrobel. "Visual analytics tools for analysis of movement data." ACM SIGKDD Explorations Newsletter 9, no. 2 (2007): 38-46.

Andrienko, N., Andrienko, G. and Gatalsky, P. 2005. "Impact of data and task characteristics on design of spatio-temporal data visualization tools". In Exploring geovisualization, Edited by: Dykes, J.A., MacEachren, A.M. and Kraak, M.J. 201–222. New York: Elsevier.

Anguita, D., Ghio, A., Oneto, L., Parra, X., & Reyes-Ortiz, J. L. (2012, December). *Human activity recognition on smartphones using a multiclass hardware-friendly support vector machin*e. In International workshop on ambient assisted living (pp. 216-223). Springer, Berlin, Heidelberg.

Anselin, L. (1995). Local indicators of spatial association—LISA. Geographical analysis, 27(2), 93-115.

Antoniadis, Anestis, Xavier Brossat, Jairo Cugliari, and Jean-Michel Poggi. "Clustering functional data using wavelets." International Journal of Wavelets, Multiresolution and Information Processing 11, no. 01 (2013): 1350003.

Aston, John AD, and Claudia Kirch. "Detecting and estimating changes in dependent functional data." Journal of Multivariate Analysis 109 (2012): 204-220.

Aue, Alexander, Robertas Gabrys, Lajos Horváth, and Piotr Kokoszka. "Estimation of a change-point in the mean function of functional data." Journal of Multivariate Analysis 100, no. 10 (2009): 2254-2269.

Auld, Joshua, Chad Williams, Abolfazl Mohammadian, and Peter Nelson. "An automated GPS-based prompted recall survey with learning algorithms." Transportation Letters 1, no. 1 (2009): 59-79.

Aziz, HM Abdul, Nicholas N. Nagle, April M. Morton, Michael R. Hilliard, Devin A. White, and Robert N. Stewart. "Exploring the impact of walk–bike infrastructure, safety perception, and built-environment on active transportation mode choice: a random parameter model using New York City commuter data." *Transportation* 45, no. 5 (2018): 1207-1229.

Ballari, Daniela, Ramón Giraldo, Lenin Campozano, and Esteban Samaniego. "Spatial functional data analysis for regionalizing precipitation seasonality and intensity in a sparsely monitored region: Unveiling the spatio-temporal dependencies of precipitation in Ecuador." International Journal of Climatology 38, no. 8 (2018): 3337-3354.

Barbiero, Pietro, Giovanni Squillero, and Alberto Tonda. "Modeling Generalization in Machine Learning: A Methodological and Computational Study." arXiv preprint arXiv:2006.15680 (2020).

Barraquand, F., & Benhamou, S. (2008). "Animal movements in heterogeneous landscapes: identifying profitable places and homogeneous movement bouts." Ecology, 89(12), 3336–48. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/19137941

Barraquand, F., & Benhamou, S. (2008). "Animal movements in heterogeneous landscapes: identifying profitable places and homogeneous movement bouts." Ecology, 89(12), 3336–48. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/19137941

Benhamou, S. (2004). "How to reliably estimate the tortuosity of an animal's path: straightness, sinuosity, or fractal dimension?" Journal of Theoretical Biology, 229(2), 209–20. doi:10.1016/j.jtbi.2004.03.016

Benkert, Marc, Joachim Gudmundsson, Florian Hübner, and Thomas Wolle. "Reporting flock patterns." Computational Geometry 41, no. 3 (2008): 111-125.

Berndt, Donald J., and James Clifford. "Using dynamic time warping to find patterns in time series." In KDD workshop, vol. 10, no. 16, pp. 359-370. 1994.

Bíl, M.; Andrášik, R.; Kubeček, J. How comfortable are your cycling tracks? A new method for objective bicycle vibration measurement. Transp. Res. Part C: Emerg. Technol. 2015, 56, 415–425.

Bleisch, Susanne, Matt Duckham, Antony Galton, Patrick Laube, and Jarod Lyon. "Mining candidate causal relationships in movement patterns." International Journal of Geographical Information Science 28, no. 2 (2014): 363-382.

Böcker, Lars, Martin Dijst, Jan Faber, and Marco Helbich. "En-route weather and place valuations for different transport mode users." Journal of Transport Geography 47 (2015): 128-138.

Bohte, Wendy, and Kees Maat. "Deriving and validating trip purposes and travel modes for multi-day GPS-based travel surveys: A large-scale application in the Netherlands." Transportation Research Part C: Emerging Technologies 17, no. 3 (2009): 285-297.

Boss, D., Nelson, T., & Winters, M. (2018). Monitoring city wide patterns of cycling safety. Accident Analysis & Prevention, 111, 101-108.

Boss, Darren, Trisalyn Nelson, Meghan Winters, and Colin J. Ferster. "Using crowdsourced data to monitor change in spatial patterns of bicycle ridership." Journal of Transport & Health (2018): 226-233.

Boulange, Claire, Lucy Gunn, Billie Giles-Corti, Suzanne Mavoa, Chris Pettit, and Hannah Badland. "Examining associations between urban design attributes and transport mode choice for walking, cycling, public transport and private motor vehicle trips." Journal of Transport & Health 6 (2017): 155-166.

Bourbonnais, Mathieu L., Trisalyn A. Nelson, Gordon B. Stenhouse, Michael A. Wulder, Joanne C. White, Geordie W. Hobart, Txomin Hermosilla, Nicholas C. Coops, Farouk Nathoo, and Chris Darimont. "Characterizing spatial-temporal patterns of landscape disturbance and recovery in western Alberta, Canada using a functional data analysis approach and remotely sensed data." Ecological informatics 39 (2017): 140-150.

Bourbonnais, Mathieu L., Trisalyn A. Nelson, Gordon B. Stenhouse, Michael A. Wulder, Joanne C. White, Geordie W. Hobart, Txomin Hermosilla, Nicholas C. Coops, Farouk Nathoo, and Chris T. Darimont. "A functional data analysis approach for characterizing spatial-temporal patterns of landscape disturbance and recovery from remotely sensed data." In CEUR Workshop Proceedings, vol. 2323, p. 4. CEUR-WS, 2019.

Brandenburg, Christiane, Andreas Matzarakis, and Arne Arnberger. "Weather and cycling—a first approach to the effects of weather conditions on cycling." Meteorological Applications: A journal of forecasting, practical applications, training techniques and modelling 14, no. 1 (2007): 61-67.

Bricka, S., & Bhat, C. R. (2006). Comparative analysis of Global Positioning System–based and travel survey–based data. Transportation Research Record, 1972(1), 9-20.

Brum-Bastos, Vanessa, Colin J. Ferster, Trisalyn Nelson, and Meghan Winters. "Where to put bike counters? Stratifying bicycling patterns in the city using crowdsourced data." Transport Findings (2019).

Buchin, M., Driemel, A., Van Kreveld, M., & Sacristán, V. (2010, November). An algorithmic framework for segmenting trajectories based on spatio-temporal criteria. In Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems (pp. 202-211). ACM.

Buchin, M., Driemel, A., Van Kreveld, M., & Sacristán, V. (2010, November). An algorithmic framework for segmenting trajectories based on spatio-temporal criteria. In Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems (pp. 202-211). ACM.

Buehler, Ralph, and John Pucher. "Cycling to work in 90 large American cities: new evidence on the role of bike paths and lanes." Transportation 39, no. 2 (2012): 409-432.

Calenge, C. (2006). “The package “adehabitat” for the R software: A tool for the analysis of space and habitat use by animals.” Ecological Modelling, 197(3-4), 516–519. doi:10.1016/j.ecolmodel.2006.03.017

Calenge, Clement, and Maintainer Clement Calenge. "Package ‘adehabitat’." R package version 1 (2015): 18.

Calenge, Clément, Stéphane Dray, and Manuela Royer-Carenzi. "The concept of animals' trajectories from a data analysis perspective." Ecological informatics 4, no. 1 (2009): 34-41.

Calenge, Clément. “The package “adehabitat” for the R software: a tool for the analysis of space and habitat use by animals.” Ecological modelling 197, no. 3-4 (2006): 516-519.

Carlson, Jordan A., Brian E. Saelens, Jacqueline Kerr, Jasper Schipperijn, Terry L. Conway, Lawrence D. Frank, Jim E. Chapman, Karen Glanz, Kelli L. Cain, and James F. Sallis. "Association between neighborhood walkability and GPS-measured walking, bicycling and vehicle time in adolescents." Health & Place 32 (2015): 1-7.

Chawla, Nitesh V., Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. "SMOTE: synthetic minority over-sampling technique." Journal of Artificial Intelligence Research 16 (2002): 321-357.

Chen, Cynthia, Hongmian Gong, Catherine Lawson, and Evan Bialostozky. "Evaluating the feasibility of a passive travel survey collection in a complex urban environment: Lessons learned from the New York City case study." Transportation Research Part A: Policy and Practice 44, no. 10 (2010): 830-840.

Chen, Cynthia, Jingtao Ma, Yusak Susilo, Yu Liu, and Menglin Wang. "The promises of big data and small data for travel behavior (aka human mobility) analysis." Transportation Research Part C: Emerging Technologies 68 (2016): 285-299.

Chen, L., Özsu, M. T. & Oria, V. 2005. “Robust and Fast Similarity Search for Moving Object Trajectories.” Proceedings of the ACM SIGMOD International Conference on Management of Data, ACM. 491-502.

Chen, L., Özsu, M. T. & Oria, V. 2005. “Robust and Fast Similarity Search for Moving Object Trajectories.” Proceedings of the ACM SIGMOD International Conference on Management of Data, ACM. 491-502. 32. Lin, B. and Su, J. 2008. “One Way Distance: For Shape Based Similarity Search of Moving Object Trajectories.” GeoInformatica, 12, 117-142.

Chen, P.; Shen, Q. Built environment effects on cyclist injury severity in automobile-involved bicycle crashes. Anal. Prev. 2016, 86, 239–246.

Cheng, Long, Xuewu Chen, Jonas De Vos, Xinjun Lai, and Frank Witlox. "Applying a random forest method approach to model travel mode choice behavior." Travel Behaviour and Society 14 (2019): 1-10.

Choi, Hongjun, Qiao Wang, Meynard Toledo, Pavan Turaga, Matthew Buman, and Anuj Srivastava. "Temporal alignment improves feature quality: an experiment on activity recognition with accelerometer data." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 349-357. 2018.

City of Phoenix, Bicycle Master Plan report, 2015 https://www.phoenix.gov/streetssite/Documents/Bicycle%20Master%20Plan/2014bikePHX_Final_web.pdf . Accessed October 7, 2020.

City of Tempe. Tempe Transportation Master Plan. 2015. Available online: http://www.tempe.gov/home/showdocument?id=30317 (accessed on 20 April 2018).

City of Vancouver (2018), Transportation panel survey 2018, https://vancouver.ca/files/cov/2018-transportation-panel-survey.pdf

Codling, E. a, Plank, M. J., & Benhamou, S. (2008). “Random walk models in biology.” Journal of the Royal Society, Interface / the Royal Society, 5(25), 813–34. doi:10.1098/rsif.2008.0014

Codling, E. a, Plank, M. J., & Benhamou, S. (2008). “Random walk models in biology.” Journal of the Royal Society, Interface / the Royal Society, 5(25), 813–34. doi:10.1098/rsif.2008.0014

Colberg, S.R.; Sigal, R.J.; Fernhall, B.; Regensteiner, J.G.; Blissmer, B.J.; Rubin, R.R.; Chasan-Taber, L.; Albright, A.L.; Braun, B.; American College of Sports Medicine; et al. Exercise and type 2 diabetes: The American College of Sports Medicine and the American Diabetes Association: joint position statement executive summary. Diabetes Care 2010, 33, 2692–2696.

Correa, Carlos D., Tarik Crnovrsanin, Christopher Muelder, Zeqian Shen, Ryan Armstrong, James Shearer, and Kwan-Liu Ma. "Visual analytics of cell phone data using MobiVis and OntoVis." In IEEE symposium on visual analytics science and technology, pp. 211-212. 2008.

Cortes, C., & Vapnik, V. (1995). Support vector machine. Machine learning, 20(3), 273-297.

Crawley, M.J. Statistics: An Introduction Using R; Wiley: Hoboken, NJ, USA, 2005; ISBN 0 470.02298.

Cui, Boer, Geneviève Boisjoly, Luis Miranda-Moreno, and Ahmed El-Geneidy. "Accessibility matters: Exploring the determinants of public transport mode share across income groups in Canadian cities." Transportation Research Part D: Transport and Environment 80 (2020): 102276.

Demšar, Urška, and Kirsi Virrantaus. "Space–time density of trajectories: exploring spatio-temporal patterns in movement data." International Journal of Geographical Information Science 24.10 (2010): 1527-1542.

Demšar, Urška, Kevin Buchin, Francesca Cagnacci, Kamran Safi, Bettina Speckmann, Nico Van de Weghe, Daniel Weiskopf, and Robert Weibel. "Analysis and visualisation of movement: an interdisciplinary review." Movement Ecology 3, no. 1 (2015): 5.

Dennis, Todd E., William C. Chen, Inigo Koefoed, Shabana F. Shah, Michael M. Walker, Patrick Laube, and Pip Forer. "Performance characteristics of small global-positioning-system tracking collars." Wildlife Biology in Practice 6, no. 1 (2010): 14-31.

Dill, Jennifer, and Theresa Carr. "Bicycle commuting and facilities in major US cities: if you build them, commuters will use them." Transportation Research Record 1828, no. 1 (2003): 116-123.

Dobson, A.J.; Barnett, A.G. An Introduction to Generalized Linear Models; CRC Press: Boca Raton, FL, USA, 2008; ISBN 9781138741515.

Dodge, S., Weibel, R. and Lautenschütz, A.-K. 2008. Towards a Taxonomy of Movement Patterns. Journal of Information Visualization, 7, 240-252.

Dodge, Somayeh, Patrick Laube, and Robert Weibel. "Movement similarity assessment using symbolic representation of trajectories." International Journal of Geographical Information Science 26, no. 9 (2012): 1563-1588.

Dodge, Somayeh, Robert Weibel, and Anna-Katharina Lautenschütz. "Towards a taxonomy of movement patterns." Information visualization 7, no. 3-4 (2008): 240-252.

Dodge, Somayeh, Robert Weibel, and Ehsan Forootan. "Revealing the physics of movement: Comparing the similarity of movement characteristics of different types of moving objects." Computers, Environment and Urban Systems 33, no. 6 (2009): 419-434.

Dodge, Somayeh, Robert Weibel, Sean C. Ahearn, Maike Buchin, and Jennifer A. Miller. "Analysis of movement data." (2016): 825-834.

Eccles, Ryan, Thomas Kapler, Robert Harper, and William Wright. "Stories in geotime." Information Visualization 7, no. 1 (2008): 3-17.

El Esawey, M.; Mosa, A.I.; Nasr, K. Estimation of daily bicycle traffic volumes using sparse data. Comput. Environ. Urban Syst. 2015, 54, 195–203.

Ellis, K., Godbole, S., Marshall, S., Lanckriet, G., Staudenmayer, J., & Kerr, J. (2014). Identifying active travel behaviors in challenging environments using GPS, accelerometers, and machine learning algorithms. Frontiers in public health, 2, 36.

Embling, Clare B., Janine Illian, Eric Armstrong, Jeroen van der Kooij, Jonathan Sharples, Kees CJ Camphuysen, and Beth E. Scott. "Investigating fine-scale spatio-temporal predator–prey patterns in dynamic marine ecosystems: a functional data analysis approach." Journal of Applied Ecology 49, no. 2 (2012): 481-492.

Ewing, Reid, and Robert Cervero. "Travel and the built environment: A meta-analysis." Journal of the American Planning Association 76, no. 3 (2010): 265-294.

Feick, R.; Roche, S. Understanding the Value of VGI. In Crowdsourcing Geographic Knowledge; Sui, D., Elwood, S., Goodchild, M., Eds.; Springer: Dordrecht, The Netherlands, 2013; pp. 15–29.

Feng, T., & Timmermans, H. J. (2013). Transportation mode recognition using GPS and accelerometer data. Transportation Research Part C: Emerging Technologies, 37, 118-130.

Ferster, Colin, Trisalyn Nelson, Karen Laberee, and Meghan Winters. "Mapping bicycling exposure and safety risk using Strava Metro." Applied Geography 127 (2021): 102388.

Fithian, W.; Elith, J.; Hastie, T.; Keith, D.A. Bias correction in species distribution models: pooling survey and collection data for multiple species. Methods Ecol. Evol. 2015, 6, 424–438.

Fleming, C. H., Calabrese, J. M., Mueller, T., Olson, K. a, Leimgruber, P., & Fagan, W. F. (2014). "From fine-scale foraging to home ranges: a semivariance approach to identifying movement modes across spatiotemporal scales." The American Naturalist, 183(5), 154–67. doi:10.1086/675504

Ford, Alistair C., Stuart L. Barr, Richard J. Dawson, and Philip James. "Transport accessibility analysis using GIS: Assessing sustainable transport in London." ISPRS International Journal of Geo-Information 4, no. 1 (2015): 124-149.

Forrest, Timothy L., and David F. Pearson. "Comparison of trip determination methods in household travel surveys enhanced by a global positioning system." Transportation Research Record 1917, no. 1 (2005): 63-71.

Fournier, Nicholas, Eleni Christofa, and Michael A. Knodler Jr. "A mixed methods investigation of bicycle exposure in crash rates." Accident Analysis & Prevention 130 (2019): 54-61.

Friedman, Jerome H. "Greedy function approximation: a gradient boosting machine." Annals of Statistics (2001): 1189-1232.

Fryxell, J. M., Hazell, M., Börger, L., Dalziel, B. D., Haydon, D. T., Morales, J. M., … Rosatte, R. C. (2008). "Multiple movement modes by large herbivores at multiple spatiotemporal scales." Proceedings of the National Academy of Sciences of the United States of America, 105(49), 19114–9. doi:10.1073/pnas.0801737105

Gaffney, S., & Smyth, P. (1999, August). Trajectory clustering with mixtures of regression models. In Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 63-72). ACM.

GENEActiv accelerometer device. https://www.activinsights.com/products/geneactiv/

Gong, Hongmian, Cynthia Chen, Evan Bialostozky, and Catherine T. Lawson. "A GPS/GIS method for travel mode detection in New York City." Computers, Environment and Urban Systems 36, no. 2 (2012): 131-139.

Goodrich, M.T., Mitchell, J.S.B., & Orletsky, M.W. 1999. "Approximate geometric pattern matching under rigid motions." IEEE Transaction on Pattern Analysis and Machine Intelligence, 21, 4, 371-379.

Graham, K., & Stenhouse, G. B. (2014). "Home Range, Movements, and Denning Chronology of the Grizzly Bear (Ursus arctos) in West-Central Alberta." The Canadian Field-Naturalist, 128(3), 223–234.

Griffin, G.; Nordback, K.; Götschi, T.; Stolz, E.; Kothuri, S. Monitoring Bicyclist and Pedestrian Travel and Behavior: Current Research and Practice; Transportation Research Circular E-C18; 2014. http://www.trb.org/Publications/Blurbs/170452.aspx (Accessed 31 May 2018)

Griffin, G.P.; Jiao, J. Where does bicycling for health happen? Analysing volunteered geographic information through place and plexus. J. Transp. Health 2015, 2, 238–247.

Griffin, Greg P., and Junfeng Jiao. "Where does bicycling for health happen? Analysing volunteered geographic information through place and plexus." Journal of Transport & Health 2, no. 2 (2015): 238-247.

Griswold, J.B.; Medury, A.; Schneider, R.J.; Information, R. Pilot Models for Estimating Bicycle Intersection Volumes. Transp. Res. Rec. J. Transp. Res. Board 2011, 2247, 1–7.

Guardiola, Ivan G., Teresa Leon, and Fermin Mallor. "A functional approach to monitor and recognize patterns of daily traffic profiles." Transportation Research Part B: Methodological 65 (2014): 119-136.

Gurarie, E. (2013). “Behavioral Change Point Analysis in R : The bcpa package”, 1–16.

Gurarie, E., Andrews, R. D., & Laidre, K. L. (2009). “A novel method for identifying behavioural changes in animal movement data.” Ecology Letters, 12(5), 395–408. doi:10.1111/j.1461-0248.2009.01293.x

Gurarie, E., Bracis, C., Delgado, M., Meckley, T. D., Kojola, I., & Wagner, C. M. (2015). “What is the animal doing? Tools for exploring behavioral structure in animal movements.” Journal of Animal Ecology, n/a–n/a. doi:10.1111/1365-2656.12379

Gurarie, Eliezer, Russel D. Andrews, and Kristin L. Laidre. "A novel method for identifying behavioural changes in animal movement data." Ecology letters 12, no. 5 (2009): 395-408.

Hägerstrand, T., 1985. Time geography: focus on the corporeality of man, society and environment. In: S. Aida, ed. The science and Praxis of complexity. Tokyo: The United Nations University, 193–216.

Hamann, C.; Peek-Asa, C. On-road bicycle facilities and bicycle crashes in Iowa, 2007–2010. Accid. Anal. Prev. 2013, 56, 103–109.

Hamel, L. H. (2011). Knowledge discovery with support vector machines (Vol. 3). John Wiley & Sons.

Han, J., Pei, J., & Yin, Y. (2000, May). “Mining frequent patterns without candidate generation.” In ACM sigmod record(Vol. 29, No. 2, pp. 1-12). ACM.

Han, S. Y., Tsou, M. H., Knaap, E., Rey, S., & Cao, G. (2019). How do cities flow in an emergency? Tracing human mobility patterns during a natural disaster with big data and geospatial data science. Urban Science, 3(2), 51.

Hand, D. J., & Till, R. J. (2001). A simple generalisation of the area under the ROC curve for multiple class classification problems. Machine learning, 45(2), 171-186.

Hankey, S.; Lindsey, G.; Wang, X.; Borah, J.; Hoff, K.; Utecht, B.; Xu, Z. Estimating use of non-motorized infrastructure: Models of bicycle and pedestrian traffic in Minneapolis, MN. Landsc. Urban Plan. 2012, 107, 307–316.

Heesch, K.C.; Langdon, M. The usefulness of GPS bicycle tracking data for evaluating the impact of infrastructure change on cycling behaviour. Health Promot. J. Aust. 2016, 27, 222–229.

Heinen, Eva, Kees Maat, and Bert Van Wee. "Day-to-day choice to commute or not by bicycle." Transportation Research Record 2230, no. 1 (2011): 9-18.

Hemminki, S., Nurmi, P., & Tarkoma, S. (2013, November). Accelerometer-based transportation mode detection on smartphones. In Proceedings of the 11th ACM conference on embedded networked sensor systems (pp. 1-14).

Hornsby, Kathleen, and Max J. Egenhofer. "Modeling moving objects over multiple granularities." Annals of Mathematics and Artificial Intelligence 36, no. 1-2 (2002): 177-194.

Huang, L.; Stinchcomb, D.G.; Pickle, L.W.; Dill, J.; Berrigan, D. Identifying Clusters of Active Transportation Using Spatial Scan Statistics. Am. J. Prev. Med. 2009, 37, 157–166.

J. Gudmundsson, P. Laube, and T. Wolle. Movement patterns in spatio-temporal data. In Encyclopedia of GIS 2008.

Jahangiri, A., & Rakha, H. (2014, January). Developing a support vector machine (SVM) classifier for transportation mode identification by using mobile phone sensor data. In Transportation Research Board 93rd Annual Meeting (No. 14-1442).

Jestico, B.; Nelson, T.; Winters, M. Mapping ridership using crowdsourced cycling data. J. Geogr. 2016, 52, 90–97.

Jonsen, I., Flemming, J., & Myers, R. (2005). “Robust state-space modeling of animal movement data.” Ecology, 86(11), 2874–2880. Retrieved from http://www.esajournals.org/doi/abs/10.1890/04-1852

Jordahl, K. GeoPandas: Python Tools for Geographic Data. Geopandas, 2014. Available online: https://github.com/geopandas/geopandas (accessed on 31 May 2019).

Kalnis, P., Mamoulis, N., & Bakiras, S. (2005, August). On discovering moving clusters in spatio-temporal data. In International Symposium on Spatial and Temporal Databases(pp. 364-381). Springer, Berlin, Heidelberg.

Kang, Jeon-Young, and Jared Aldstadt. "Using multiple scale spatio-temporal patterns for validating spatially explicit agent-based models." International Journal of Geographical Information Science 33, no. 1 (2019): 193-213.

Kapler, Thomas, and William Wright. "GeoTime information visualization." Information visualization 4, no. 2 (2005): 136-146.

Kareiva, P. M., & Shigesada, N. (1983). “Analyzing insect movement as a correlated random walk.” Oecologia, 56(2-3), 234–238.

Kim, Kyusik, Kyusang Kwon, and Mark W. Horner. "Examining the Effects of the Built Environment on Travel Model Choice across Different Age Groups in Seoul using a Random Forest Method." Transportation Research Record (2021): 03611981211000750.

Kim, Young-Long. "Data-driven approach to characterize urban vitality: how spatiotemporal context dynamically defines Seoul’s nighttime." International Journal of Geographical Information Science 34, no. 6 (2020): 1235-1256.

Kirkeby, Carsten, Maren Wellenreuther, and Mikkel Brydegaard. "Observations of movement dynamics of flying insects using high resolution lidar." Scientific reports 6 (2016): 29083.

Kluyver, T.; Ragan-Kelley, B.; Pérez, F.; Granger, B.E.; Bussonnier, M.; Frederic, J.; Kelley, K.; Hamrick, J.B.; Grout, J.; Corlay, S.; et al. Jupyter Notebooks—A publishing format for reproducible computational workflows. In Positioning and Power in Academic Publishing: Players, Agents and Agendas; IOS Press: Amsterdam, The Netherlands, 2016; pp. 87–90.

Kohavi, R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI’95), Montreal, QC, Canada, 20–25 August 1995; Morgan Kaufmann Publishers Inc.: San Francisco, CA, USA, 1995; Volume 2.

Kraak, Menno-Jan. "Geovisualization and time–new opportunities for the space–time cube." Geographic visualization: concepts, tools and applications (2008): 293-306.

Kushi, L.H.; Doyle, C.; McCullough, M.; Rock, C.L.; Demark-Wahnefried, W.; Bandera, E.V.; Gapstur, S.; Patel, A.V.; Andrews, K.; Gansler, T.; et al. American Cancer Society guidelines on nutrition and physical activity for cancer prevention: Reducing the risk of cancer with healthy food choices and physical activity. CA Cancer J. Clin. 2012, 62, 30–67.

Kwan, M.-P. and Neutens, T., 2014. Space-time research in GIScience. International Journal of Geographical Information Science, 28 (5), 851–854. doi:10.1080/13658816.2014.889300

Kwan, M.-P., 2002. Time, information technologies, and the geographies of everyday life. Urban Geography, 23 (5), 471–482. doi:10.2747/0272-3638.23.5.471

Kwan, Mei-Po. "Interactive geovisualization of activity-travel patterns using three-dimensional geographical information systems: a methodological exploration with a large data set." Transportation Research Part C: Emerging Technologies 8, no. 1-6 (2000): 185-203.

Laberee, Karen, Trisalyn A. Nelson, Benjamin P. Stewart, Tracy McKay, and Gordon B. Stenhouse. "Oil and gas infrastructure and the spatial pattern of grizzly bear habitat selection

in Alberta, Canada." The Canadian Geographer/Le Géographe canadien 58, no. 1 (2014): 79-94.

Larsen, J.; Patterson, Z.; El-Geneidy, A. Build It. But Where? The Use of Geographic Information Systems in Identifying Locations for New Cycling Infrastructure. Int. J. Sustain. Transp. 2013, 7, 299–317.

Laube, P., and T. Dennis. "Exploratory analysis of movement trajectories." In GeoCart,.(2006). National Cartographic Conference. Auckland, NZ. 2006.

Laube, Patrick, and Ross S. Purves. "How fast is a cow? cross-scale analysis of movement data." Transactions in GIS15, no. 3 (2011): 401-418.

Laube, Patrick, Marc van Kreveld, and Stephan Imfeld. "Finding REMO—detecting relative motion patterns in geospatial lifelines." In Developments in spatial data handling, pp. 201-215. Springer, Berlin, Heidelberg, 2005.

Laube, Patrick, Matt Duckham, and Thomas Wolle. "Decentralized movement pattern detection amongst mobile geosensor nodes." In International Conference on Geographic Information Science, pp. 199-216. Springer, Berlin, Heidelberg, 2008.

Laube, Patrick, Matt Duckham, Mike Worboys, and Tony Joyce. "Decentralized spatial computing in urban environments." In Geospatial Analysis and Modelling of Urban Structure and Dynamics, pp. 53-74. Springer, Dordrecht, 2010.

Laube, Patrick, Todd Dennis, Pip Forer, and Mike Walker. "Movement beyond the snapshot–dynamic analysis of geospatial lifelines." Computers, Environment and Urban Systems 31, no. 5 (2007): 481-501.

Laube, Patrick, Todd Dennis, Pip Forer, and Mike Walker. "Movement beyond the snapshot–dynamic analysis of geospatial lifelines." Computers, Environment and Urban Systems 31, no. 5 (2007): 481-501.

Laube, Patrick. Computational movement analysis. Berlin, Germany: Springer, 2014.

Lee, D-J., Zhengyuan Zhu, and P. Toscas. "Spatio-temporal functional data analysis for wireless sensor networks data." Environmetrics 26, no. 5 (2015): 354-362.

Li, C. H., Kuo, B. C., Lin, C. T., & Huang, C. S. (2011). A spatial–contextual support vector machine for remotely sensed image classification. IEEE Transactions on Geoscience and Remote Sensing, 50(3), 784-799.

Li, Zhenhui, Ming Ji, Jae-Gil Lee, Lu-An Tang, Yintao Yu, Jiawei Han, and Roland Kays. "MoveMine: mining moving object databases." In Proceedings of the 2010 ACM SIGMOD International Conference on Management of data, pp. 1203-1206. ACM, 2010.

Lieske, S.N.; Leao, S.Z.; Conrow, L.; Pettit, C.J. Validating Mobile Phone Generated Bicycle Route Data in Support of Active Transportation. In Proceedings of the SOAC 2017—State

of Australian Cities (SOAC) National Conference, Adelaide, South Australia, 28–30 November 2017.

Lin, B. and Su, J. 2008. “One Way Distance: For Shape Based Similarity Search of Moving Object Trajectories.” GeoInformatica, 12, 117-142.

Lin, Miao, and Wen-Jing Hsu. "Mining GPS data for mobility patterns: A survey." Pervasive and Mobile Computing 12 (2014): 1-16.

Liu, Xinyi, Qunying Huang, and Song Gao. "Exploring the uncertainty of activity zone detection using digital footprints with multi-scaled DBSCAN." International Journal of Geographical Information Science 33, no. 6 (2019): 1196-1223.

Loidl, Martin, Gudrun Wallentin, Robin Wendel, and Bernhard Zagel. "Mapping bicycle crash risk patterns on the local scale." Safety 2, no. 3 (2016): 17.

Long, J.A. and Nelson, T.A., 2013. A review of quantitative methods for movement data. International Journal of Geographical Information Science, 27 (2), 292–318. doi:10.1080/13658816.2012.682578

Lovelace, Robin, Anna Goodman, Rachel Aldred, Nikolai Berkoff, Ali Abbas, and James Woodcock. "The Propensity to Cycle Tool: An open source online system for sustainable transport planning." (2017),Vol 10 No. 1 [2017] pp. 505–528

Lunke, Erik Bjørnson. "Commuters’ satisfaction with public transport." Journal of Transport & Health 16 (2020): 100842.

Lyons, W.; Peckett, H.; Morse, L.; Khurana, M.; Nash, L. Metropolitan Area Transportation Planning for Healthy Communities; No. DOT-VNTSC-FHWA-13-01; John A. Volpe National Transportation Systems Center: Boston, MA, USA, 2012.

Mäenpää, Heikki, Andrei Lobov, and Jose L. Martinez Lastra. "Travel mode estimation for multi-modal journey planner." Transportation Research Part C: Emerging Technologies 82 (2017): 273-289.

Mansfield, T.J.; Gibson, J.M. Estimating Active Transportation Behaviors to Support Health Impact Assessment in the United States. Front. Public Health 2016, 4, 591.

Maricopa Association of Governments. MAG Bike Counts Initiative 2016. Available online: azmag.gov/Portals/0/Documents/BaP_2014-09-16_Item-07_MAG-Bicycles-Count-Project-Presentation.pdf?ver=2017-04-06-110803 (accessed on 6 April 2017).

Marra, Alessio D., Henrik Becker, Kay W. Axhausen, and Francesco Corman. "Developing a passive GPS tracking system to study long-term travel behavior." Transportation Research Part C: Emerging Technologies 104 (2019): 348-368.

Marron, James Stephen, James O. Ramsay, Laura M. Sangalli, and Anuj Srivastava. "Functional data analysis of amplitude and phase variation." Statistical Science (2015): 468-484.

Mayo, Francis L., and Evelyn B. Taboada. "Ranking factors affecting public transport mode choice of commuters in an urban city of a developing country using analytic hierarchy process: The case of Metro Cebu, Philippines." Transportation Research Interdisciplinary Perspectives 4 (2020): 100078.

McKinney, Wes. "pandas: a foundational Python library for data analysis and statistics." Python for High Performance and Scientific Computing 14, no. 9 (2011): 1-9.

Meinshausen, N.; Bühlmann, P. Stability selection. J. R. Stat. Soc. Ser. B (Methodol.) 2010, 72, 417–473.

Mennis, J., & Guo, D. (2009). Spatial data mining and geographic knowledge discovery—An introduction. Computers, Environment and Urban Systems, 33(6), 403-408.

Michelot, Théo, Roland Langrock, and Toby A. Patterson. "moveHMM: An R package for the statistical modelling of animal movement data using hidden Markov models." Methods in Ecology and Evolution 7.11 (2016): 1308-1315.

Miller, H. J., & Goodchild, M. F. (2015). “Data-driven geography.” GeoJournal, 80(4), 449-461.

Miller, H.J., 2005. A measurement theory for time geography. Geographical Analysis, 37 (1), 17–45. doi:10.1111/gean.2005.37.issue-1

Miller, Harvey J., and Jiawei Han, eds. Geographic data mining and knowledge discovery. London: Taylor & Francis, 2001.

Miranda-Moreno, Luis F., and Thomas Nosal. "Weather or not to cycle: Temporal trends and impact of weather on cycling in an urban environment." Transportation research record 2247, no. 1 (2011): 42-52.

Miranda-Moreno, Luis F., Thomas Nosal, Robert J. Schneider, and Frank Proulx. "Classification of bicycle traffic patterns in five North American Cities." Transportation research record 2339, no. 1 (2013): 68-79.

Morales, J., Haydon, D., & Frair, J. (2004). “Extracting more out of relocation data: building movement models as mixtures of random walks.” Ecology, 85(9), 2436–2445. Retrieved from http://www.esajournals.org/doi/abs/10.1890/03-0269

Moudon, A.V.; Lee, C.; Cheadle, A.D.; Collier, C.W.; Johnson, D.; Schmid, T.L.; Weather, R.D. Cycling and the built environment, a US perspective. Transp. Res. Part D Transp. Environ. 2005, 10, 245–261.

Mueller, T., Olson, K. a., Dressler, G., Leimgruber, P., Fuller, T. K., Nicolson, C., … Fagan, W. F. (2011). “How landscape dynamics link individual- to population-level movement patterns: A multispecies comparison of ungulate relocation data.” Global Ecology and Biogeography, 20(5), 683–694.

Murakami, Elaine, David P. Wagner, and David M. Neumeister. "Using global positioning systems and personal digital assistants for personal travel surveys in the United States." In International Conference on Transport Survey Quality and Innovation. 2004.

Nams, V. O. (2014). “Combining animal movements and behavioural data to detect behavioural states.” Ecology Letters, 1228–1237. doi:10.1111/ele.12328

Nehme, E.K.; Pérez, A.; Ranjit, N.; Amick, B.C., III; Kohl, H.W., III. Sociodemographic factors, population density, and bicycling for transportation in the United States. J. Phys. Act. Health 2016, 13, 36–43.

Nelson, T.A.; DenOuden, T.; Jestico, B.; Laberee, K.; Winters, M. BikeMaps.org: A Global Tool for Collision and Near Miss Mapping. Front. Public Health 2015, 3, 53.

Nelson, T., Ferster, C., Laberee, K., Fuller, D., & Winters, M. (2021). Crowdsourced data for bicycling research and practice. Transport Reviews, 41(1), 97-114.

Nelson, Trisalyn A., and Barry Boots. "Detecting spatial hot spots in landscape ecology." Ecography 31, no. 5 (2008): 556-566.

Nelson, Trisalyn, Avipsa Roy, Colin Ferster, Jaimy Fischer, Vanessa Brum-Bastos, Karen Laberee, Hanchen Yu, and Meghan Winters. "Generalized model for mapping bicycle ridership with crowdsourced data." Transportation Research Part C: Emerging Technologies 125 (2021): 102981.

Nguyen, Minh Hieu, and Jimmy Armoogum. "Hierarchical process of travel mode imputation from GPS data in a motorcycle-dependent area." Travel Behaviour and Society 21 (2020): 109-120.

Nordback, K.; Marshall, W.E.; Janson, B.N.; Stolz, E. Estimating annual average daily bicyclists: Error and accuracy. Transp. Res. Rec. J. Transp. Res. Board 2013, 2339, 90–97.

Páez, Antonio, and Kate Whalen. "Enjoyment of commute: A comparison of different transportation modes." Transportation Research Part A: Policy and Practice 44, no. 7 (2010): 537-549.

Palomino, J., Muellerklein, O. C., & Kelly, M. (2017). A review of the emergent ecosystem of collaborative geospatial tools for addressing environmental challenges. Computers, Environment and Urban Systems, 65, 79-92.

Patterson, T. a, Thomas, L., Wilcox, C., Ovaskainen, O., & Matthiopoulos, J. (2008). “State-space models of individual animal movement.” Trends in Ecology & Evolution, 23(2), 87–94. doi:10.1016/j.tree.2007.10.009

Patterson, Toby A., Marinelle Basson, Mark V. Bravington, and John S. Gunn. "Classifying movement behaviour in relation to environmental conditions using hidden Markov models." Journal of Animal Ecology 78, no. 6 (2009): 1113-1123.

Patterson, Zachary, Kyle Fitzsimmons, Stewart Jackson, and Takeshi Mukai. "Itinerum: The open smartphone travel survey platform." SoftwareX 10 (2019): 100230.

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel et al. "Scikit-learn: Machine learning in Python." the Journal of Machine Learning Research 12 (2011): 2825-2830.

Piatkowski, D.P.; Marshall, W.E. Not all prospective bicyclists are created equal: The role of attitudes, socio-demographics, and the built environment in bicycle commuting. Travel Behav. Soc. 2015, 2, 166–173.

Plaut, P.O. Non-motorized commuting in the US. Transp. Res. Part D Transp. Environ. 2005, 10, 347–356.

Pucher, John, and Ralph Buehler. "Safer cycling through improved infrastructure." (2016): 2089-2091.

Purves, R.S., et al., 2014. Moving beyond the point: an agenda for research in movement analysis with real data. Computers, Environment and Urban Systems, 47, 1–4. doi:10.1016/j.compenvurbsys.2014.06.003

Purves, R.S., et al., 2014. Moving beyond the point: an agenda for research in movement analysis with real data. Computers, Environment and Urban Systems, 47, 1–4. doi:10.1016/j.compenvurbsys.2014.06.003

Python Software Foundation. Python Language Reference, version 3.5, http://www.python.org (Accessed on: 31 May 2019).

R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, http://www.R-project.org (Accessed on: 31 May 2019).

Rafiei, D. and Mendelzon, A. 2002. “Efficient retrieval of similar shapes.” The VLDB, The International Journal on Very Large Data Bases, 11, 17-27. DOI= http://dx.doi.org/10.1007/s007780100059.

Rafiei, D. and Mendelzon, A. 2002. “Efficient retrieval of similar shapes.” The VLDB, The International Journal on Very Large Data Bases, 11, 17-27. DOI= http://dx.doi.org/10.1007/s007780100059.

Ramsay, J. O., & Silverman, B. W. (200). Functional data analysis. Springer.

Reddy, Sasank, Min Mun, Jeff Burke, Deborah Estrin, Mark Hansen, and Mani Srivastava. Using mobile phones to determine transportation modes. ACM Transactions on Sensor Networks (TOSN) 6, no. 2 (2010): 1-27.

Reddy, Sasank, Min Mun, Jeff Burke, Deborah Estrin, Mark Hansen, and Mani Srivastava. "Using mobile phones to determine transportation modes." ACM Transactions on Sensor Networks (TOSN) 6, no. 2 (2010): 1-27.

Reis, R.S.; Hino, A.A.; Parra, D.C.; Hallal, P.C.; Brownson, R.C.; Hino, A.A. Bicycling and Walking for Transportation in Three Brazilian Cities. Am. J. Prev. Med. 2013, 44, e9–e17.

Rodríguez, J. P., Fernández-Gracia, J., Thums, M., Hindell, M. A., Sequeira, A. M., Meekan, M. G., ... & Muelbert, M. (2017). Big data analyses reveal patterns and drivers of the movements of southern elephant seals. Scientific reports, 7(1), 112.

Rodríguez, Daniel A., and Joonwon Joo. "The relationship between non-motorized mode choice and the local physical environment." Transportation Research Part D: Transport and Environment 9, no. 2 (2004): 151-173.

Romanillos, Gustavo, Martin Zaltz Austwick, Dick Ettema, and Joost De Kruijf. "Big data and cycling." Transport Reviews 36, no. 1 (2016): 114-133.

Root, R., & Kareiva, P. (1984). "The search for resources by cabbage butterflies (Pieris rapae): ecological consequences and adaptive significance of Markovian movements in a patchy environment." Ecology, 65(1), 147–165.

Roy, Avipsa, and Edzer Pebesma. "A Machine Learning Approach to Demographic Prediction using Geohashes." In Proceedings of the 2nd International Workshop on Social Sensing, pp. 15-20. ACM, 2017.

Roy, Avipsa, Daniel Fuller, Kevin Stanley, and Trisalyn Nelson. "Classifying Transport Mode from Global Positioning Systems and Accelerometer Data: A Machine Learning Approach." Transport Findings (2021).

Roy, Avipsa, Trisalyn A. Nelson, A. Stewart Fotheringham, and Meghan Winters. "Correcting bias in crowdsourced data to map bicycle ridership of all bicyclists." Urban Science 3, no. 2 (2019): 62.

Roy, A., & Kar, B. "Characterizing the spread of COVID-19 from human mobility patterns and SocioDemographic indicators." In *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Advances in Resilient and Intelligent Cities* (2020) :39-48.

Ryus, P.; Ferguson, E.; Laustsen, K.M.; Schneider, R.J.; Proulx, F.R.; Hull, T.; Miranda-Moreno, L. Methods and Technologies for Pedestrian and Bicycle Volume Data Collection; The National Academies Press: Washington, DC, USA, 2014.

Saalfeld, A. (1999). "Topologically consistent line simplification with the Douglas-Peucker algorithm." Cartography and Geographic Information Science, 26(1), 7-18.

Saelens, B.E.; Sallis, J.F.; Frank, L.D. Environmental correlates of walking and cycling: Findings from the transportation, urban design, and planning literatures. Ann. Behav. Med. 2003, 25, 80–91.

Saha, Dibakar, Priyanka Alluri, Albert Gan, and Wanyang Wu. "Spatial analysis of macro-level bicycle crashes using the class of conditional autoregressive models." Accident Analysis & Prevention 118 (2018): 166-177.

Sakurai, Y., Yoshikawa, M., and Faloutsos, C. 2005. “FTW: Fast similarity search under the time warping distance”, In Proceedings of ACM Symposium on Principles of Database Systems, 326-337.

Sallis, J.F.; Conway, T.L.; Dillon, L.I.; Frank, L.D.; Adams, M.A.; Cain, K.L.; Saelens, B.E. Environmental and Demographic Correlates of Bicycling. Prev. Med. 2013, 57, 456–460.

Sallis, J.F.; Floyd, M.F.; Rodríguez, D.A.; Saelens, B.E. Role of built environments in physical activity, obesity, and cardiovascular disease. Circulation 2012, 125, 729–737.

Sallis, J.F.; Saelens, B.E.; Frank, L.D.; Conway, T.L.; Slymen, D.J.; Cain, K.L.; Chapman, J.E.; Kerr, J. Neighborhood Built Environment and Income: Examining Multiple Health Outcomes. Soc. Sci. Med. 2009, 68, 1285–1293.

Sangalli, Laura M., Piercesare Secchi, Simone Vantini, and Valeria Vitelli. "K-mean alignment for curve clustering." Computational Statistics & Data Analysis 54, no. 5 (2010): 1219-1233.

Sauerländer-Biebl, Anke, Elmar Brockfeld, David Suske, and Eric Melde. "Evaluation of a transport mode detection using fuzzy rules." Transportation Research Procedia 25 (2017): 591-602.

Schapire, Robert E., Yoram Singer, and Amit Singhal. "Boosting and Rocchio applied to text filtering." In Proceedings of the 21st annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 215-223. 1998.

Schick, R. S., Loarie, S. R., Colchero, F., Best, B. D., Boustany, A., Conde, D. a, … Clark, J. S. (2008). “Understanding movement data and movement processes: current and emerging directions.” Ecology Letters, 11(12), 1338–50. doi:10.1111/j.1461-0248.2008.01249.x

Schuessler, Nadine, and Kay W. Axhausen. "Processing raw data from global positioning systems without additional information." Transportation Research Record 2105, no. 1 (2009): 28-36.

Schwanen, Tim, and Patricia L. Mokhtarian. "What affects commute mode choice: neighborhood physical structure or preferences toward neighborhoods?." Journal of Transport Geography 13, no. 1 (2005): 83-99.

Semanjski, Ivana, Sidharta Gautama, Rein Ahas, and Frank Witlox. "Spatial context mining approach for transport mode recognition from mobile sensed big data." Computers, Environment and Urban Systems 66 (2017): 38-52.

Shah, Rahul C., Chieh-yih Wan, Hong Lu, and Lama Nachman. "Classifying the mode of transportation on mobile phones using GIS information." In Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing, pp. 225-229. 2014.

Shen, L.; Stopher, P.R. Review of GPS Travel Survey and GPS Data-Processing Methods. Transp. Rev. 2014, 34, 316–334.

Soleymani, Ali, Frank Pennekamp, Somayeh Dodge, and Robert Weibel. "Characterizing change points and continuous transitions in movement behaviours using wavelet decomposition." Methods in Ecology and Evolution 8, no. 9 (2017): 1113-1123.

Solymosi, R.; Bowers, K.J.; Fujiyama, T. Crowdsourcing Subjective Perceptions of Neighbourhood Disorder: Interpreting Bias in Open Data. Br. J. Criminol. 2017, 58, 944–967.

Srivastava, Anuj, Wei Wu, Sebastian Kurtek, Eric Klassen, and James Stephen Marron. "Registration of functional data using Fisher-Rao metric." arXiv preprint arXiv:1103.3817 (2011).

Stenneth, Leon, Ouri Wolfson, Philip S. Yu, and Bo Xu. Transportation mode detection using mobile phones and GIS information. In Proceedings of the 19th ACM SIGSPATIAL international conference on advances in geographic information systems, pp. 54-63. 2011.

Stewart, Benjamin P., Trisalyn A. Nelson, Karen Laberee, Scott E. Nielsen, Michael A. Wulder, and Gordon Stenhouse. "Quantifying grizzly bear selection of natural and anthropogenic edges." The Journal of Wildlife Management77, no. 5 (2013): 957-964.

Stopher, Peter, Camden FitzGerald, and Jun Zhang. "Search for a global positioning system device to measure person travel." Transportation Research Part C: Emerging Technologies 16, no. 3 (2008): 350-369.

Strauss, J.; Miranda-Moreno, L.F. Spatial modeling of bicycle activity at signalized intersections. J. Transp. Land Use 2013, 6, 47.

Strava Metro. (2017). Comprehensive User Guide: Version 5.01. Strava LLC. San Francisco, California.

Strava.com. Strava Metro. Strava, April 2018. Available online: https://metro.strava.com/ (accessed on 28 April 2018).

Suhaila, Jamaludin, Abdul Aziz Jemain, Muhammad Fauzee Hamdan, and Wan Zawiah Wan Zin. "Comparing rainfall patterns between regions in Peninsular Malaysia via a functional data analysis technique." Journal of hydrology 411, no. 3-4 (2011): 197-206.

Sun, Y.; Mobasheri, A. Utilizing Crowdsourced Data for Studies of Cycling and Air Pollution Exposure: A Case Study Using Strava Data. Int. J. Environ. Res. Public Health 2017, 14, 274.

Sur, M., Skidmore, A. K., Exo, K.-M., Wang, T., J. Ens, B., & Toxopeus, a. G. (2014). “Change detection in animal movement using discrete wavelet analysis.” Ecological Informatics, 20, 47–57. doi:10.1016/j.ecoinf.2014.01.007

Teixeira, Inaian Pignatti, Antônio Nélson Rodrigues da Silva, Tim Schwanen, Gustavo Garcia Manzato, Linda Dörrzapf, Peter Zeile, Luc Dekoninck, and Dick Botteldooren. "Does cycling infrastructure reduce stress biomarkers in commuting cyclists? A comparison of five European cities." Journal of Transport Geography 88 (2020): 102830.

Thums, Michele, et al. "How Big Data Fast Tracked Human Mobility Research and the Lessons for Animal Movement Ecology." Frontiers in Marine Science 5 (2018): 21.

Tibshirani, R. Regression Shrinkage and Selection via the Lasso. J. R. Stat. Soc. Ser. B (Methodol.) 1996, 58, 267–288.

Tibshirani, Robert, Guenther Walther, and Trevor Hastie. "Estimating the number of clusters in a data set via the gap statistic." Journal of the Royal Statistical Society: Series B (Statistical Methodology) 63, no. 2 (2001): 411-423.

Ton, D.; Duives, D.; Cats, O.; Hoogendoorn, S. Evaluating a data-driven approach for choice set identification using GPS bicycle route choice data from Amsterdam. Travel Behav. Soc. 2018, 13, 105–117.

Urban Planning and Mobility Department, City of Montréal. (2017, September 05). Se déplacer... et y gagner. Retrieved April 05, 2021, from https://ville.montreal.qc.ca/mtltrajet/

US Census Bureau Geography. Cartographic Boundary Shapefiles—Counties. 1 September 2012. Available online: https://www.census.gov/geo/maps-data/data/cbf/cbf_counties.html (accessed on 6 April 2018).

US Census Bureau, American Community Survey 2015, https://www.phoenixopendata.com/dataset/phoenix-az-demographic-data/resource/6f460cd1-d0aa-4005-aadb-c371772cbd7b . Accessed October 7, 2020.

Van Vugt, Mark, Paul AM Van Lange, and Ree M. Meertens. "Commuting by car or public transportation? A social dilemma analysis of travel mode judgements." European Journal of Social Psychology 26, no. 3 (1996): 373-395.

Vanparijs, Jef, Jelle Van Cauwenberg, L. Int Panis, Etienne Van Hecke, Dominique Gillis, Sidharta Gautama, R. Meeusen, and Bas de Geus. "Cycling exposure and infrastructural correlates in a Flemish adolescent population." Journal of Transport & Health 16 (2020): 100812.

Vlachos, M., Kollios, G. & Gunopulos, D. 2002. “Discovering similar multidimensional trajectories,” ICDE '02: Proceedings of the 18th International Conference on Data Engineering, IEEE Computer Society, 673–684 3

Vrotsou, Katerina, Jimmy Johansson, and Matthew Cooper. "Activitree: Interactive visual exploration of sequences in event-based data using graph similarity." IEEE Transactions on Visualization and Computer Graphics 15, no. 6 (2009): 945-952.

Wagner-Muns, Isaac Michael, Ivan G. Guardiola, V. A. Samaranayke, and Wasim Irshad Kayani. "A functional data analysis approach to traffic volume forecasting." IEEE Transactions on Intelligent Transportation Systems 19, no. 3 (2017): 878-888.

Wang, Bao, Linjie Gao, and Zhicai Juan. "Travel mode detection using GPS data and socioeconomic attributes based on a random forest classifier." IEEE Transactions on Intelligent Transportation Systems 19, no. 5 (2017): 1547-1558.

Wang, Bijun, Yulong Wang, Kun Qin, and Qizhi Xia. "Detecting transportation modes based on LightGBM classifier from GPS trajectory data." In 2018 26th International Conference on Geoinformatics, pp. 1-7. IEEE, 2018.

Wener, Richard E., and Gary W. Evans. "A morning stroll: levels of physical activity in car and mass transit commuting." Environment and Behavior 39, no. 1 (2007): 62-74.

Widhalm, P., Nitsche, P., & Brändie, N. (2012, November). Transport mode detection with realistic smartphone sensor data. In Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012) (pp. 573-576). IEEE.

Winters, M.; Teschke, K.; Grant, M.; Setton, E.M.; Brauer, M. How far out of the way will we travel?: Built environment influences on route selection for bicycle and car travel. Transp. Res. Rec. J. Transp. Res. Board 2010, 2190, 1–10.

Winters, Meghan, Michael Brauer, Eleanor M. Setton, and Kay Teschke. "Built environment influences on healthy transportation choices: bicycling versus driving." Journal of Urban Health 87, no. 6 (2010): 969-993.

World Health Organization (WHO). Global Recommendations on Physical Activity for Health: World Health Organization; World Health Organization (WHO): Geneve, Switzerland, 2010.

Wu, T. F., Lin, C. J., & Weng, R. C. (2004). Probability estimates for multi-class classification by pairwise coupling. Journal of Machine Learning Research, 5(Aug), 975-1005.

Wyner, Abraham J., Matthew Olson, Justin Bleich, and David Mease. "Explaining the success of adaboost and random forests as interpolating classifiers." The Journal of Machine Learning Research 18, no. 1 (2017): 1558-1590.

Xiao, Guangnian, Zhicai Juan, and Chunqin Zhang. "Travel mode detection based on GPS track data and Bayesian networks." Computers, Environment and Urban Systems 54 (2015): 14-22.

Yanagisawa, Y., Akahani, J., and Satoh, T. 2003. “Shape-Based Similarity Query for Trajectory of Mobile Objects.” In Proceedings of the 4th international Conference on Mobile Data Management. M. Chen, P. K. Chrysanthis, M. Sloman, and A. B. Zaslavsky, Eds. Lecture Notes In Computer Science, Springer-Verlag, London, 2574. 63-77.

Yang, C., et al., 2017b. Utilizing cloud computing to address big geospatial data challenges. Computers, Environment and Urban Systems, 61, 120–128. doi:10.1016/j.compenvurbsys.2016.10.010

Yang, X., Stewart, K., Tang, L., Xie, Z., & Li, Q. (2018). A review of GPS trajectories classification based on transportation mode. Sensors, 18(11), 3741.

Yang, Xue, Luliang Tang, Kathleen Stewart, Zhen Dong, Xia Zhang, and Qingquan Li. "Automatic change detection in lane-level road networks using GPS trajectories." International Journal of Geographical Information Science 32, no. 3 (2018): 601-621.

Zhao, Jinfeng, Pip Forer, and Andrew S. Harvey. "Activities, ringmaps and geovisualization of large human movement fields." Information visualization 7, no. 3-4 (2008): 198-209.

Zhao, K., & Jurdak, R. (2016). Understanding the spatiotemporal pattern of grazing cattle movement. Scientific reports, 6, 31967.

Zheng, Yu, Quannan Li, Yukun Chen, Xing Xie, and Wei-Ying Ma. "Understanding mobility based on GPS data." In Proceedings of the 10th International Conference on Ubiquitous Computing, pp. 312-321. 2008.

Zheng, Yu, Yukun Chen, Quannan Li, Xing Xie, and Wei-Ying Ma. "Understanding transportation modes based on GPS data for web applications." ACM Transactions on the Web (TWEB) 4, no. 1 (2010): 1-36.

Zheng, Yu. “Trajectory data mining: an overview.” ACM Transactions on Intelligent Systems and Technology (TIST) 6, no. 3 (2015): 29.

Zhou, H., & Hu, H. (2008). Human motion tracking for rehabilitation—A survey. Biomedical signal processing and control, 3(1), 1-18.

APPENDIX A

WORKFLOWS FOR MODELING FRAMEWORKS IN CHAPTERS 2-4

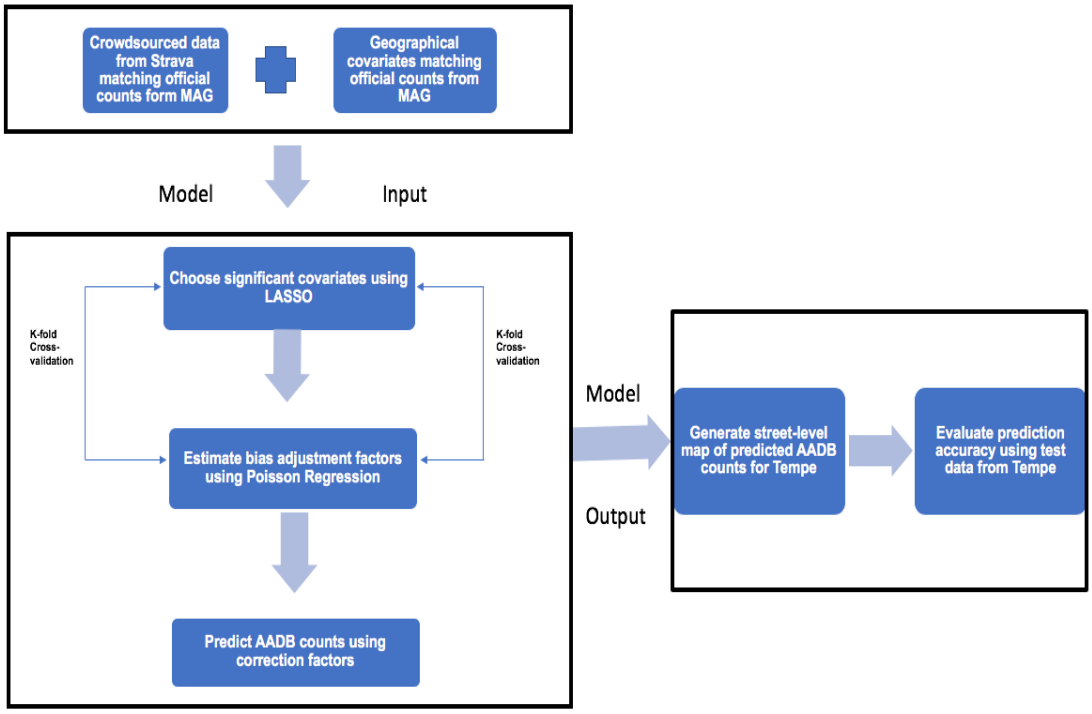


Figure A1: Model design for bicycle ridership prediction using Poisson regression.

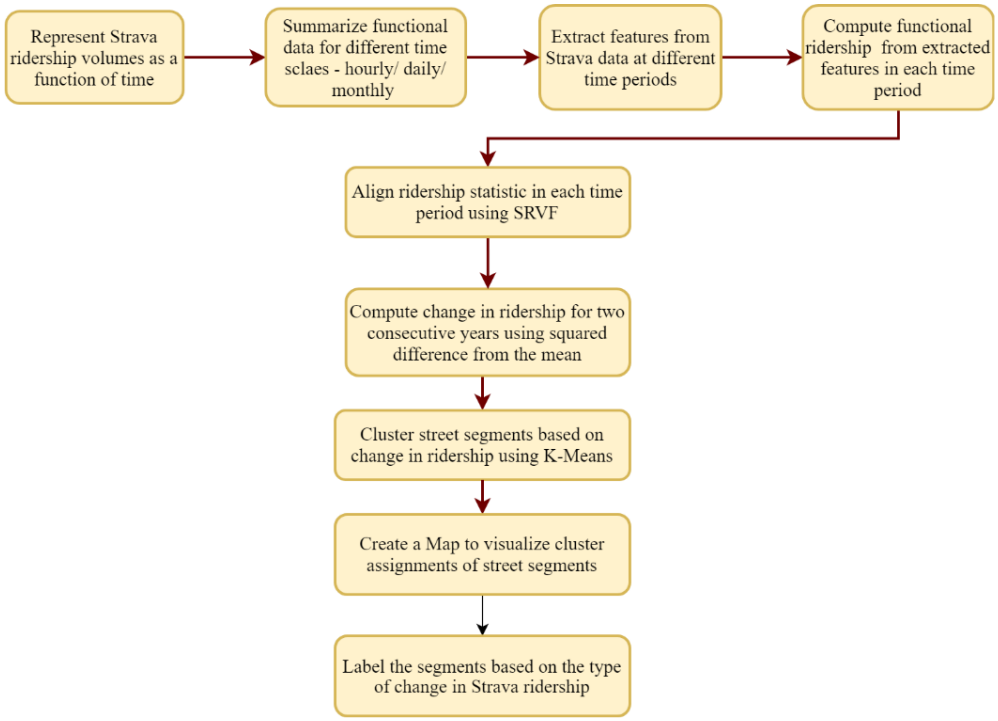


Figure A2: A general workflow for change detection in Strava ridership using functional data analysis

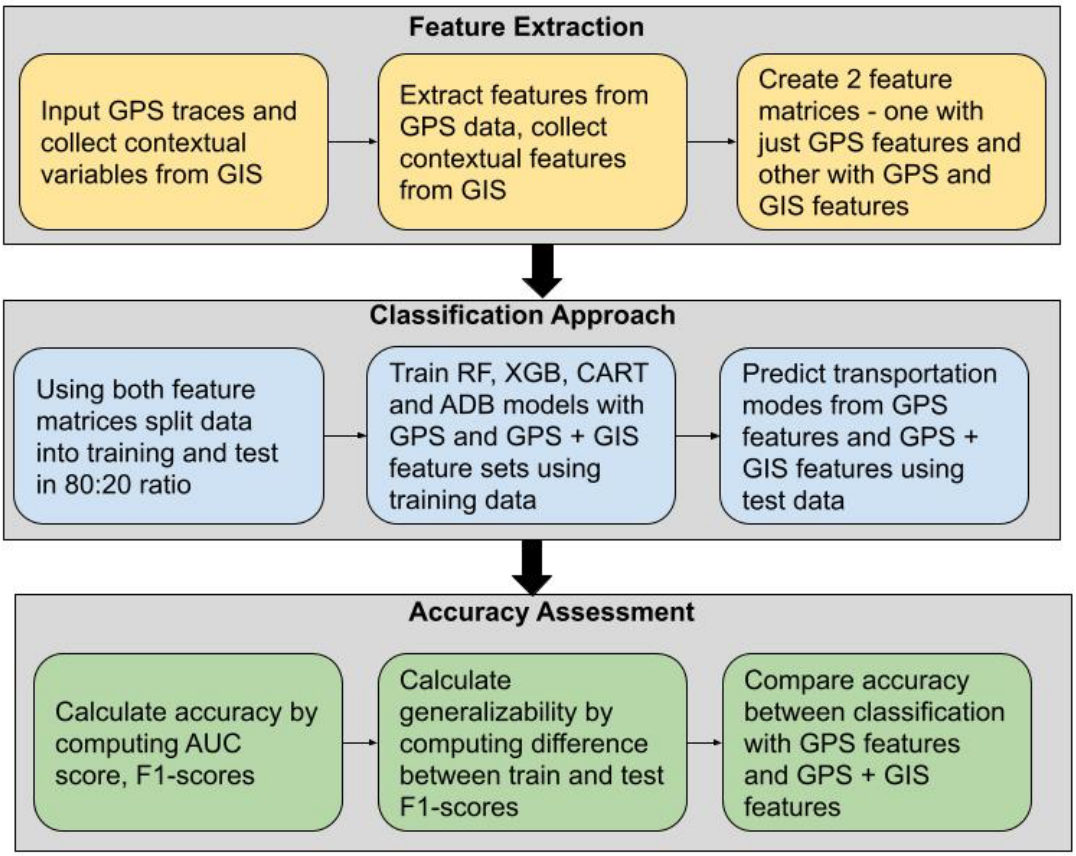


Figure A3: Overall workflow showing transportation mode detection from GPS data and geographic context and assessment of accuracy .

APPENDIX B

ALGORITHMS / CODE FOR CHAPTERS 2-4

Algorithms and code developed as part of the methods described in Chapters 2- 4 are available on Figshare and can be accessed using the following links:

CHAPTER 2: https://doi.org/10.6084/m9.figshare.13171862.v1

CHAPTER 3: https://doi.org/10.6084/m9.figshare.13171862.v1

CHAPTER 4: https://doi.org/10.6084/m9.figshare.13171862.v1

APPENDIX C

CO-AUTHORSHIP STATEMENT FOR CHAPTERS 2- 4

All the manuscripts contained in this dissertation (Chapters 3-5) were co-authored, and the following outlines each of the authors' contributions, as well as that of the doctoral candidate. An appropriate reference conferring the publishing status of each chapter is provided.

CHAPTER 2

Roy, A., Nelson, T. A., Fotheringham, A. S., & Winters, M. (2019). Correcting bias in crowdsourced data to map bicycle ridership of all bicyclists. *Urban Science*, *3*(2), 62.

CHAPTER 3

Roy A, Turaga P and Nelson TA. (2021). Functional data analysis approach for mapping change in time series: A case study using bicycle ridership patterns. *International Journal of Geographical Information Science*, Submitted October 28, 2020, Revised February 12, 2021.

CHAPTER 4

Roy A, Nelson TA, Fuller D and Kedron PJ. (2021). Assessing the role of geographic context in transportation mode detection from GPS data. *Computers, Environment and Urban Systems*, Submitted May 11, 2021.