LINK-FOCUSED PREDICTION OF BIKE SHARE TRIP VOLUME USING GPS DATA: A GIS-BASED APPROACH

LINK-FOCUSED PREDICTION OF BIKE SHARE TRIP VOLUME USING GPS DATA: A GIS-BASED APPROACH

By MATTHEW BROWN, B.Sc. (HONS.)

A THESIS

SUBMITTED TO THE SCHOOL OF GEOGRAPHY AND EARTH SCIENCES

AND THE SCHOOL OF GRADUATE STUDIES

IN PARTIAL FULFILMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

MASTER OF SCIENCE

MASTER OF SCIENCE (2019) McMaster University

(School of Geography and Earth Sciences) Hamilton, Ontario, Canada

TITLE: Link-focused prediction of bike share trip volume using GPS data: A GIS based approach

AUTHOR: Matthew Brown, B.Sc. (McMaster University)

SUPERVIOR: Dr. Darren M. Scott

NUMBER OF PAGES: ix, 80

Abstract

Modern bike share systems (BSSs) allow users to rent from a fleet of bicycles at hubs across the designated service area. With clear evidence of cycling being a health-positive form of active transport, furthering our understanding of the underlying processes that affect BSS ridership is essential to continue further adoption. Using 286,587 global positioning system (GPS) trajectories over a 12-month period between January 1st, 2018 and December 31st, 2018 from a BSS called SoBi (Social Bicycles) Hamilton, the number of trips on every traveled link in the service area are predicted. A GIS-based map-matching toolkit is used to generate cyclists' routes along the cycling network of Hamilton, Ontario to determine the number of observed unique trips on every road segment (link) in the study area. To predict trips, several variables were created at the individual link level including accessibility measures, distances to important locations in the city, proximity to active travel infrastructure (SoBi hubs, bus stops), and bike infrastructure. Linear regression models were used to estimate trips. Eigenvector spatial filtering (ESF) was used to explicitly model spatial autocorrelation. The results suggest the largest positive predictors of cycling traffic in terms of cycling infrastructure are those that are physically separated from the automobile network (e.g., designated bike lanes). Additionally, hub-trip distance accessibility, a novel measure, was found to be the most significant variable in predicting trips. A demonstration of how the model can be used for strategic planning of road network upgrades is also presented.

Keywords: Bike share, Cycling, Active Travel, Eigenvector Spatial Filtering, Network Analysis

Acknowledgements

This thesis would not be possible without the support and contributions of several individuals. Chief among them is my supervisor, Dr. Darren M. Scott, who I would like to thank for readily providing patience, guidance, and support for the past several years of my academic career. I would also like to thank Dr. Antonio Paez for his advice with the spatial statistics components of this work. Thank you to Pat DeLuca for not only teaching me GIS fundamentals, but being someone who has constantly pushed me towards greater opportunities to deepen my educational experience and provide feedback. Thank you to Jay Brodeur and Vivek Jadon from the Maps, Data, and GIS Library for providing data support. Thank you to Esri Canada's Higher Education team member Mike Leahy for providing feedback and suggestions on my research methods several times – your advice was instrumental in my progress. I would also like to extend thanks to staff at the City of Hamilton and SoBi Hamilton for providing quality data that enriched my thesis. Special thanks to Dr. Xun Li and the GeoDa team from the University of Chicago's Center for Spatial Data Science for providing their expertise and making their software freely available for all.

I would like to thank my TransLAB mates: Nosheen Alamgir, Christina Borowiec, Jayden Choi, Samira Hamiditehrani, and Michele Tsang for making time spent in the lab enjoyable and being a constant source of support. I would like to thank former TransLAB member Dr. Ron Dalumpines for his tech support and providing the basis upon which this thesis became possible – the GPS map-matching toolkit. Finally, I would like to thank my

friends (especially Chris, Alex, and Laura) and family for their unwavering support. I could not have done this without you!

Funding for this research was provided by the Natural Sciences and Engineering Research Council of Canada (Grant number RGPIN-2016-06153), awarded to my supervisor. Additional financial support for this thesis was provided by the Social Sciences and Humanities Research Council of Canada through the Joseph Armand Bombardier Canada Graduate Scholarship-Master's award.

I would also like to thank the teams behind the Python and R languages for making their tools available for all. The following R packages were used in this research: `sf` was used for data importing (Pebesma, 2018), `tidyverse` (Wickham, 2017), `dplyr` (Wickham et al., 2019), `rgdal` (R. Bivand et al., 2019), and `spdep` (Bivand, Pebesma, & Gómez-Rubio, 2013) were used in various aspects of data manipulation and calculations. `ggplot2` was used for creating various graphics to further explore data (Wickham, 2016), `caret` (Wing et al., 2019), `leaps` (Miller, 2017) and `MASS` (Venables & Ripley, 2002) were used for stepwise regression modeling for determining important variables and cross-validation. The package `adespatial` was used to acquire the Moran eigenvector maps that the spatial filters were created from (Dray et al., 2019). The packages `spatialreg` (Bivand et al., 2013) and `zeligverse` (Gandrud, 2017) were used to explore different types of statistical models.

Table of Contents

List of Tables

List of Figures

1. Introduction

1.1. Research Problem

Many parts of the world have seen a renaissance in not only cycling as a travel mode, but also in the emergence of bike sharing systems (BSSs) (Pucher, Buehler, et al., 2011; Pucher, Garrard, et al., 2011; Pucher & Buehler, 2008). Before 2008, public bike docking hubs were almost non-existent, which starkly contrasts to the present situation where, as of October 2019, there were over 2100 BSSs operating worldwide (Meddin & DeMaio, 2019; Midgley, 2011). Post-World War II, many countries across the world became reliant on the private automobile as a travel mode. With modern developments in understanding air pollutant health risk factors and advances in environmental monitoring, the negative impacts of motor vehicles and the oil industry have become increasingly well-known. In comparison to cycling, automobiles create air and noise pollution, cause approximately 23-24 million annual injuries worldwide (Feleke et al., 2018), and limit the amount of physical activity needed to maintain a healthy lifestyle. In fact, fear of injury from automobiles has been found to be the most significant barrier to bicycling (Manton et al., 2016). Cycling, on the other hand, has grown to become considered an environmentally friendly, energy efficient, healthy and economical alternative to automobiles (Ryu et al., 2018). Hence, much of the existing bike share literature is focused on the goal of better understanding the role of BSSs to induce a modal shift. However, switching travel modes from automobile to bicycle confers a greater risk of traffic accident and increases exposure to air pollution at the individual level, but the beneficial effects (reduced emissions, increased physical activity, etc.) are generally considered to be more impactful at a societal

level (de Hartog et al., 2010). In terms of influencing modal shifts, past literature generally shows that most of the trips that BSSs replace are those that were previously completed by either walking or via public transportation (Fishman et al., 2015). Therefore, more work needs to be done to assess methods of increasing the attractiveness of BSSs for longer trips and changing the perception that bike share travel is merely a solution to the first and last mile issue of transit connectivity. To successfully establish more BSSs and confer a modal shift, it is imperative that governments, researchers, and BSS operators work together to ease the ability for bikes to replace the automobile. Predictive models, such as those presented in this thesis, help to achieve the goal of furthering the understanding of impacts that different factors have on bike share ridership.

Modern BSSs, such as SoBi (Social Bicycles) Hamilton – the trip data provider for this thesis – are an extremely promising data source in the current climate of bike share research. Given that they are capable of recording information about trip start and end locations (often at bike share hubs) and route GPS tracks, modern BSSs act as a big data solution that can illuminate route details for user trips between the origin and destination. Because of high spatial precision, cycling GPS route data can be used to create models for predicting bicycle traffic, which can answer important questions in several policy contexts. Such models can be important predictors of the efficacy of new cycling infrastructure, especially in a link-based analysis (where a link/edge is defined as an individual road or trail segment that is connected to at least one other segment by a shared junction/node), meaning the effects can be predicted with high precision. Such analysis is of great importance to policymakers when trying to make strategic decisions about developing the

transportation network, or to determine which built and social environment variables play the biggest role in determining cycling traffic. Answers to such questions could ultimately increase BSS ridership and empower users to change their travel mode. As stated on the FAQ page of the SoBi Hamilton website, "Using the anonymous data we collect from the bikes, SoBi Hamilton looks forward to sharing information with the City so they can be better informed as to where cyclists are travelling, how infrastructure can be improved, and where to prioritize bike lanes." (*SoBi Hamilton FAQ*, 2019). This thesis seeks to not only demonstrate the usefulness of collecting bike share GPS information and its applications, but also to provide valuable insights about cycling usage to help achieve the goal of improving sustainable transportation.

1.2. Research Objectives

This thesis presents a GIS-based approach for creating a valid predictive model for assessing cycling traffic at the individual link level using GPS trajectories collected by the SoBi Hamilton bike fleet between January 1st and December 31st, 2018. This thesis seeks to demonstrate a use case for the presented model by predicting the effects of planned infrastructure projects in the City of Hamilton. To satisfy this goal, the following objectives are met:

- Create a comprehensive cycling network for the study area by combining multiple datasets and manually adding new trails revealed by GPS trajectories
- Generate cyclist's actual routes between hubs along the road network by processing GPS trajectories

- Identify the total number of trips occurring on each link in the network using the processed 2018 GPS data
- Conduct regression analysis that controls for spatial effects to predict the number of trips on a link for the study period, including identification and creation of relevant variables such as network features and accessibility to hubs
- Use the model to predict effects on cycling traffic of high priority planned cycling infrastructure projects in the city at the individual link level

By meeting these objectives, this thesis contributes to the existing body of literature on the following topics: bike share, predicting cycling trips, eigenvector spatial filtering, and network analysis. This thesis seeks to fill research gaps in demonstrating both the visualization and predictive power of GPS data for road network analysis. To properly understand the spatial variation of bicycle ridership, it is necessary to make use of data with acceptable spatial detail and temporal coverage (Jestico et al., 2016). Crowd-sourced GPS data from mobile applications (e.g. Strava, Endomondo, MapMyRide, etc.) have been an emerging data source for many recent studies on BSSs. While smartphone popularity has seen a dramatic increase in the past decade, the sample of contributors from crowd-sourced apps tends to be small compared to the actual cycling population in the studied area, meaning it is prone to user selection bias or geographic bias (Romanillos et al., 2016). Crowd-sourced data also lacks the same quality assurance procedures that exist for data sourced from more well-established collectors, such as government agencies (Goodchild & Li, 2012). Many travel survey-based studies have also been done to assess cycling

behavior that collect GPS data as a compliment to detailed trip reporting; however, often as a subset that is more biased towards younger, tech savvy participants (Bricka et al., 2012). Furthermore, survey studies are inefficient and the areas able to be studied by them are limited (Winters et al., 2016). Since findings in this thesis are based on GPS route data for the SoBi user base for an entire year, problems associated with under sampling and user selection bias are alleviated. Furthermore, a hub-trip distance accessibility measure was constructed as an explanatory variable in the modeling process and found to be the most significant explanatory variable. This calculation of accessibility, to the author's knowledge, is a novel way to measure accessibility in the context of the BSS prediction literature. Additionally, this work expands on previous analysis techniques by integrating eigenvector spatial filtering (ESF) as a method of handling spatial autocorrelation in the model residuals.

1.3. Thesis Outline

Including the introduction, this thesis consists of 6 chapters. Chapter 2 contains a literature review on the history of bike share programs and their impacts, GPS big data analysis, the determinants of bike share demand, and how spatial autocorrelation can be accounted for in predictive models. Chapter 3 describes the study area for context and describes the utilized data sources. Chapter 4 explains all the different research methodologies applied to the data including GPS processing, creation of model variables, and specification of the model. Chapter 5 presents the results obtained after research methods were applied, namely the distribution of SoBi trips across Hamilton for 2018 and model results, including a use-case scenario for planned road infrastructure projects. This

section also includes a discussion and interpretation of the results. Chapter 6 summarizes major findings and contributions of this thesis, as well as limitations, assumptions, and recommendations for future areas of research.

2. Background

2.1. History of Bike Share Systems

The history of BSSs, much like many forms of consumer technology, can be broken into distinct generations of growth. As several authors have pointed out, there are four different generations of BSSs that can be identified (Fishman, 2016; Parkes et al., 2013). In 1965, Witte Fietsen (Dutch for "White Bicycles") was introduced in Amsterdam as the first BSS (Davis, 2014). Representing the first generation of BSSs, the premise was simple – white painted bikes that were free for anyone to use. Unfortunately, without any security measures in place, the bikes were stolen and damaged, leading to the program's failure (DeMaio, 2009). Security measures had to be developed for BSSs to become more popular. The second generation of BSSs included a coin-operated payment system, which meant that the bikes had a maintenance budget. However, since cash payment allows for near complete anonymity, theft was a common occurrence (DeMaio, 2009). The third generation of BSSs introduced mandatory credit card or electronic payment to track customers, location tracking devices (e.g., GPS), and stationary docking ports (where users pick-up and drop-off their bikes). These improvements led to reduced theft and made BSSs more resilient for global adoption (Shaheen et al., 2013). The fourth/current generation of BSSs has not been fully realized yet. According to Parkes et al. (2013) the fourth generation could include features such as dock-less systems (for improved user convenience), more

payment options (e.g., public transit smartcard integration), and improved dock designs. Today, most bikes in the BSS fleet are located within China, a trend that is driven by population size and the relatively slow adoption and participation of BSSs in North America (Fishman, 2016). Figure 1 shows the number of bicycles that exist in the bike share fleet across many cities around the world. From this figure, North America is clearly lagging in terms of adoption. This gap between continents is slowly being bridged though, as increased awareness and continued research on the benefits of BSSs and other forms of active travel is done, specifically in North America. These advancements can potentially influence decision makers to upgrade active transport infrastructure, leading to increased attractiveness of these travel modes and increasing adoption.

2.2. Impacts of Bike Share Programs

2.2.1. Improving Health

Many studies attempt to quantify the health impacts of bike share. Several studies have made strides towards quantifying how BSSs improve levels of physical activity and other health-related outcomes (Babagoli et al., 2019; Bauman et al., 2017; Otero et al., 2018). Promoting cycling use is often a controversial task, as automobile users are concerned about competing road space and safety. In fact, the decision to install new bike lanes has been a major issue in municipal elections for many Canadian cities, such as Toronto, Vancouver, and even the study area of this thesis – Hamilton. As such, many studies have sought to establish an understanding of the risks and benefits of promoting cycling. Teschke et al. (2012) explored this issue directly and found that cycling confers significant health benefits compared to automobiles, such that the net risk of mortality is

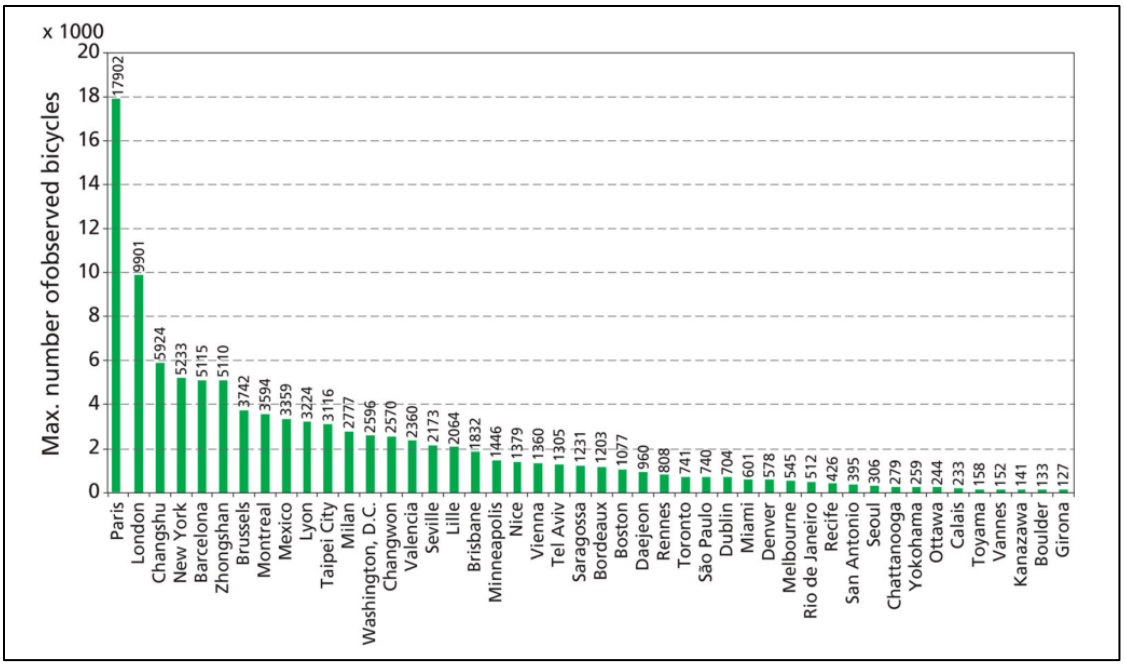

Figure 1: Bike share in cities across the world. Not all cities with bike share are shown.

Note. Reprinted from "Bikeshare: A Review of Recent Literature" by Fishman, E., 2016, *Transport Reviews*, *36*(1), 92–113, p. 95.

outweighed. They also suggested that if enough trips were converted from automobile to active forms of travel, there would be reductions in overall traffic fatalities. To accomplish this, they suggest that North American cities follow models set by some European cities, where bicycle-specific facilities have been created (i.e., separated or protected bike lanes) with promising results in terms of injury reduction. Furthermore, ecological studies have shown that areas with higher levels of active travel correspond to greater amounts of physical activity and lower rates of obesity/diabetes (Pucher et al., 2010). An oft-touted benefit of bikes over cars is the lack of pollution emissions. However, there is evidence that cyclists may experience increased automobile-related air pollutant inhalation (de Hartog et al., 2010). Road transportation comprises a significant portion of the air

pollutants that humans are exposed to in most urban areas across the world (Réquia et al., 2015; Slezakova et al., 2013; Tessum et al., 2014). A comprehensive review of bike share health impacts that focused on physical activity, crashes, and air pollution exposure was conducted by Woodcock et al. (2014). This review modeled health impacts using trip data in London, England and created scenarios where the BSS both did and did not exist to allow for comparison. Models with the BSS were shown to most significantly reduce ischemic heart disease in men and depression in women. London's BSS was shown to have positive health benefits overall, but the benefits were more significant for men than women and for older users than for younger users. The modelled benefits of cycling were found to be much higher for older users as they are more vulnerable to incidence of disease – an effect mitigated by physical activity via bike share. Although older users are more susceptible to injury/fatalities when cycling and have fewer years to live compared to younger users, the beneficial effect bike share had on combating higher disease incidence was found to have a greater effect in the model. Therefore, older users have a better benefit-cost trade-off ratio and were shown to have the largest increase in health benefits for models that altered the age structure of the biking population. For air pollution, the study found that cycling and walking routes have lower average $PM_{2.5}$ (particles with a diameter of $\leq$ 2.5μm) concentrations than for road-based automobile trips, however, this was counterbalanced by higher ventilation rates. Compared to other modes of travel, the impact of using BSSs was found to be small in terms of pollution exposure.

2.2.2. Traffic Injuries

Safety perception is perhaps the leading issue that deters the use of bicycles. Emerging cycling cities still face a significant hamper in their development because cyclists are forced to share the road with automobiles. This has created the association of the cycling experience with a sense of danger and stress, which puts off many potential users (O'Connor & Brown, 2010). Therefore, assessing the safety of bikeshare has received a lot of interest, even from mainstream media, which has sparked a semi-volatile debate between pro-cycle and anti-cycle groups with conflicting research. Graves et al. (2014) found, using hospital injury data from five American cities, that there was a reduction in bicycle injuries after the implementation of BSSs. Their study also recommended that bikeshare operators include helmets for users, although this was criticized by other researchers who showed that studied cities with BSSs had declines in non-head and head injury risk (Teschke & Winters, 2014).

A study by Fuller et al. (2013) examined the likelihood of collisions before and after the implementation of a BSS in Montreal, Canada. The study found that bikeshare users had the same risk level of collision as private cyclists. The study also found that the BSS did not reduce the risk of collision over a two-year sample period, but also stated that the results should be taken with caution as the sample size lacked power. Despite this, the study found overall that the implementation of the BSS conferred an overall modal shift towards active transport.

Fishman & Schepers (2016) conducted a study comparing both cycling injury data between cities with BSSs and those without, and crash data between bike share users and

other cyclists. This research found that when introduced, bike share programs reduce overall cycling injury risk and that bike share users are less likely than other cyclists to sustain fatal or severe injuries. Reasons suggested for this finding were that bike share cycling speeds are slower than other cyclists, providing increased reaction time for both cyclists and motorists to avoid collision. Furthermore, with the introduction of a BSS, driver awareness and cautiousness towards cyclists tends to increase as cycling becomes more common. Also, motorists tend to perceive bike share users as being less experienced and/or tourists, and therefore behave more cautiously in their encounters. Although evidence that bike share may be safer than private cycling exists, it needs to be studied further to understand the underlying mechanics.

2.3. Big Data

The term ‘Big Data’ has become a buzzword in modern scientific rhetoric for its ability to solve the problem of under sampling. Big Data refers to the processing and analysis of extremely large datasets to reveal trends, associations, and ultimately human behavior. Such datasets are curated by several data vendors and can commonly be millions of records. Historically, big data has been formally defined in the framework of the ‘3Vs’: volume (size of the dataset), velocity (how quickly data are generated or collected), and variety (composed from several different sources) (De Mauro et al., 2016). Over time this framework has expanded to include other terms, with veracity (quality of the data) being most important in a research context. While an ideal dataset would exemplify all the Vs, the reality is that most providers only meet the volume criteria. For example, the SoBi Hamilton bikeshare dataset does not meet the “variety” criteria since it gathers data from a

single platform – only SoBi bikeshare users. Furthermore, depending on accuracy of the GPS receiver being used to collect data, additional processing work may need to be done to ensure the veracity criterion is met. In this thesis, the GPS data was processed to filter out invalid trajectories, thus mitigating the issue of veracity (described in section 4.1).

2.3.1. GPS Data

Cycling demand has been an area of intense study over the past decade. As such, researchers have attempted to obtain information about cycling volume and trip counts from multiple different data sources including interviews, household travel surveys, observations in the field, census data, and with recent advances in technology, GPS receivers. With the emergence of smartphones and development of GPS technology, wide scale geographic analysis through big data is becoming a reality. GPS data can be collected using smartphones, or embedded GPS units within the bicycle itself. The data are typically collected by participants contributing to the study or in the context of a specific lifestyle (e.g., using bikeshare as a travel mode to and from work, exercise apps, etc.). Real-time GPS data are possible to collect, but most analysis is done on uploaded user data. GPS trackers typically record location points every 3-5 seconds, creating hundreds or thousands of points for a single trip. GPS data do have accuracy problems though, as locational accuracy can be several meters off, especially with more affordable devices. GPS data analysis for cycling has become mainstream within the past decade, with the first study analyzing cycle mobility being released in 2007 (Harvey & Krizek, 2007). This study explores route choice behavior and found that positional inaccuracies with GPS data require map-matching, a process where GPS points are matched with street infrastructure

to become accurate. Other studies from the early stages of GPS data started trying to use larger and larger samples to try and capture as much variation as possible. One such study was conducted by Menghini et al. (2010), where over 2500 journeys were analyzed. This study estimated a route choice model, but also recognized that by failing to collect any variables pertaining to the actual cyclist or the trips they made meant several simplifying assumptions had to be put in place. Major app companies also collect user GPS data and license this data to researchers. Cintia, Pappalardo, & Pedreschi (2013) conducted a study using GPS tracks from over 30,000 cyclists extracted from the *Strava* API. *Strava* is a popular athletic activity tracker which allows users to upload and share their trips. This study examined user speeds, ride durations, and average heart rates to quantify user training activity. The validity of such apps as a proxy for actual cycling rates has been an area of concern. Studies have examined the correlations between the GPS-tracked commuters in these apps versus actual ridership, and generally found that there is a moderate correlation and amount of explained variation between the two, suggesting they could be an effective proxy (Hochmair et al., 2019; Whitfield et al., 2016).

Data sourced directly from a BSS provider is generally considered to have the highest level of detail and scale for analysis (Romanillos et al., 2016). The detail of such data allows it to be an effective proxy for cycling commuter behavior and allows for thorough examination of different topics. Commonly, studies explore ways to improve efficiency in the placement of bike hubs (e.g. Carlos Garcia-Palomares, Gutierrez, & Latorre, 2012; Park & Sohn, 2017; Soriguera & Jiménez-Meroño, 2020), the factors influencing ridership (e.g., Kutela & Teng, 2019; Mattson & Godavarthy, 2017; Scott &

Ciuro, 2019), and understanding route choice behavior (e.g. Chen et al., 2019; Kou et al., 2020; Wei et al., 2019). Lu et al. (2018) explore cyclist route choice behavior using 161,426 GPS trajectories collected from SoBi Hamilton bikes. This paper found that bike share users are willing to make significant detours in their route compared to the shortest path route in order to have greater access to bicycle facilities and lower traffic. Using data from New York City's BSS, a study found that when trying to predict bike usage during the morning rush, aggregating trips at the neighborhood level gives substantially worse predictions than when looking at individual bike hubs, providing an argument for keeping analysis at the high spatial detail bike share data affords (Ghanem et al., 2017).

User privacy is also a significant concern in GPS data. In many cases, individual user trips are not able to be extracted. Data providers usually aggregate results to preserve anonymity, which means it becomes impossible to know the trip purpose or route choice at an individual level for those datasets. As privacy laws are becoming ever more relevant in mainstream attention (e.g., 2018 Facebook-Cambridge Analytica data scandal), the future of GPS data collection and the level of individual detail that can be extracted is uncertain.

2.4. Determinants of Bike Share Demand

2.4.1. Built Environment

Built environment characteristics such as presence of cycling infrastructure, land-use, street network connectivity, aesthetics, and destination accessibility have consistently been identified as primary indicators of cycling trips (Eren & Uz, 2019; Saberi et al., 2018;

Ton et al., 2019; Zhao & Li, 2017). The type of cycling infrastructure available affects travel behavior in different ways (Buehler & Pucher, 2012). For example, on-street bike lanes are generally more attractive to cyclists than off-road facilities, even if that means a higher travel time (Lu et al., 2018; Tilahun et al., 2007). Using census commuter data from 2000 to 2010, one study found that block groups which had on-road bicycle lanes installed during the study period saw significant increases in cycling commuter traffic compared to block groups which had no infrastructure or just shared lane markings (Ferenchak & Marshall, 2016). Proximity to bike-friendly pathways have also been linked to increased housing prices, suggesting there is an inherent monetary value to having increased cycling infrastructure accessibility (Welch et al., 2016). Guidon et al. (2018) examined the efficacy of electric bicycles and the factors that dictate demand using eight months of transaction data from a BSS based in Zurich, Switzerland. They found that most of the trips were for commuting purposes and had distances that were directly comparable to traditional public transportation modes and taxis. Using spatial regression models, they found that economic activity, social activity, public transportation service quality, and cycling infrastructure availability had the largest impact on demand. Larsen et al. (2013) used a GIS-based approach to create a tool that can be used to determine optimal locations for new cycling infrastructure using cycling data in Montréal. They used grid cells to aggregate several different measures that were found to be related to areas of higher cycling infrastructure need. Namely, they used trip volumes, survey reported “priority” road candidates for infrastructure upgrades, and collision frequency. They also looked at the presence of “dangling nodes”, or places where biking infrastructure ends to identify improvement

candidates. Their results found that Montréal's road network would be significantly improved if roads that increase the connectivity of high traffic areas were updated with biking infrastructure.

In terms of the local context of this research, the decision-making process for SoBi hub placement in Hamilton involved examining housing density, commuter trends, and consultations with both city planners and the public alike (*SoBi Hamilton FAQ*, 2019). Using multilevel statistical models, Scott & Ciuro (2019) explored the factors that influence daily ridership at SoBi hubs in Hamilton. They found that proximity to key areas of the city (i.e., McMaster University and the downtown core) had significant impacts on ridership. Interestingly, the population count between ages 15 - 64 within 200m of the hubs was found to be insignificant, whereas employee counts within the same buffer distance was significant. This implies increased access to activities is a strong motivator for ridership at a hub, but the actual number of people living in the immediate vicinity is not.

2.4.2. Accessibility

Transportation infrastructure systems can be considered as a type of network. An important consideration during network analysis is the underlying spatial characteristics of the network and how they become affected by planning decisions. A key component that dictates active travel usage is the relative accessibility of links in the network for that mode of travel. For example, cycling accessibility is believed to be a major determining factor for transit use (Handy, 2005). Gravity-based measures of accessibility (also called Hansen-type or potential measures) were first defined as "the potential of opportunities for interaction" (Hansen, 1959). Gravity-based measures assume that opportunities are

complementary to each other, and that travel time/distance is a cost that should be minimized or kept within a threshold (Saghapour et al., 2017; Vale et al., 2015). Thus, this type of measure can be applied to forms of active travel, such as cycling and walking (Vale et al., 2015). An example of this would be accessibility to jobs, where being closer to areas with higher employment opportunities corresponds to higher accessibility. Gravity-based accessibility measures typically take the form of the following expression, introduced by Hansen (1959):

$$A_i = \sum_j O_j f(C_{ij}) \quad (1)$$

where A_i is the accessibility of place i, O_j are opportunities found at place j, C_{ij} is the cost of traveling between i and j, and $f(C_{ij})$ is an impedance function (also called distance decay function). The impedance function that is used to weight opportunities can vary, however much of the cycling literature uses travel distance as the impedance function. As a major driving factor behind travel behavior of individuals, many papers have proposed different methods of quantifying gravity-based accessibility and subsequently how to use these methods to understand trends in travel behavior and urban form. As Tobler's first law of geography states, "everything is related to everything else, but closer things are more related than distant things" (Tobler, 1970). Therefore, finding a way to control for spatial effects when measuring accessibility is a key consideration. Scott & Horner (2008) used accessibility indices to investigate if the locations of goods, services, and other opportunities are distributed across American cities in an equitable way for different socio-economic groups. They made use of gravity-based measures in conjunction with a distance

decay model. Interestingly, they found that groups conventionally thought to suffer from worsened accessibility (i.e., low-income households) experienced higher accessibility compared to counterparts in the city tested at the time of the study. Novak & Sullivan (2014) demonstrated a measure for evaluating accessibility to emergency services at the individual link level, as much of the existing literature had previously evaluated accessibility at the scale of nodes or zones in the network, which is inadequate for describing inherently link-based roadway infrastructure (e.g., Xie & Levinson, 2007). Their measure evaluated the system-wide contribution of every link in the network in terms of closeness to critical facilities, topology of the road network, and physical/spatial characteristics of the link, as well as its neighbors.

Due to a surge of popularity over the past two decades, researchers have started making measures specific to cycling. Lowry et al. (2012) introduced the first methodology specifically for measuring bicycle accessibility, termed "bikeability". This measure assesses the entire bikeway network for comfort and convenience by first calculating the bicycle level of service for the entire network, and then using the result in a Hansen-type accessibility model. In a similar vein, Winters et al. (2016) created the "Bike Score®" metric to try and predict the amount of within-city variability of cycling commuters. Bike Score is calculated from a weighted average of 3 different metrics: amount of biking facilities, topography, and connectivity. As this measure was based on existing methodology for a popular walking metric called "Walk Score®", the study found synergy between promoting both walking and cycling as active travel methods, as both shared similar results in terms of which variables had the most predictive power for cycling mode

share. Although walking and cycling have similarities, speed and distance of trips vary significantly between these modes and can be in direct competition with each other when the trip distance is short (Muhs & Clifton, 2015; Wu et al., 2019).

Another way to define the importance of a link in a transportation network is by quantifying how centrally located it is. This idea can take the form of a centrality index by considering the number of shortest paths that connect the nodes within the network that pass-through a given link. This concept of a stress centrality index was first introduced by Shimbel (1953). However, this measure has issues when applied in the context of a transportation network. Namely, not all nodes generate the same number of trips, and, in line with Tobler's first law of geography, as distance increases, the amount of interaction between nodes decreases significantly. Sarlas & Axhausen (2016) addressed both concerns by creating a measure called "Accessibility-weighted centrality". Distance interactions were handled by using a distance decay function that was derived from Halás et al. (2014). In their application of this measure, they used 2010 census work commute trip data for Switzerland and found it to perform significantly better than the unaltered stress centrality measure, providing the best fit of all model comparisons.

Although distance-decay is the most prominent impedance function used in cycling accessibility studies, many studies exist that use different decay formulations. Wu et al. (2019) tested different forms of impedance functions and their ability to predict the number of cycling trips using BSS GPS data from Shenzhen, China. They compared 3 different regression models: one that used accessibility without a distance-decay function, one that used exponential distance decay, and one with a logarithmic normal distribution function

proposed by the authors. They found that their formulation of distance decay performed slightly better than the exponential decay model and suggest that researchers should think closely about the underlying behavior of cyclists when deciding which impedance function to use.

2.5. The "Spatial" Problem

It has been long understood that when dealing with geographic/spatial data, the analyst must account for spatial effects during modeling in order to make valid predictions. In regression analysis, independence of the residuals is a fundamental model assumption that must be met. In geographic data (observations collected from points or regions located in space), it is often the case that measured phenomena experience correlation to nearby observations. This becomes an issue when spatial data is incorporated into a nonspatial statistical model, as it usually results in spatial autocorrelation of the residuals. When spatial dependence is ignored, it can lead to models with coefficient estimation bias, as well as bias in the standard errors (Anselin, 1988b). The problem of spatial autocorrelation has been understood for decades, as Geary (1954) pioneered this work through his study of mapped residuals. The usage of autocorrelation diagnostic tools, such as Moran's I (Moran, 1948), have been implemented in thousands of papers, and the bike share literature is no exception (e.g., McKenzie, 2019).

Over the years, different strategies have emerged on how to handle spatial autocorrelation. Early work on spatial econometric models was pioneered by Anselin (1988), which gave rise to the spatial auto-regression (SAR) model. In the field of econometrics, there are generally two different approaches: "bottom-up" or "top-down",

with the "bottom-up" or "specific to general" approach being predominant in the literature (Larch & Walde, 2008). In the "bottom-up" approach, a model without any spatially lagged variables is created (commonly ordinary least squares linear regression). Next, Lagrange Multiplier tests (Anselin, 1988a) are used to see if a spatial error or lag model fits the data more appropriately. If a statistical test on the residuals to test for spatial autocorrelation has a rejected null hypothesis, then either a spatial lag or error model is specified. For most spatial techniques, a weighting scheme that defines neighbors must be specified, such as taking the inverse distance between observations or through identification of nearest neighbors.

The concept of eigenvector spatial filtering (ESF) was first introduced by Griffith (1978) as a doctoral thesis. Over time, the concept has evolved and been demonstrated to work well in regression analysis for the purpose of removing (i.e., "filtering") spatial effects from the variables in a model (Getis & Griffith, 2002). Using a spatial weights matrix and its associated Moran's I coefficients (MCs), a series of latent map patterns and their corresponding eigenvectors can be derived and used as a proxy independent variable in the model (Cupido et al., 2019). Each eigenvector has a different corresponding uncorrelated map pattern that displays systematic variation, with high and low extreme values corresponding to high and low values of MC for the corresponding spatial weights matrix. Typically, candidate eigenvectors are selected to create a spatial filter that absorbs autocorrelation of the residuals. The resulting spatial filter becomes an independent variable in the model that absorbs autocorrelation of the residuals. ESF has benefits over standard spatial econometric approaches, with the most significant one being that it allows

analysts to use standard linear regression models with OLS, while also controlling for the assumption that residuals are uncorrelated. This benefit is considerable over other approaches because the OLS model is simple (i.e., avoiding the usage of non-standard probability functions), has a well-established theoretical backing, and a litany of diagnostics to aid in interpretation (Griffith, 2017). Chun & Griffith (2014) formally addressed the quality of parameter estimates from regression models using ESF. They found that ESF demonstrated the statistical properties of unbiasedness, efficiency, and consistency. Chun, Griffith, Lee, & Sinha (2016) explored the processes of eigenvector selection. Their paper found that when the candidate eigenvector set was well specified, it can effectively account for spatial autocorrelation. As demonstrated by Paez (2019), ESF can be used as an exploratory tool to identify potentially omitted significant model variables and was argued to be more effective at doing this than examination of model residuals.

3. Data

3.1. Study Area

Hamilton is a densely populated city located in the southern region of the province of Ontario, Canada. Hamilton is located at the westernmost end of Lake Ontario with most of the city, including the downtown core, being near the southern shore. Hamilton Harbour exists at the northern extent, and the Niagara Escarpment bisects the city along the East-West direction, creating upper and lower regions from the sharp elevation change. The downtown core of the city is located entirely below the Niagara Escarpment. As of 2016, the population of Hamilton was 536,917, making it the ninth most populous city in Canada

(Statistics Canada, 2017). Before 2015, the only major form of public transportation existing within the city was busses operated by the Hamilton Street Railway Company. In an effort to complement existing public transit with an affordable and sustainable option, as well as providing greater first and "last mile" connectivity by filling in transit gaps, counsellors at the City of Hamilton voted to spend $1.6 million using a grant from Metrolinx to cover the start-up costs for the SoBi (Social Bicycles) Hamilton Bike Share program (Craggs, 2013). By March of 2015, the program officially launched after passing a trial period starting in January 2015. Since the program's inception, it has continued to grow. The most recent expansion of the service was completed in late 2017 through the "Everyone Rides Initiative", resulting in the creation of new hubs and expansion of the bike fleet (Vize, 2017). SoBi Hamilton bike share works by allowing users to unlock bikes from a network of hubs located across the city at strategic locations. Users choose a payment plan, unlock their bike from the hub, and may ride it anywhere within the official service area. Users end their trips by returning the bike to any of the hubs located in the city and locking it back up. Users can optionally decide to leave their rented bike anywhere in the service area and incur a service fee. In order to use one of the bikes, users create an account using the website or mobile app, and then pay electronically. Each bike is equipped with a GPS receiver to track its location. Real-time public web maps are also available that update how many bikes each hub has at a given moment. Using this hub volume information, SoBi employees gather and redistribute the bikes across the network to accommodate supply and demand. Due to these properties, SoBi Hamilton would be classified as a Generation III bike share program in a transitionary state towards becoming

Generation IV. As of September 2019, there are 132 hubs and approximately 825 bikes in operation.

The official SoBi service area is located primarily in downtown Hamilton, below the Niagara Escarpment, but also extends across to Dundas in the west, covering the entire area of Westdale surrounding McMaster University. There is also a small strip of service area that exists along Van Wagner's Beach. The SoBi Service area and hub network is shown in Figure 2.

3.2. Cycling Network

A cycling network was created for this thesis using both road and trail data to capture accurate cycling routes using the map-matching tool provided as part of Dalumpines and Scott's (2011, 2018) GIS-based Episode Reconstruction Toolkit (GERT). Open data from Hamilton's Open Data Portal was used for the road network and was subsequently enriched with another open dataset containing bikeway information, such as the specific bike lane classification. (*Open Hamilton*, 2018). Private data sets were acquired from the McMaster Library's Maps, Data, & GIS Centre and from the City of Hamilton to further enrich the network with accurate trail features. The CanMap® Content Suite was additionally used to enrich the network with detail at various locations around Hamilton, including more accurate line topology in some areas (DMTI Spatial, 2016). Further, manual digitization of the network was done in high traffic areas (e.g., McMaster University) to create "unofficial" trail features that were commonly used by cyclists, as informed by satellite imagery and raw GPS tracks. The network was then thoroughly inspected to ensure accurate topology and then manually edited accordingly. The 2016

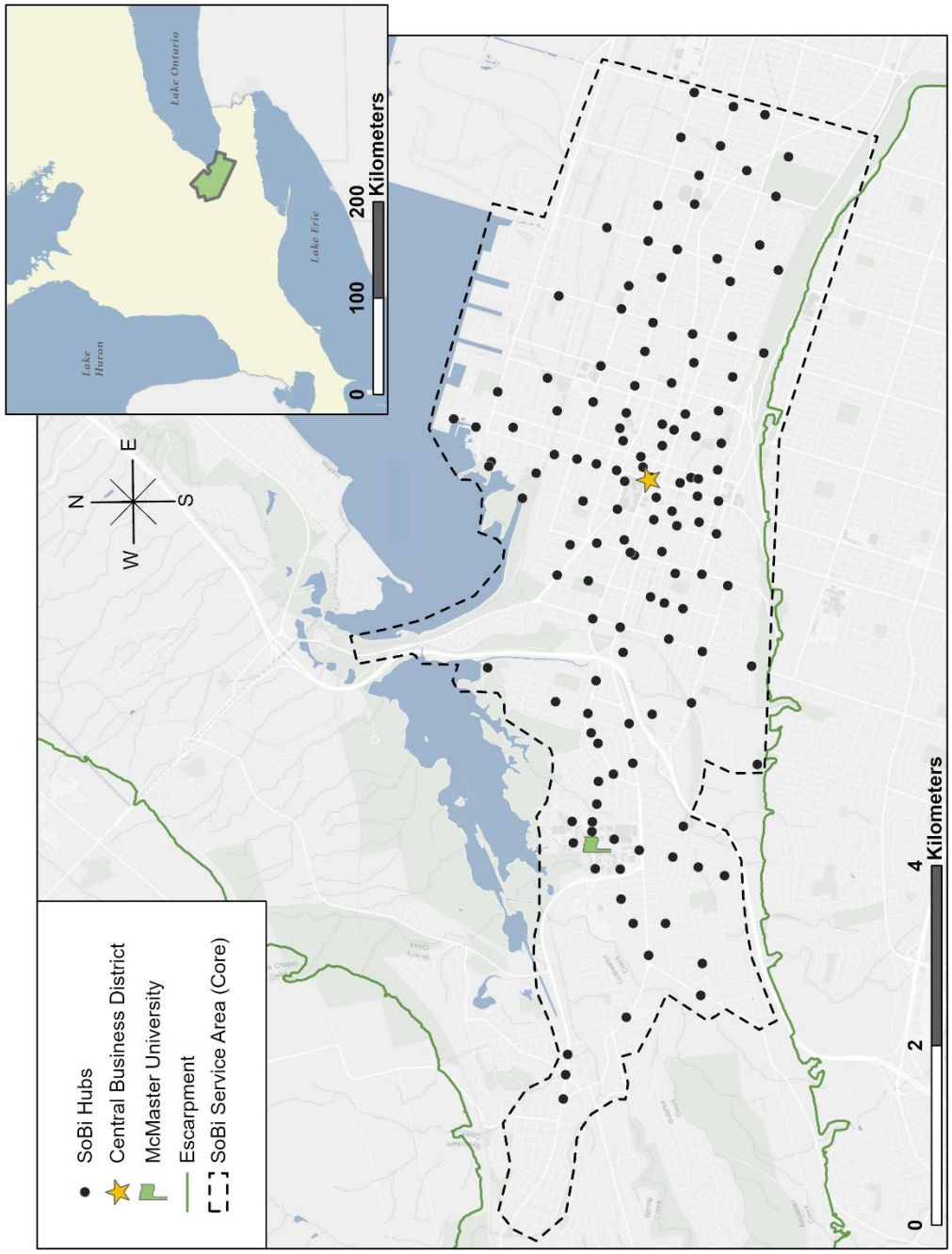

Figure 2: Study Area

Canadian Census and the City of Hamilton's parcel data was used to further enrich the network with population and employment information that was used to test different types of accessibility variables. This process is described in section 4.5.2. Upon completion, the network was converted into a network dataset in order to support network analysis within the ArcGIS® environment. ArcGIS® is a desktop geographic information system (GIS) software developed by Esri. A network dataset is a GIS dataset created within the ArcGIS® Network Analyst framework. Such datasets typically consist of lines representing traffic routes (e.g., roads and trails), junction points (e.g., road intersections), and attributes that are relevant to network analysis such as impedance and flow capacity, as well as topology. For this thesis, the network dataset constructed allowed for bi-directional travel along the network in order to capture trips more accurately, as cyclists are not as strictly constrained to the rules of traffic as automobiles are. The final network dataset consisted of 22,172 links and 16,834 junctions.

3.3. GPS Dataset

As SoBi Hamilton is a third generation BSS that is currently in a transitionary state towards becoming fourth generation, its bike fleet is GPS receiver equipped. This means that each bike can have its XY coordinates tracked in real time. This data is then stored for every trip that occurs from its origin to destination. Despite many past studies using stated and revealed preference surveys, which are cost effective but limited in what information can be collected about trips, this thesis takes advantage of SoBi user GPS data, which reveals the actual routes that users take. To ensure activity from all seasons were captured

and a large sample size was used, GPS data from all trips in the year 2018 were analyzed (January 1st to December 31st, 2018).

GPS trajectories were obtained from SoBi Hamilton. The original 2018 dataset contained 347,079 unique GPS trajectories. Upon processing the data, which is described in detail in the "Methods" section of this thesis, the total number of trips considered in the analysis was 286,587. This study used the map-matching tool provided in GERT to generate individual routes travelled by SoBi riders that exactly follow the underlying cycling network described in section 3.2.

4. Methods

4.1. Extracting Trip Segments from Raw GPS Data

Initially stored as .GPX files, the data required significant processing before it could be analyzed in a meaningful way. To accomplish this, the GIS-based Episode Reconstruction Toolkit (GERT) developed by Dalumpines & Scott (2018) was used. The data were first converted to a useable format (.CSV) using Python. Then, they were updated to have accurate trip time information using accompanying trip episodes with RStudio®. Next, the data were run through a series of GERT modules in ArcMap. First, the GPS Preprocessing Module was used. This module removes invalid points from GPS data using data cleaning algorithms to filter out valid trajectories. From the documentation, invalid points include redundant points (points with the same coordinates), and outliers (points with speed greater than or equal to 50 m/s). Next, the data were run through the TUD-GPS Trip Segments Extraction Module. This module extracts trip segments (sequences of GPS

points in a travel episode) from valid GPS trajectories using start and end times of trip episodes. The output at this stage is a point shapefile for each unique GPS trip segment. This data can be used to visualize travel routes and serves as the input to the map-matching tool discussed later.

4.2. Extracting Network-Matched Route Using GPS Trip Segments

The GIS-based map-matching tool, developed originally by Dalumpines & Scott (2011) and incorporated later into GERT (Dalumpines and Scott, 2018), was used to generate cycling routes from the GPS trip segments along the road network. This tool, developed using the Python scripting language, also generates various route attributes for each input trip segment, including route distance, route directness, number of turns, mean distance between intersections, and the longest road travelled for each specific route. The tool produces route shapefiles by executing a series of steps based on the shortest path algorithm in the ArcGIS® Network Analyst extension. The only inputs required for the map-matching toolkit are GPS trajectories with a start and end point, and an input that can successfully have the shortest path algorithm run on it, such as a network dataset. The general process for map-matching is demonstrated in Figure 3. First, origin and destination points of the trip are identified as stops, and then a polyline feature is generated between the stops and all the GPS points comprising the trip. This polyline represents a hub-to-hub trip (Figure 3a). Next, a buffer is created around the polyline using a default distance specified by the user. The buffer's purpose is to create a series of "barrier" points around the route by performing an intersection with the road network (Figure 3b). The observed route is then created along the road network using the ArcGIS® Network Analyst

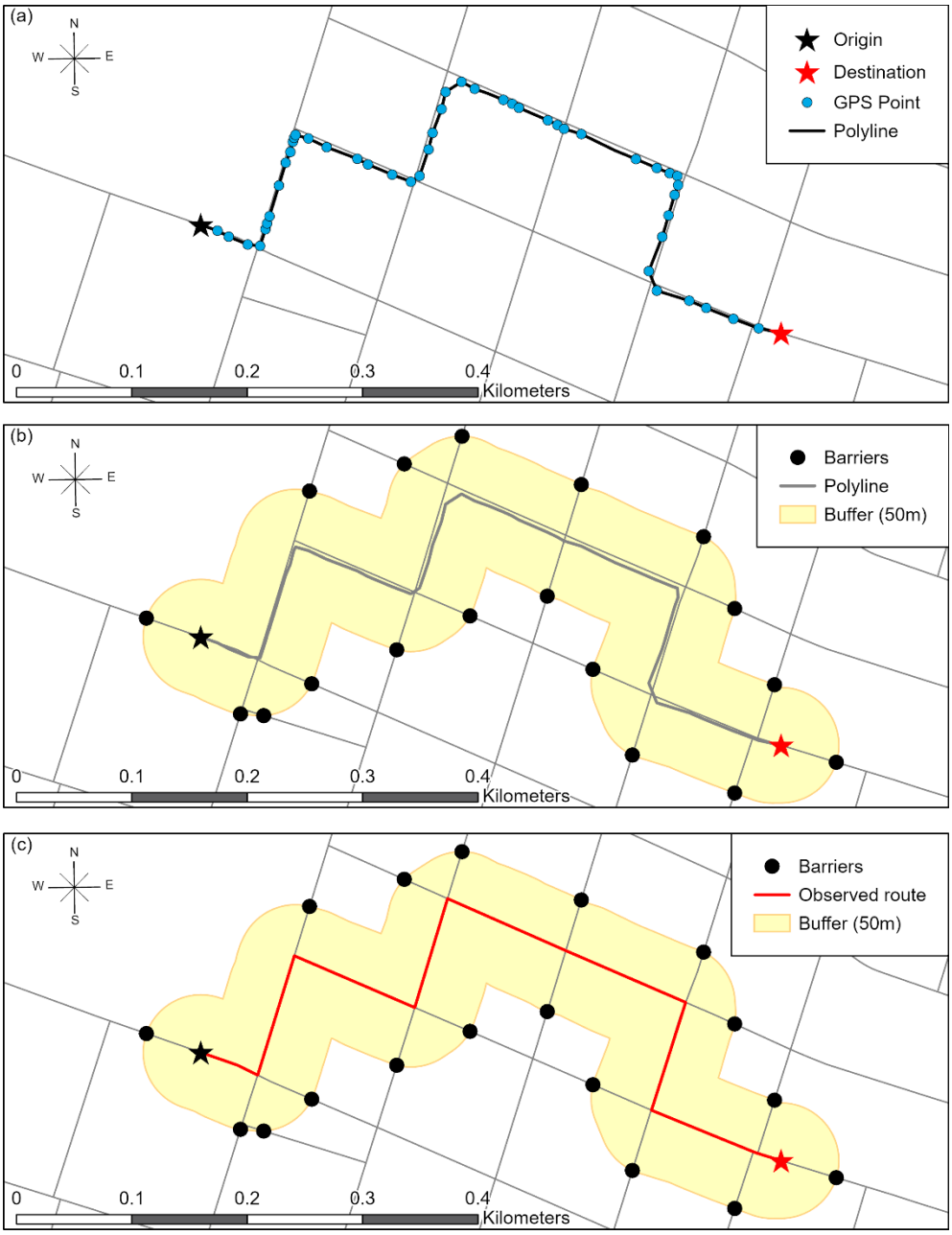

Figure 3: The conversion process from GPS trajectories to an observed route using the map-matching tool

Note. Adapted from "Understanding bike share cyclist route choice using GPS data: Comparing dominant routes and shortest paths", by Lu, W., Scott, D. M., & Dalumpines, R., 2018, *Journal of Transport Geography*, *71*, 176.

extension's shortest path algorithm to generate a route between the two stops constrained by the links in the barriered area (Figure 3c). Dalumpines & Scott (2011) found that a 50m buffer distance produces accurate results for GPS data with a horizontal accuracy of 10m, which is comparable to the quality of the SoBi GPS data. For this reason, 50m was chosen as the default buffer distance for this study.

4.3. Generating Trip Counts

The output generated from the map-matching process was a folder containing Esri shapefiles for each individual route. With such a large volume of data, it is prudent to convert it into a more efficient file format. The data was converted into the Esri Geodatabase format, allowing for improved organization, and decreased storage space requirements. Subsequent processing was also shifted to ArcGIS® Pro which uses Python 3.6 to drastically improve processing speeds. Once converted, each route polyline feature class was individually intersected with the road network. This process "breaks" each route into the network road segments (i.e., links) that comprise it using the network junctions of the underlying road network as breakpoints. Before intersecting, each route has solely a start and end point at the respective beginning and end locations of the trip. After intersecting, there are start and end points at every location where the route polyline overlaps with a junction in the road network. This process is shown in Figure 4. Once this process is complete, each polyline feature is then merged. Since every segment of the intersected route is assigned an ID based on the road network, merging the intersected result together creates a record for every time a route crossed a link in the network. To find

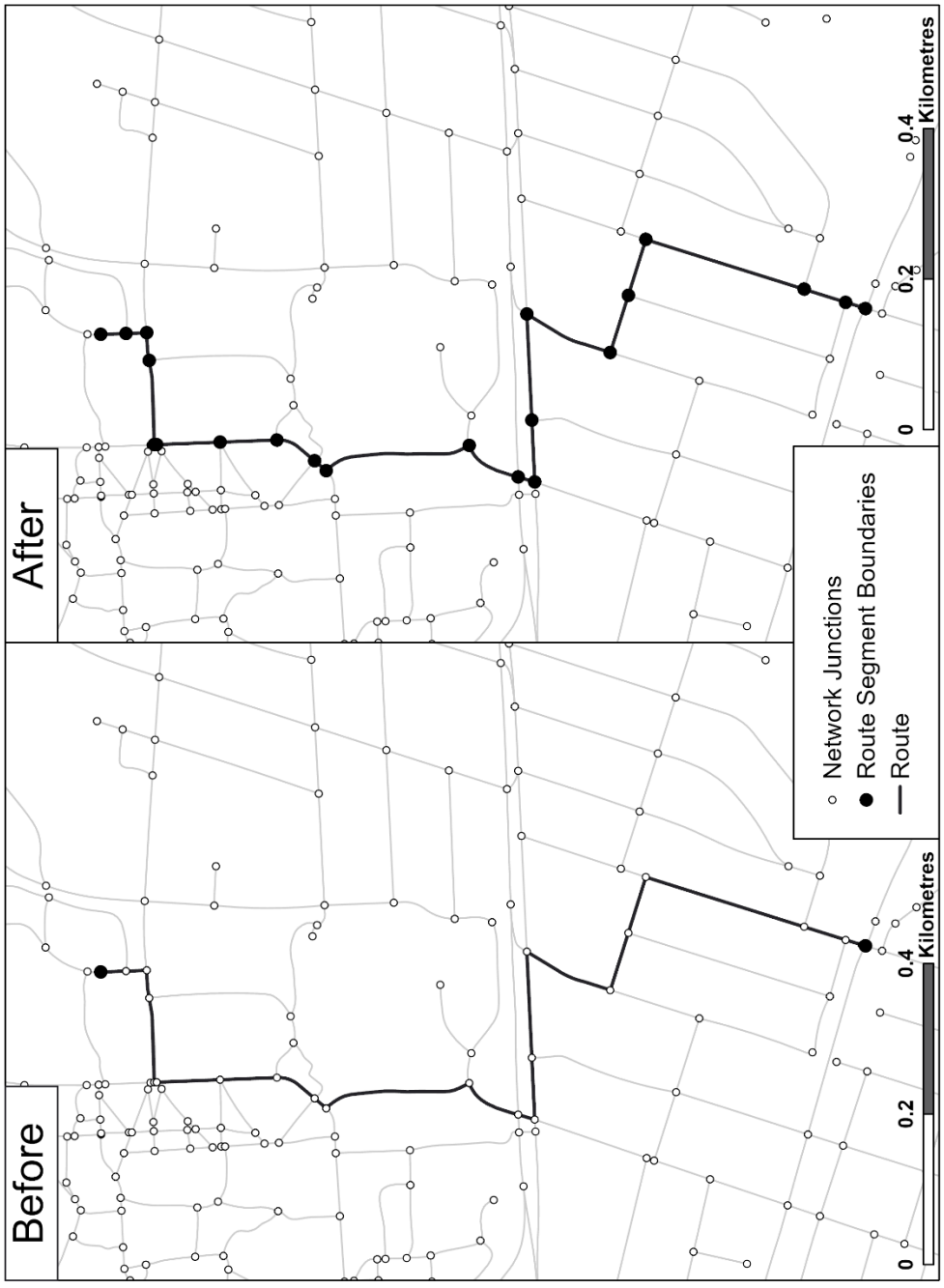

Figure 4: Demonstrating route breakpoints before and after the intersect process

the total number of observed trips on each link, each unique occurrence of a route segment that travelled on each link was summed.

4.4. Creating the Dependent Variable

Regression analysis was used to predict the response variable: 2018 SoBi bicycle traffic (each unique trip) on every link in the cycling network. The original response variable was very positively skewed with a mean of 873 and a median of 33. The minimum value was 1 and the maximum was 29,677. When trips outside of the SoBi service area were removed, the data remained positively skewed, but the average shifted to 1,378, and the median became 320. Thus, the response variable did not follow a normal distribution. As a result, a count-data model (e.g., Poisson, negative binomial, etc.) could be used, or the response variable could be transformed to allow a linear regression model to be valid. The latter option was chosen in this analysis because of the underlying spatial structure of the data and since spatial models for count data are rare in the literature (Guidon et al., 2018). Given that the number of trips on a link is a non-negative variable, the dependent variable was transformed to $\ln(R_i + 1)$, where R_i is the number of unique SoBi trips in 2018 on link i. One was added to this value to ensure that no zeroes are included in the analysis.

4.5. Creating the Independent Variables

4.5.1. Built-Environment

Two categories of built-environment explanatory variables were derived from the road network data. The first was the classification of links as being either major, minor, or

a trail. The second was the bike infrastructure classification. Because these two variables are closely interrelated (e.g., major links often have some type of infrastructure on them), these two variables were combined as a sequence of dummy variables. The reference level for the dummy variables was set to be major roads with no biking infrastructure, as these links would experience high levels of automobile traffic without any safety measures, making them much less attractive to cyclists, as they generally prioritize safer routes when making travel decisions. Table 1 shows the different types of bikeway infrastructure classifications and the number of trips taken by cyclists on each unique link segment for each classification. It also shows the average number of trips, the total number of trips per kilometer travelled, the number of total link segments, and the percentage of links that are located on either a major or minor link segment for each bikeway classification. The trail links are located on neither of these road designations, as they do not experience automobile traffic.

Initially, a sequence of 17 different levels of bikeway classification dummy variables were built. Stepwise regression was implemented to determine which combination of dummy variables resulted in the lowest prediction error. All levels of the dummy variable that were left out of the analysis were insignificant in the model and therefore assumed to be the same statistically as the dummy reference level.

4.5.2. Accessibility

Three different accessibility measures were created and tested in the model based on different criteria: population, employment, and hub-trip distance accessibility. Population and employment accessibility were created using 2016 Canadian census data

Table 1: Bikeway classification statistics

Bikeway type	Total Trips	Mean # Trips	Trips per Km	#Links	%Major	%Minor
Designated BL	2,198,516	4,388.26	41,024	501	**75.05**	24.95
Cautionary Un-Signed BR-HT	74,171	1,765.98	21,608	42	**100**	0
Signed On-Street BR-MHT	13,432	447.73	5,789	30	**100**	0
Unmarked Paved Shoulder BL	342	19	123	18	**100**	0
Cautionary Un-Signed BR-MT	161,509	2,990.91	25,265	54	12.96	**87.04**
Signed On-Street BR-LT	806,677	2,216.15	21,947	364	16.48	**83.52**
Cautionary Un-Signed BR-LT	33,874	294.56	2,952	115	10.43	**89.57**
Trails on McMaster Campus	228,167	1,741.73	27,271	131	0	0
Paved Multi-Use Pathways	305,352	1,411.76	6,911	216	0	0
Unpaved Multi-Use Pathways	23,194	644.28	2,585	36	0	0
Trails in Public Parks	153,355	215.70	4,219	709	0	0
No Infrastructure	2,295,316	460.08	3,935	4989	29.99	**70.01**

Note. BL = Designated bike lane, BR = Bike route, HT = High traffic, MT = Medium traffic, MHT = Medium/high traffic, LT = Low traffic

for the number of people between ages 15 and 64 living in residential areas and the number of people working in employment areas respectively. The population measure was chosen because it represents the amount of people near the hub that might use a BSS. The employment measure was chosen because it acts as a proxy for the attractiveness of an area

near a hub. This data was created using the methodology described in Scott & Ciuro (2019), which is summarized as follows. First, 200m buffers were created around each of the SoBi hubs in the network using ArcGIS® Pro. Then, the data was filtered such that only people between the ages of 15 and 64 were included, as these ages are assumed to correspond to those who most actively use bikeshare. For the population statistic, instead of assuming that people are evenly distributed throughout a dissemination area (DA), the lowest level of geography that data is released from the Canadian Census, the population data was allocated to residential area polygons. The buffers were then intersected with the residential area polygons to create a cross-tabulation of population from DAs based on the proportion of residential area inside the buffer. The cross-tabulation was subsequently aggregated to the hub level. This process was also used for obtaining the working population inside the buffer, but employment areas were used instead of residential. Furthermore, census records were individually aggregated by their 'place of work' at the DA level. The employment and residential area polygons were created using City of Hamilton parcel data. The third accessibility measure, called hub-trip distance accessibility, was derived from SoBi hub trip count data. SoBi keeps track of the number of starting and ending trips at each hub across the city. In this thesis, every unique trip to a hub (either start or end) was considered in the creation of this variable to ascertain the overall usage trend.

Population, employment, and hub-trip distance accessibility were all calculated using the same general process. A methodology by Scott & Horner (2008) was used to create gravity-based accessibility measures using a negative exponential distance decay

impedance function. As described in section 2.4.2, Accessibility is calculated according to the following model:

$$A_i = \sum_j O_j f(C_{ij}) \quad (1)$$

where A_i is the accessibility of place i, O_j are opportunities found at place j, C_{ij} is the cost of traveling between i and j, and $f(C_{ij})$ is an impedance function. In this study, $f(C_{ij})$ is given as $\exp(-\beta C_{ij})$, which is an exponential decrease function controlled by the decay parameter β. Instead of choosing an arbitrary value for β, a value was computed using the unique hub-to-hub trip distances according to the following model:

$$I_k = \alpha \exp(-\beta t_k) \quad (2)$$

where k is the distance category, I_k is the number of trips for category k, and t_k is the trip distance in 100-metre increments for category k. The β value determined using this model was 0.000628, as trip counts sharply decrease as the distance of the trip increases.

After determining the decay function for the accessibility model, centroids were created for every link that had at least one trip in the network. Next, an origin-destination (OD) cost matrix was constructed using ArcGIS® Network Analyst between each link and SoBi hub. This data was then exported into RStudio®, where accessibilities relative to population, employment, and the total number of unique hub trips were calculated using Equation (1). In addition to the accessibility measures, the OD cost matrix was used to create variables for network distance to the closest bus stop and closest hub, as these two factors were assumed to be large motivators of BSS usage for a specific link. Additionally,

the network distance to McMaster University and Hamilton's Central Business District (CBD) were calculated, as these variables were found to be significant in prior research on SoBi Hamilton (Scott & Ciuro, 2019). After examining multiple model diagnostics, including variable correlations and overall significance, the hub-trip distance accessibility variable was determined to be the best predictor of trips out of all tested accessibility measures, as it captured the overall usage trend the best and contributed by far the highest increase in R^2. Looking at the VIF measure, both the population and employment accessibility measures were found to be multicollinear with hub-trip distance accessibility, and as such were dropped from the model. Both the distance to nearest hub and bus stop variables were found to be significant explanatory variables and were kept in the model. Interestingly, although identified as significant variables in a past study of SoBi Hamilton cycling determinants (Scott & Ciuro, 2019), both network distance to McMaster and to the CBD were not as effective predictors of traffic as the hub-trip distance accessibility measure, and thus dropped from the model. This is likely due to the hub-trip distance accessibility variable capturing the spatial trend of increased usage around McMaster University and the downtown core more precisely than simply measuring a base distance to a single point.

4.6. Spatial Predictive Modeling

Two statistical models were created to predict total trips on links in the network. Model 1 includes all links with at least 1 trip and Model 2 includes only links intersecting or inside the official SoBi service area. This choice is discussed more thoroughly in section 5.2. In this thesis, a "bottom-up" approach was used for statistical analysis. First, an OLS

regression model was created and specified for all significant variables. Next, a Moran's I test for residual autocorrelation was conducted. The *p*-value returned was statistically significant, thus rejecting the null hypothesis that the spatial processes promoting the observed values is due to random chance. The returned MC was 0.6, indicating a clustering of similar values. Therefore, the assumptions of an OLS model were violated and spatial effects needed to be controlled for. A spatial weights matrix was used to define spatial relationships between links across the study area. Since the data used is in the form of a road network, commonly used contiguity-based weighting schemes (e.g., Queen's case, Rook's case, etc.) and distance-based schemes (e.g., Inverse distance, *k*-nearest neighbors, etc.) do not sufficiently capture true spatial neighbors for roads. This is because two roads may be very close in proximity, but not actually connect at a shared vertex. Take for example, a bridge running over a highway – even though the two road segments intercept geometrically, there is not connection point for an actual user of the road network to move between them. The choice of weight matrix is a modelling decision, and in this study a first-order neighbor spatial weights matrix is used where links that share a start or end node with a given link are considered neighbors.

Eigenvector spatial filtering was used to mitigate autocorrelation of the residuals to produce a final model. A spatial filter was constructed using a linear combination of candidate eigenvectors through an iterative stepwise regression procedure that is formally described by Le Gallo & Páez (2013). The procedure occurs as follows:

1. Compute eigenvectors using spatial weights matrix and Moran's I

2. Create a series of counter variables to keep track of eigenvector candidates and the spatial filter
3. Select a candidate eigenvector for inclusion in the spatial filter and estimate an OLS model using it
4. If the eigenvector coefficient meets a pre-determined significance level (e.g., *p*-value ≤ 0.10 was used in this study), then combine the existing spatial filter with the selected eigenvector and proceed to the next step, otherwise return to step 3 and update the eigenvector counter
5. Regress the model again using the existing iteration of the spatial filter. Check the MC value, and if it is less than the specified tolerance level (0.5 was used in this study, indicating no spatial autocorrelation) then stop checking candidate eigenvectors, otherwise, update the spatial filter counter and return to step 3
6. The final spatial filter is used as a single explanatory value in the model that removes autocorrelation from the residuals and always has a model coefficient of 1.0.

5. Results

5.1. Data Processing

In terms of total data processed, 286,587 individual trips were successfully map-matched, meaning 60,492 (~17%) of the original GPS trajectories were removed during processing. As described in section 4.1, GIS-based Episode Reconstruction Toolkit (GERT) modules filter out invalid GPS trajectories. In this analysis, GPS extraction through GERT removed 48,693 (~14%) of the GPS trajectories from the analysis. The

remaining 11,799 (~3%) trips were lost during the map-matching process. This loss could have been due to issues with network topology (including specified buffer size during map-matching), or GPS errors that were unsuccessfully filtered out during the extraction process. Figure 5 shows the distribution of trips by month. Most trips occur between the months of May and October, which corresponds to the warmest months in the study area. This is consistent with research showing a positive relationship between warmer weather and BSS usage (Miranda-Moreno & Nosal, 2011; Tin Tin et al., 2012). Although most students at McMaster University are not present in the study area from the end of April to the beginning of September, this is the period with the highest ridership. This suggests that although student populations are significant users of the BSS, they are not necessarily the core users.

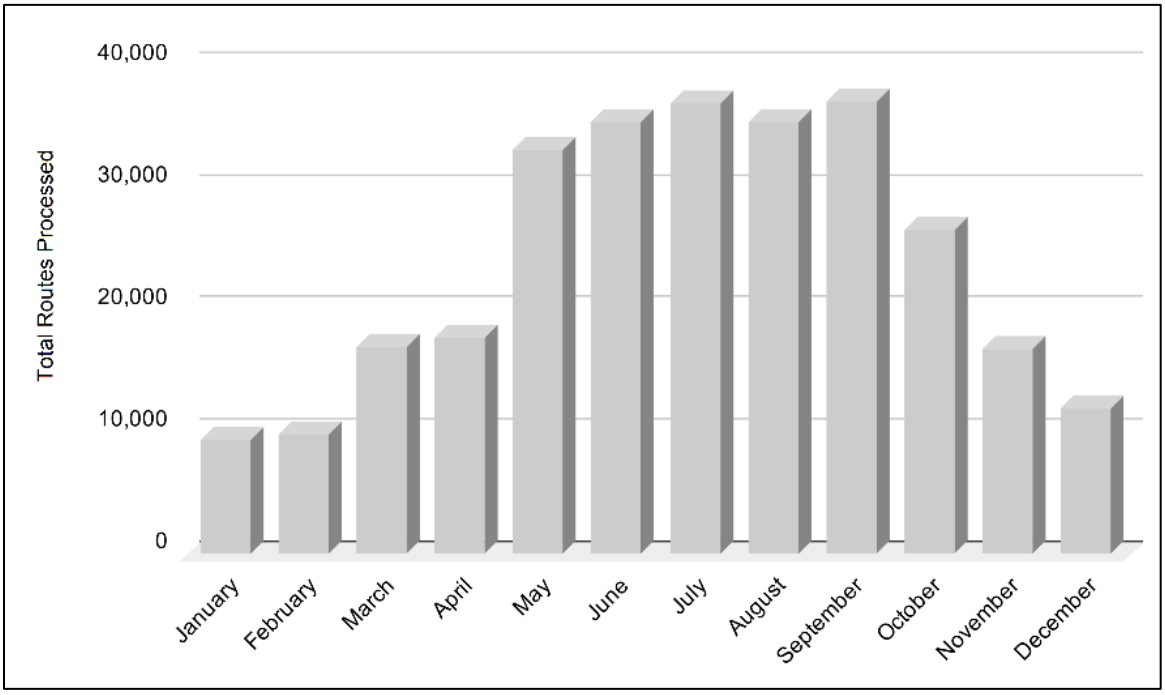

Figure 5: Frequency of processed SoBi trips by month, 2018

5.2. Trip Counts Across Hamilton for 2018

7205 links in the network were observed to have at least 1 SoBi bicycle trip in Hamilton during 2018. Official SoBi service areas have been defined and made publicly available by the City of Hamilton. For clarity, trip count results are discussed with respect to these boundaries, with links divided into categories of either being inside or outside of the service areas. Table 2 highlights how the number of unique trip counts by link differs depending on its location.

Table 2: Breakdown of the unique number of trips across the network, 2018

Division of Links	# of Links	Mean # Trips	Median # Trips
Entire Dataset	7,205	873	33
Inside SA	4,553 (63%)	1,338	319
Outside SA	2,652 (37%)	7	2

From the table, there is a discrepancy in how much SoBi riders used the road network outside of the city-defined service areas, as the average and median number of trips declines significantly. This demonstrates that riders who tend to go outside of the service areas are much fewer in number. Because almost 37% of trips occurred on links outside of the service areas, riders that choose to do so appear to be making extensive use of the network, meaning these trips could be longer in general and not necessarily utilitarian in purpose. From this finding, the modeling process was divided into two separate streams –

one that considers all trips successfully map-matched in 2018, and another that removes all links outside of the service areas. A map showing the location of all links with at least one recorded trip relative to the service areas is shown in Figure 6. In terms of trips occurring outside of the service areas, they occur mostly in areas above and below the Niagara Escarpment in the East Hamilton area.

Figure 7 depicts unique SoBi trip counts for individual links across the city in 2018. In this visualization, the general trends of cycling usage for 2018 are revealed. SoBi riders tend to make higher use of links that connect them to areas near McMaster University and the downtown core. There is a distinct spatial pattern of usage in latitudinal directions (east/west), which is heavily affected by the presence of the Niagara Escarpment, which acts as a significant physical barrier to cycling travel because of the sharp elevation increase. This map reveals that users typically do not make use of these bikes for large numbers of recreational trips, as popular attractions such as Bayfront and Pier 4 Park all contain mostly links with fewer than 966 trips.

. Table 3 shows the top 5 roads in the city based on the total number of unique trips across all segments of the road. All roads in this table are considered major roads by the City of Hamilton and contain several segments with cycling infrastructure, the main category being designated bike lanes. All the roads in this table are in central locations of Hamilton's downtown core, corresponding to areas of high employment opportunities and population, except for one. The exception, Sterling Street, is in a residential student area and serves to connect the center of McMaster University's campus to King St. Student use in this area has significantly increased usage rates, making it the most popular link.

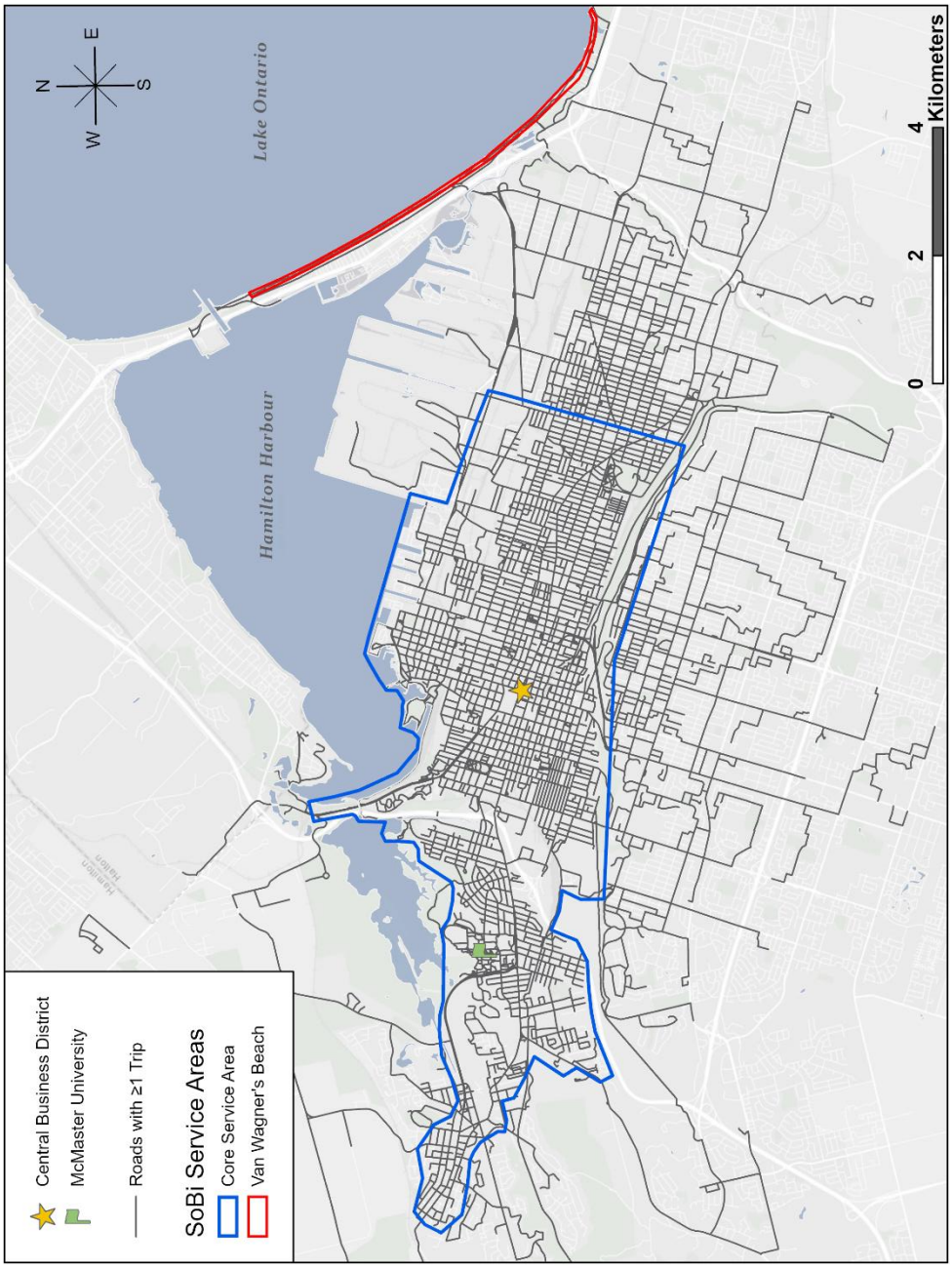

Figure 6: Locations of SoBi trips relative to official service are

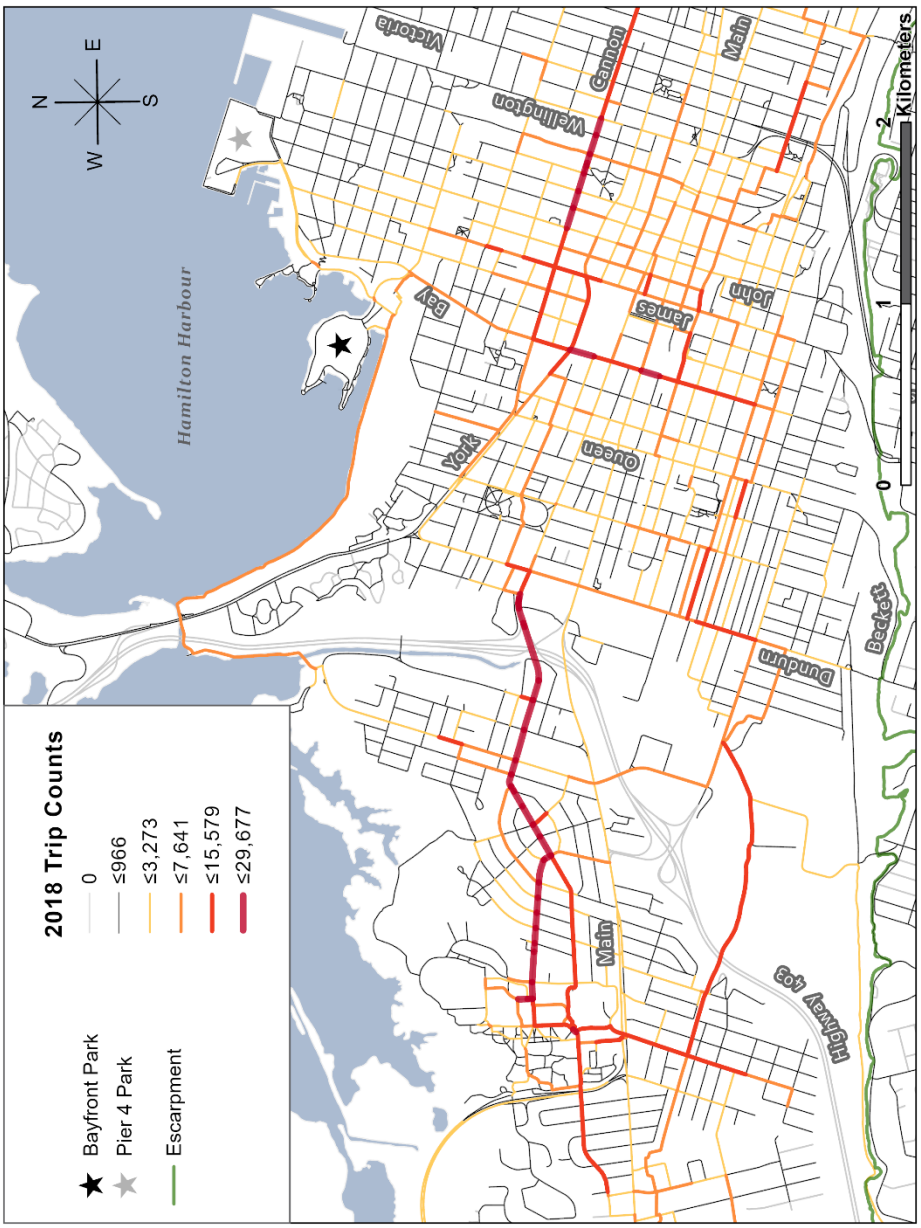

Figure 7: Cycling trip counts across the city. The darker the red and thicker the line is, the more unique trips occurred on it

Table 3: Top 5 highest SoBi traffic roads in Hamilton, 2018

Rank	Name	# Segments	Unique Trips	Mean Trips	Infrastructure
1	King Street	210	581,416	2,769	BL, SBR, HCOS
2	Cannon Street	80	543,386	6,792	BL
3	Bay Street	39	273,660	7,017	BL, SBR
4	Sterling Street	11	255,476	**23,225**	BL, SBR
5	Main Street	192	243,069	1,266	BL, HCOS

Note. The bolded value in the "Mean Trips" column highlights how Sterling Street has a significantly higher average trip value than other high traffic roads. BL = Designated bike lane, SBR = Signed bike route, HCOS = High traffic cautionary un-signed bike route

5.3. Model Specification and Results

Two OLS regression models were estimated with the use of eigenvector spatial filtering (ESF) to predict the log-transformed number of bicycle trips on each link in the network for 2018 – one for all links with at least one trip (Model 1), and one for only links inside the officially designated SoBi Hamilton service areas (Model 2). Explanatory/independent variables that were removed from the model were found to either be multicollinear with other variables (through examination of VIF), or have general model insignificance (i.e., p-value > 0.05). Furthermore, after filtering spatial autocorrelation, some levels of the bike infrastructure dummy variable became insignificant, implying that they may have only been significant in the original model due to autocorrelation. Model 1

and 2 coefficient estimates/summary statistics are presented both before and after the implementation of a spatial filter for comparative purposes. These results are shown in Tables 4 and 5 respectively.

Minor roads with a signed on-street bike route were found to be significant in Model 2, but not in Model 1. This is likely explained by the removal of several links from this classification that had a low number of trips in peripheral areas. The average number of trips on this classification inside the service areas was 3,320. When links outside of the service areas were included, the average decreased to 2,501, demonstrating a substantial usage decrease for this road classification outside of the service areas. This may explain why minor roads with signed on-street bike routes was a significant predictor in Model 2, but not in Model 1. In both model scenarios the Moran's I coefficient (MC) indicated that spatial autocorrelation of residuals was present in the model before implementation of the spatial filter. In both cases, once the spatial filter was included, the new MC indicated that the null hypothesis (i.e., the spatial processes promoting the observed pattern of values due to random chance) could not be rejected. Thus, eigenvector spatial filtering alleviated issues of autocorrelation in both models.

For conciseness, only the Model 1 (filter applied) output will be interpreted since it is the more inclusive form of the model in terms of network coverage. However, it should be noted that Model 2 (filter applied) performed well, with a high adjusted R squared value (0.78), full explanatory variable significance at the $p < 0.001$ level, and reasonable standard error. The adjusted R squared value of Model 1 was 0.89, indicating approximately 89% of the variance in the response variable can be explained by the explanatory variables.

Table 4: Regression output for bike trip prediction by link (Model 1)

Dep. variable: log (bike trips + 1)	**No Filter**		**Filter Applied**	
	Coef.	*t* statistic	Coef.	*t* statistic
(Intercept)	3.642	59.957***	3.789	105.572***
Spatial Filter	-	-	1.000	115.942***
Hub-Trip Accessibility ($\times 10^{-3}$)	0.026	57.631***	0.026	94.417***
Net. Dist. to Closest Hub ($\times 10^{-3}$)	-0.724	-37.689***	-0.780	-68.505***
Net. Dist. to Closest Bus Stop ($\times 10^{-3}$)	-0.234	-3.502***	-0.329	-8.342***
Cycling Infrastructure on Roads (Reference: Major Roads with No Infrastructure)				
Major Road with BL	0.680	7.890***	0.928	18.202***
Minor Road with BL	1.077	7.529***	1.126	13.295***
Trails on McMaster Campus	-0.981	-6.960***	-1.438	-17.226***
Paved Multi-Use Pathways	0.599	5.377***	0.306	4.645***
Trails in Public Parks	-1.897	-27.741***	-1.917	-47.494***
Minor Road with No Infrastructure	-1.072	-25.004***	-1.089	-43.205***
Major Signed On-Street BR-MHT	-1.553	-5.495***	-	-
Major Unmarked Paved Shoulder	-1.103	-3.027**	-	-
Observations	7205		7205	
Adjusted R^2	0.68		0.89	
Residual Std. Error	1.535 (df=7193)		0.909 (df = 7194)	
F-Statistic	1388*** (df = 11; 7193)		5188*** (df = 10; 7194)	
Moran's I Coefficient	0.612		0.004	

Note. Hub-Trip Accessibility and Network Distance to Closest Hub/Bus Stop were scaled by a factor of 10^{-3} to improve coefficient interpretation.

*** $p < 0.001$, ** $p < 0.01$

Table 5: Regression output for bike trip prediction by link (Model 2)

Dep. variable: log (bike trips + 1)	**No Filter**		**Filter Applied**	
	Coef.	*t* statistic	Coef.	*t* statistic
(Intercept)	5.382	61.052***	5.538	91.306***
Spatial Filter	-	-	1.000	71.081***
Hub-Trip Accessibility ($\times 10^{-3}$)	0.017	28.852***	0.016	39.436***
Net. Dist. to Closest Hub ($\times 10^{-3}$)	-2.129	-29.733***	-2.238	-43.863***
Net. Dist. to Closest Bus Stop ($\times 10^{-3}$)	-0.594	-4.388***	-0.701	-7.510***
Cycling Infrastructure on Roads (Reference: Major Roads with No Infrastructure)				
Major Road with BL	1.062	9.035***	1.133	14.005***
Minor Road with BL	1.123	7.110***	1.178	10.821***
Trails on McMaster Campus	-1.110	-7.325***	-1.325	-12.685***
Paved Multi-Use Pathways	1.003	7.053***	0.477	4.851***
Trails in Public Parks	-2.243	-26.103***	-2.251	-38.113***
Minor Road with No Infrastructure	-1.215	-19.027***	-1.201	-27.503***
Minor Road with Signed BR	0.255	2.152*	0.280	3.433***
Major Signed On-Street BR-MHT	-1.631	-5.332***	-	-
Observations	4551		4551	
Adjusted R^2	0.54		0.78	
Residual Std. Error	1.588 (df = 4539)		1.096 (df = 4539)	
F-Statistic	481*** (df = 11, 4539)		1464*** (df = 11; 4539)	
Moran's I Coefficient	0.477		0.004	

Note. Hub-Trip Accessibility and Network Distance to Closest Hub/Bus Stop were scaled by a factor of 10^{-3} to improve coefficient interpretation.

*** $p < 0.001$, * $p < 0.05$

Furthermore, all explanatory variables in the model achieved significance at the 0.001 level. Before applying the spatial filter, Model 1's R^2 value was 0.68, demonstrating a modest, but not substantially large increase in model fit solely due to the filter. Since only the dependent/response variable in this model was log-transformed, interpretation of model coefficients is done by exponentiating the coefficient, subtracting one from the result, and then multiplying by 100. This yields a percent increase/decrease in the dependent variable for every one-unit increase in the independent variable.

Hub trip distance accessibility and the two network distance variables were scaled by 1000 in order to improve coefficient interpretation and standardize observation values. In terms of hub-trip distance accessibility, for every one-unit increase in hub-trip distance accessibility, the response variable is observed to increase by approximately 3%. Since accessibility is a unitless measure, it is difficult to quantify exactly what a one-unit value increase is in concrete terms. However, this result shows that as accessibility of a link increases with respect to hub trips, so does the number of trips on the link itself. It is important to note that the hub-trip distance accessibility variable was by far the most substantial predictor in the model. Before applying the spatial filter, this variable alone was able to achieve an R^2 value greater than 0.5. Since the network distance measures were originally calculated in units of meters, the scaling effect converts them into kilometers. Understandably, as the network distance to both the closest hub and closest bus stop increases, the number of trips decreases (-54% and -28% per unit increase (one kilometer) respectively).

For the categorical/dummy road infrastructure variables, interpretation is taken relative to the reference level - major roads with no biking infrastructure. Thus, there is approximately a 153% increase in link trips when the link changes from the reference level to a major road with a bike lane. Minor roads with a bike lane demonstrate a 208% increase in link trips compared to the reference level. The larger increase observed on minor roads compared to major roads can be explained by cyclists wanting to avoid high traffic areas in the network for safety reasons, while still benefiting from the bike lane infrastructure. Minor roads with no infrastructure have a -66% decrease in link trips compared to the reference level. This is likely explained by such roads having poor access to key trip attractors (e.g., high population/employment areas) that major roads generally do have. Trails on the McMaster University campus and in public parks across the study area also have negative coefficients, corresponding to a -76% and -85% decrease in link trips compared to the reference level. The trails around McMaster University's campus are narrow and highly used by pedestrians (i.e., students going to and from classes), making them less desirable for cyclists, who would more likely follow the road infrastructure, which has restricted access to all but university vehicles and public transportation, making them very low traffic routes. SoBi users also typically use the bikes for utilitarian trips such as commuting, so unless a city park link conferred a useful shortcut, it is not unreasonable to expect these to also attract fewer trips than a major road with no infrastructure. Finally, paved multi-use pathways were observed to have a 36% increase in trips compared to the reference. Links belonging to this classification are trails that are physically separated from vehicular traffic by an open space or barrier, and usually shared with pedestrians or other

non-motorized users. Due to the attractiveness and safety such links confer to cyclists, the positive coefficient is expected. The following bikeway classifications were not found to be significant in either model:

- Major and minor links with cautionary un-signed bike routes for all traffic levels (low – high)
- Major roads with signed on-street bike routes
- All links with paved shoulders

This observation is interesting because it suggests that these types of bikeway classifications do not necessarily translate to a significant effect on increasing trips. A commonality between all the bikeway infrastructure variables with positive coefficients is the presence of a physical barrier or space between them and the actual roadway. Conversely, all the insignificant bikeway infrastructure types are ones that exist directly on the roadway where vehicular traffic travels. This result supports the idea that having increased safety via physical separation or barriers increases cycling traffic. As this thesis is concerned with modeling flows on the road network, unpaved multi-use trails were removed from the analysis. Furthermore, unpaved multi-use trails are mainly in peripheral areas of the network, few in quantity, and used mostly by recreational cyclists. These properties make them an impractical addition to the model.

Model predictions were checked for validity using a test-train framework. Specifically, a repeated *k*-fold cross-validation technique was used. This means the data was randomly split into *k*-subsets (10 were used in this thesis), and then using one subset

at a time, the model was trained and tested to record prediction errors. This process was repeated 3 times. The average value of the prediction errors for each subset was then taken to get the Root Mean Squared Error (RMSE). The RMSE for the testing set, when rounded, was identical to the training model, indicating that the model has predictive value when tested out of the sample. An RMSE of 0.909 was observed, meaning that predictions were larger or smaller than the observed value by a factor of $e^{0.909} \approx 2.48$ (since the transformation was done using the natural logarithm). Although this level of error is perhaps too large for specific operational uses, the observed trip counts span a large range of values (up to $\sim e^{10} \approx 22{,}000$, shown in Figure 8), making it useful for planning decisions concerning the effects of infrastructure upgrades, and new hubs or bus stops. Figure 8 shows the observed vs. predicted values of bike traffic for Model 1. The approximate clustering of points around the trendline suggests that the model predictions are accurate. Additionally, this figure shows that roads with lower values of trips demonstrate a higher amount of variation in predicted values.

5.4. Predictions

In order to demonstrate a real-world application of the presented modeling efforts, several locations where infrastructure improvement projects are being planned by the City of Hamilton had their cycling traffic values predicted before and after a simulated inclusion of new cycling infrastructure. Candidate locations were identified using Appendix B of Hamilton's Cycling Master Plan, which outlined proposed cycling network projects and ranked them based on priority (City of Hamilton, 2018). The cycling projects used for predictions in this thesis are presented in Table 6. These projects were chosen because they

existed within the SoBi service area and were indicated as being high priority by the City of Hamilton.

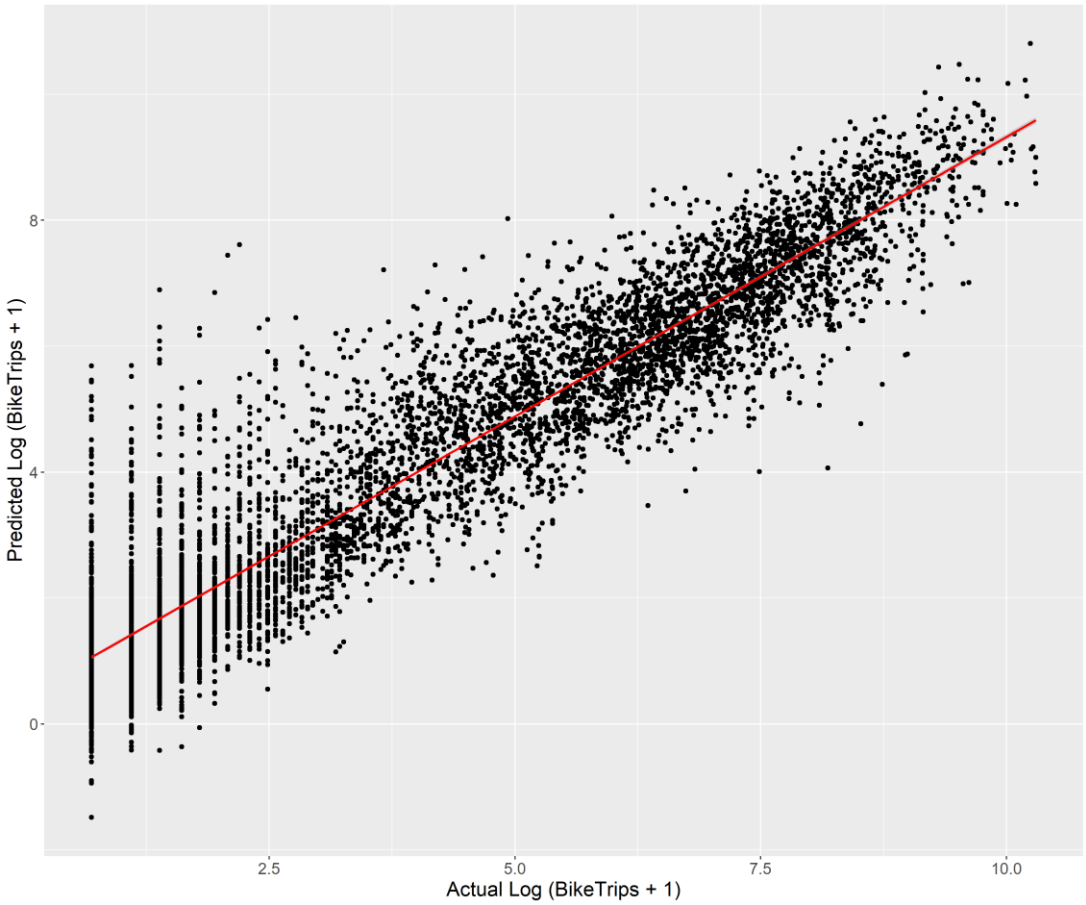

Figure 8: Observed vs. predicted bike trips across the entire network. The red line shows the regression trendline for the data.

For clarity, the predictions were split into different scenarios, all using data from Model 1. The first scenario is simply the base case, where the regression model presented in section 5.3 was used to generate predictions of link trip counts as the links exist presently in the network (i.e., existing road infrastructure). The second scenario is focused on showing the predicted changes in SoBi ridership for links in each project (numbered 1, 2,

Table 6: High priority bike infrastructure projects selected for examination

Project #	Project Name	Description	Length (m)	2017 Cost estimate
1	Hunter St. from MacNab to Catharine	Add two-way bike lane with road diet	470	$57,678
2	Hunter St. from Liberty to Claremont Access	Add two-way bike lane with road diet (west of Wellington), widen street (east of Wellington)	230	$23,071
3	Locke St. from King to Hunter	Add bike lane with road diet, contraflow lane north of Main	1275	$5,912

and 3, as per Table 7). The regression equation from Model 1 was used to calculate the predicted traffic on each link after changing them from their existing bike infrastructure classification (minor road with no infrastructure) to the infrastructure type deemed appropriate by the City of Hamilton (minor road with designated bike lane, in all cases). An overview of the predictions generated for the links involved in each scenario is presented in Table 7, where the "Absolute change" column is calculated between the base case predictions and the predictions when the project links are toggled to a minor road with a bike lane. This table shows that when link infrastructure in the project areas are changed from their original status (minor roads with no infrastructure) into bike lanes, there is a substantial predicted increase in the amount of cycling trips - approximately 815%. The centralized locations of the project links are a reason why predictions are so high. In fact, the predicted links are on average only 64 meters away from the nearest bus stop, and 158

meters away from the nearest SoBi hub. Furthermore, the range of hub trip distance accessibility values covered by the predicted links is quite high, as only the most centrally located links downtown and on McMaster's campus have higher values. As a form of validation for the project predictions, several other links with comparable attributes were manually identified that demonstrated base case predictions with similar values.

Table 7: High priority bike infrastructure projects selected for examination

Project	Link ID	Observed # of Trips	Base Case Predictions	Bike Lane Predictions	Absolute Change
1	1	3,604	1,098	**10,052**	8,954
	2	8,734	2,363	**21,633**	19,270
	3	5,101	2,076	**19,007**	16,931
	4	5,232	1,703	**15,584**	13,881
2	5	1,936	615	**5,629**	5,014
	6	2,022	811	**7,427**	6,616
3	7	3,299	3,238	**11,643**	8,405
	8	2,534	642	**5,877**	5,235
	9	1,992	884	**8,092**	7,208
	10	1,610	709	**6,488**	5,779
	11	1,606	521	**4,770**	4,249
	12	1,855	688	**6,294**	5,606

Note. The "Absolute Change" column was computed relative to the base case.

These predictions show how substantial the predicted increase in trips would be on project links, thus providing support for undergoing the project in real life. Figure 9 gives the spatial context of each of the road projects predicted. In this map, each link is labelled with the corresponding Link ID (shown in Table 7). Although the trip predictions tested

Figure 9: The spatial context of three proposed cycling infrastructure projects

are much different than the observed counterpart, likely as a limitation of the log-transformation used, the overall trend is captured. As stated before, the purpose of this model is not to provide operational predictions, but rather to give insight into the efficacy of proposed projects to make strategic decisions.

6. Conclusion

6.1. Introduction

This thesis used a combination of eigenvector spatial filtering and multiple linear regression modeling to generate predictions of unique cycling trip counts across every link in a network. In the preceding sections, the following 5 objectives were met:

- Create a comprehensive cycling network for the study area by combining multiple datasets and manually adding new trails revealed by GPS trajectories
- Generate cyclist's actual routes between hubs along the road network by processing GPS trajectories
- Identify the total number of trips occurring on each link in the network using the processed 2018 GPS data
- Conduct regression analysis that controls for spatial effects to predict the number of trips on a link for the study period, including identification and creation of relevant variables such as network features and hub trip distance accessibility

- Use the model to predict effects on cycling traffic of high priority planned cycling infrastructure projects in the city at the individual link level

In this section the findings of this thesis are summarized, key assumptions and limitations of the work are discussed, and a conclusion with recommendations for future areas of study are presented.

6.2. Summary of Findings

Past studies exploring cyclist behavior and traffic patterns commonly used household travel surveys or field observations, which are inefficient to conduct and are often limited in spatial scope (Winters et al., 2016). Furthermore, studies using GPS datasets are often based on crowd-sourced data from app providers, which tends to be biased towards younger users who are more likely to be using the technology (Bricka et al., 2012). In this thesis, problems associated with under sampling are alleviated through examination of an entire year's worth of SoBi user GPS trajectories. Predictions are based on the actual routes followed by cyclists as they moved through the network, which was generated by the GIS-based map-matching tool, and not simply start and end locations. This allows for predictions to be made at the precision of individual links, with no aggregation necessary. Several explanatory variables were created and tested to understand the underlying nature of what drives bike share cycling traffic in Hamilton. An ordinary least squares linear regression model was specified, using eigenvector spatial filtering to remove autocorrelation from the residuals. The resulting model had an adjusted R squared value of 0.89 and all explanatory variables were significant at the $p < 0.001$ level. In terms of distance metrics relative to each link, two significant built environment components

were found to be significant in the model – distance to the nearest biking hub and distance to nearest bus stop. This suggests bike share users will more likely choose routes that have greater access to active travel infrastructure that serves to optimize their trips. A novel hub-trip distance accessibility measure was created for this thesis, which was found to outperform both population and employment accessibility measures in terms of predictive power, as well as distance to key trip attractors in the city (i.e., McMaster University and the central business district). Therefore, this variable captured the overall trends of cycling across the service area more effectively than any other variable examined, even those found to be significant in previous literature (see Scott & Ciuro, 2019). In terms of biking infrastructure, all positive trip predictors were infrastructure types that are physically separated from the automobile network (bike lanes being the largest predictor). Several infrastructure types were found to be insignificant trip predictors, such as paved shoulders and signed on-street bike routes. Finally, the model was used to predict trip count changes for proposed bike lane projects. The predictions indicated that cycling traffic would increase substantially on all links tested, demonstrating how the model could be used from a strategic perspective when trying to plan network upgrades.

6.3. Assumptions and Limitations

This thesis is built upon the foundations of previous work for the processing of all GPS data used. Namely, the modules included in GERT (GIS-based Episode Reconstruction toolkit), in particular the map-matching tool (Dalumpines & Scott, 2011, 2018). As such, limitations of these tools apply to the work presented. Specifically, the buffer distance chosen for the map-matching tool affects route generation accuracy

depending on the complexity of the network and the horizontal accuracy of the GPS device used. The default value of 50 meters was used, as this gives reasonable results when used with GPS data that has a horizontal accuracy of 10 meters according to a sensitivity analysis done in the original work. This is reasonably consistent with the accuracy of SoBi equipment, and further manual inspection of processing outputs was done to confirm accuracy. However, this threshold could also explain data loss during the map-matching process. GPS devices inherently contain a margin of error when reporting locations. Routes with unrealistic distances or travel times were removed from this analysis to compensate for this. As mentioned in section 5.1, approximately 17% of the original GPS dataset was filtered out during processing. Although the lion's share of this (~14%) was removed due to GPS errors, the remaining data that was lost during map-matching could have been valid trips that were unable to be captured (e.g., unorthodox GPS trajectories that do not follow the defined cycling network). This has an implication that model predictions presented in this thesis are slightly underestimating the true dataset and could be improved by continued updates to the cycling network. Furthermore, the analysis presented in this work also assumes that cyclists travel along the network created only and that they can travel in any direction they please. Although efforts were made to incorporate links that were popular among cyclists that did not previously exist in the network according to GPS data, it is possible that some shortcuts were not incorporated into the network due to the level of complexity this adds during network creation. The assumption of multi-directional travel was made because it is accurate to cyclist behavior, which is not constrained by the same level of regulation as automobile travel (i.e., even though it is illegal to ride a bike on a

sidewalk in the study area, it is still often done). This assumption prevented many valid trips from being removed during the analysis. During statistical modeling, the use of a natural log transformation on the dependent variable was chosen due to its ease of calculation and interpretation, as well as prevalence in the literature. Although count data models (e.g., Poisson or negative binomial) are rare when it comes to spatial models, avoiding the use of a variable transformation could have led to more accurate predictions, as model predictions demonstrated high variability for low and high traffic count links. The log transformation is also quite sensitive to small amounts of change, especially at high values (e.g., the natural log of 15,000 and 20,000 only differ by ~0.288). When creating and testing accessibility variables for socioeconomic variables (population and employment), the underlying calculation of these variables assumed that accurate measurements could be derived using the proportionality of residential areas relative to the DA level of geography. These measures were also based on the most recent census data – 2016, which is not up to date with the data under examination. With these issues mitigated, the socioeconomic accessibility measures may have had higher predictive power in the model. Moreover, the underlying impedance function used in the accessibility measures was the exponential decay function, which assumes that the changes in a person's willingness to cycle as the trip distance increases follows an exponential decay curve, which is not necessarily the case for all cyclists. For ESF, a first-order edge connectivity spatial weights matrix was used. This assumes that only immediately touching links are considered neighbors. This assumption is a valid way to classify neighbors in a network and easy to calculate, but as pointed out by Ermagun & Levinson (2019), is not necessarily

the most optimal weight format when trying to capture network dependence between links. They introduce the network weight matrix as an alternative, which was demonstrated to outperform network-based models without spatial components and models that used a spatial weights matrix. Finally, in determining model variables, this analysis was conducted while keeping closely in mind that the variable being modelled was SoBi trip counts. This means that when deciding infrastructure levels of importance in the model, the reference level (major roads with no infrastructure) was based on the assumption that cyclists would be least likely to use such links because they are the most unattractive from a safety perspective. Infrastructure classifications that were insignificant in the model therefore took on the same statistical significance as these links. Finally, unpaved multi-use trails were removed from the analysis because it was assumed that since there are very few of them, and that the focus of the thesis was to model utilitarian trips, that this classification was inappropriate to include.

6.4. Concluding Remarks and Future Research

According to the model outputs of this thesis, it was found that physically separated bike infrastructure (designated bike lanes, paved multi-use pathways) were the only positive predictors of cycling traffic for SoBi riders of all tested infrastructure classifications in Model 1. Such upgrades are therefore most highly recommended to encourage the usage of BSSs in the study area. Hub trip distance accessibility was also found to be the single largest predictor of traffic for all variables examined, even more so than population and employment accessibility and simple distance metrics (i.e., distance to McMaster or downtown), which have been found to be significant in previous work. The

model presented in this thesis demonstrated a high adjusted R squared value with full variable significance at the $p < 0.001$ level. Therefore, predictive models such as the one presented in this thesis can be of great strategic importance for city planners and decision makers when trying to decide optimal locations for future improvements to the road network, as it can be used to accurately predict cycling traffic flows.

As mentioned above, a natural log transformed dependent variable was used in the analysis. However, this is not the only type of transformation that could be used in such an analysis. Future work could be done to examine different transformations and their effects on model accuracy and precision. In a similar vein, the impedance function used could be more closely examined. As pointed out by Wu et al. (2019), exponential distance decay functions may be better suited to modeling shorter distance trips (i.e., walking), and that a logarithmic normal distribution function could be more appropriate for modeling cycling trips. Such functions could be implemented during the modeling process and examined for improvements in model fit. Although an entire year's worth of GPS data was used for this thesis, the reality is that trips have been tracked since the program's inception in 2015. Thus, future work could utilize even larger datasets to formulate predictions or be used to compare trends on a year-by-year basis, even seasonal. In fact, the hub-trip distance accessibility measure created for this thesis itself could be predicted and then become an input for the trip prediction model presented in this thesis. With the emergence of bike share providers that are tracking user trips with GPS every day, models no longer must rely on infrequently updated datasets (e.g., official census data or trip diary surveys), and instead can be based on data that is constantly being updated. It is also important to note

that the results of this thesis are presented in the context of just one BSS – SoBi Hamilton. Future work can focus on comparing predictions in different cities and exploring the underlying causes for differences in variable impact.

7. References

Anselin, L. (1988a). Lagrange multiplier test diagnostics for spatial dependence and spatial heterogeneity. *Geographical Analysis*, *20*(1), 1–17. https://doi.org/10.1111/j.1538-4632.1988.tb00159.x

Anselin, L. (1988b). *Spatial Econometrics: Methods and Models* (Vol. 4). Springer Netherlands. https://doi.org/10.1007/978-94-015-7799-1

Babagoli, M. A., Kaufman, T. K., Noyes, P., & Sheffield, P. E. (2019). Exploring the health and spatial equity implications of the New York City bike share system. *Journal of Transport & Health*, *13*, 200–209. https://doi.org/10.1016/j.jth.2019.04.003

Bauman, A., Crane, M., Drayton, B. A., & Titze, S. (2017). The unrealised potential of bike share schemes to influence population physical activity levels—A narrative review. *Preventive Medicine*, *103S*, S7–S14. https://doi.org/10.1016/j.ypmed.2017.02.015

Bivand, R., Keitt, T., & Rowlingson, B. (2019). *rgdal: Bindings for the "geospatial" data abstraction library*. https://CRAN.R-project.org/package=rgdal

Bivand, R. S., Pebesma, E., & Gómez-Rubio, V. (2013). *Applied Spatial Data Analysis with R* (2nd ed.). Springer-Verlag. https://doi.org/10.1007/978-1-4614-7618-4

Bricka, S. G., Sen, S., Paleti, R., & Bhat, C. R. (2012). An analysis of the factors influencing differences in survey-reported and GPS-recorded trips. *Transportation Research Part C: Emerging Technologies*, *21*(1), 67–88. https://doi.org/10.1016/j.trc.2011.09.005

Buehler, R., & Pucher, J. (2012). Cycling to work in 90 large American cities: New evidence on the role of bike paths and lanes. *Transportation*, *39*(2), 409–432. https://doi.org/10.1007/s11116-011-9355-8

Carlos Garcia-Palomares, J., Gutierrez, J., & Latorre, M. (2012). Optimizing the location of stations in bike-sharing programs: A GIS approach. *Applied Geography*, *35*(1–2), 235–246. https://doi.org/10.1016/j.apgeog.2012.07.002

Chen, J., Zhang, Y., Zhang, R., Cheng, X., & Yan, F. (2019). Analyzing users' attitudes and behavior of free-floating bike sharing: An investigating of Nanjing. *Transportation Research Procedia*, *39*, 634–645. https://doi.org/10.1016/j.trpro.2019.06.065

Chun, Y., & Griffith, D. A. (2014). A quality assessment of eigenvector spatial filtering based parameter estimates for the normal probability model. *Spatial Statistics*, *10*, 1–11. https://doi.org/10.1016/j.spasta.2014.04.001

Chun, Y., Griffith, D. A., Lee, M., & Sinha, P. (2016). Eigenvector selection with stepwise regression techniques to construct eigenvector spatial filters. *Journal of Geographical Systems*, *18*(1), 67–85. https://doi.org/10.1007/s10109-015-0225-3

Cintia, P., Pappalardo, L., & Pedreschi, D. (2013). *"Engine matters": A first large scale data driven study on cyclists' performance*. 2013 IEEE 13th International Conference on Data Mining Workshops, Dallas, TX. 147–153. https://doi.org/10.1109/ICDMW.2013.41

City of Hamilton. (2018). *Cycling master plan review and update*. https://www.hamilton.ca/sites/default/files/media/browser/2018-06-06/draft-tmp-backgroundreport-cyclingmp-11-1.pdf

Craggs, S. (2013, December 2). *Hamilton spending $1.6M on new bike share program*. CBC. https://www.cbc.ca/news/canada/hamilton/headlines/hamilton-spending-1-6m-on-new-bike-share-program-1.2447817

Cupido, K., Jevtic, P., & Paez, A. (2019). *Spatial patterns of mortality in the united states: A spatial filtering approach* (SSRN Scholarly Paper ID 3359353). Social Science Research Network. https://papers.ssrn.com/abstract=3359353

Dalumpines, R., & Scott, D. M. (2011). GIS-based map-matching: Development and demonstration of a postprocessing map-matching algorithm for transportation research. In S. Geertman, W. Reinhardt, & F. Toppen (Eds.), *Advancing Geoinformation Science for a Changing World* (Vol. 1, pp. 101–120). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-19789-5_6

Dalumpines, R., & Scott, D. M. (2018). GIS-based episode reconstruction toolkit (GERT): A transferable, modular, and scalable framework for automated extraction of activity episodes from GPS data. *Travel Behaviour and Society*, *11*, 121–130. https://doi.org/10.1016/j.tbs.2017.04.001

Davis, L. S. (2014). Rolling along the last mile: Bike-sharing programs blossom nationwide. *Planning*, *80*(5), 10–16.

de Hartog, J. J., Boogaard, H., Nijland, H., & Hoek, G. (2010). Do the health benefits of cycling outweigh the risks? *Environmental Health Perspectives*, *118*(8), 1109–1116. https://doi.org/10.1289/ehp.0901747

De Mauro, A., Greco, M., & Grimaldi, M. (2016). A formal definition of big data based on its essential features. *Library Review*, *65*, 122–135. https://doi.org/10.1108/LR-06-2015-0061

DeMaio, P. (2009). Bike-sharing: History, impacts, models of provision, and future. *Journal of Public Transportation*, *12*(4), 41–56. https://doi.org/10.5038/2375-0901.12.4.3

DMTI Spatial. (2016). *CanMap Content Suite*. https://www.dmtispatial.com/canmap/

Dray, S., Bauman, D., Blanchet, G., Borcard, D., Clappe, S., Guenard, G., Jombart, T., Larocque, G., Legendre, P., Madi, N., & Wagner, H. H. (2019). *adespatial: Multivariate multiscale spatial analysis*. https://CRAN.R-project.org/package=adespatial

Eren, E., & Uz, V. E. (2019). A review on bike-sharing: The factors affecting bike-sharing demand. *Sustainable Cities and Society*, 101882. https://doi.org/10.1016/j.scs.2019.101882

Ermagun, A., & Levinson, D. M. (2019). Development and application of the network weight matrix to predict traffic flow for congested and uncongested conditions. *Environment and Planning B: Urban Analytics and City Science*, *46*(9), 1684–1705. https://doi.org/10.1177/2399808318763368

Feleke, R., Scholes, S., Wardlaw, M., & Mindell, J. S. (2018). Comparative fatality risk for different travel modes by age, sex, and deprivation. *Journal of Transport & Health*, *8*, 307–320. https://doi.org/10.1016/j.jth.2017.08.007

Ferenchak, N. N., & Marshall, W. E. (2016). *The relative (in)effectiveness of bicycle sharrows on ridership and safety outcomes*. Transportation Research Board 95th Annual Meeting.

Fishman, E. (2016). Bikeshare: A review of recent literature. *Transport Reviews*, *36*(1), 92–113. https://doi.org/10.1080/01441647.2015.1033036

Fishman, E., & Schepers, P. (2016). Global bike share: What the data tells us about road safety. *Journal of Safety Research*, *56*, 41–45. https://doi.org/10.1016/j.jsr.2015.11.007

Fishman, E., Washington, S., Haworth, N., & Watson, A. (2015). Factors influencing bike share membership: An analysis of Melbourne and Brisbane. *Transportation Research Part A: Policy and Practice*, *71*, 17–30. https://doi.org/10.1016/j.tra.2014.10.021

Fuller, D., Gauvin, L., Morency, P., Kestens, Y., & Drouin, L. (2013). The impact of implementing a public bicycle share program on the likelihood of collisions and near misses in Montreal, Canada. *Preventive Medicine*, *57*(6), 920–924. https://doi.org/10.1016/j.ypmed.2013.05.028

Gandrud, C. (2017). *zeligverse: Easily install and load stable zelig packages*. https://CRAN.R-project.org/package=zeligverse

Geary, R. C. (1954). The contiguity ratio and statistical mapping. *The Incorporated Statistician*, *5*(3), 115–146. https://doi.org/10.2307/2986645

Getis, A., & Griffith, D. A. (2002). Comparative spatial filtering in regression analysis. *Geographical Analysis*, *34*(2), 130–140. https://doi.org/10.1111/j.1538-4632.2002.tb01080.x

Ghanem, A., Elhenawy, M., Almannaa, M., Ashqar, H. I., & Rakha, H. A. (2017). *Bike share travel time modeling: San Francisco bay area case study*. 586–591. https://doi.org/10.1109/MTITS.2017.8005582

Goodchild, M. F., & Li, L. (2012). Assuring the quality of volunteered geographic information. *Spatial Statistics*, *1*, 110–120. https://doi.org/10.1016/j.spasta.2012.03.002

Graves, J. M., Pless, B., Moore, L., Nathens, A. B., Hunte, G., & Rivara, F. P. (2014). Public bicycle share programs and head injuries. *American Journal of Public Health*, *104*(8), e106–e111. https://doi.org/10.2105/AJPH.2014.302012

Griffith, D. A. (1978). *The impact of configuration and spatial autocorrelation on the specification and interpretation of geographical models* [Master's Thesis]. University of Toronto.

Griffith, D. A. (2017). Some robustness assessments of Moran eigenvector spatial filtering. *Spatial Statistics*, *22*, 155–179. https://doi.org/10.1016/j.spasta.2017.09.001

Guidon, S., Becker, H., Dediu, H., & Axhausen, K. W. (2018). Electric bicycle-sharing: A new competitor in the urban transportation market? An empirical analysis of transaction data. *ETH Zurich*. https://doi.org/10.3929/ethz-b-000279562

Halás, M., Klapka, P., & Kladivo, P. (2014). Distance-decay functions for daily travel-to-work flows. *Journal of Transport Geography*, *35*, 107–119. https://doi.org/10.1016/j.jtrangeo.2014.02.001

Handy, S. (2005). *Planning for Accessibility: In Theory and in Practice* (pp. 131–147). https://doi.org/10.1108/9780080460550-007

Hansen, W. G. (1959). How accessibility shapes land use. *Journal of the American Institute of Planners*, *25*(2), 73–76. https://doi.org/10.1080/01944365908978307

Harvey, F. J., & Krizek, K. J. (2007). *Commuter bicyclist behavior and facility disruption*. https://trid.trb.org/view.aspx?id=811576

Hochmair, H. H., Bardin, E., & Ahmouda, A. (2019). Estimating bicycle trip volume for Miami-Dade county from Strava tracking data. *Journal of Transport Geography*, *75*, 58–69. https://doi.org/10.1016/j.jtrangeo.2019.01.013

Jestico, B., Nelson, T., & Winters, M. (2016). Mapping ridership using crowdsourced cycling data. *Journal of Transport Geography*, *52*, 90–97. https://doi.org/10.1016/j.jtrangeo.2016.03.006

Kou, Z., Wang, X., Chiu, S. F. (Anthony), & Cai, H. (2020). Quantifying greenhouse gas emissions reduction from bike share systems: A model considering real-world trips and transportation mode choice patterns. *Resources, Conservation and Recycling*, *153*, 104534. https://doi.org/10.1016/j.resconrec.2019.104534

Kutela, B., & Teng, H. (2019). The influence of campus characteristics, temporal factors, and weather events on campuses-related daily bike-share trips. *Journal of Transport Geography*, *78*, 160–169. https://doi.org/10.1016/j.jtrangeo.2019.06.002

Larch, M., & Walde, J. (2008). Lag or error? — Detecting the nature of spatial correlation. In C. Preisach, H. Burkhardt, L. Schmidt-Thieme, & R. Decker (Eds.), *Data Analysis, Machine Learning and Applications* (pp. 301–308). Springer. https://doi.org/10.1007/978-3-540-78246-9_36

Larsen, J., Patterson, Z., & El-Geneidy, A. (2013). Build it. But where? The use of geographic information systems in identifying locations for new cycling infrastructure. *International Journal of Sustainable Transportation*, *7*(4), 299–317. https://doi.org/10.1080/15568318.2011.631098

Le Gallo, J., & Páez, A. (2013). Using synthetic variables in instrumental variable estimation of spatial series models. *Environment and Planning A: Economy and Space*, *45*(9), 2227–2242. https://doi.org/10.1068/a45443

Lowry, M. B., Callister, D., Gresham, M., & Moore, B. (2012). Assessment of communitywide bikeability with bicycle level of service. *Transportation Research Record*, *2314*(1), 41–48. https://doi.org/10.3141/2314-06

Lu, W., Scott, D. M., & Dalumpines, R. (2018). Understanding bike share cyclist route choice using GPS data: Comparing dominant routes and shortest paths. *Journal of Transport Geography*, *71*, 172–181. https://doi.org/10.1016/j.jtrangeo.2018.07.012

Manton, R., Rau, H., Fahy, F., Sheahan, J., & Clifford, E. (2016). Using mental mapping to unpack perceived cycling risk. *Accident Analysis & Prevention*, *88*, 138–149. https://doi.org/10.1016/j.aap.2015.12.017

Mattson, J., & Godavarthy, R. (2017). Bike share in Fargo, North Dakota: Keys to success and factors affecting ridership. *Sustainable Cities and Society*, *34*, 174–182. https://doi.org/10.1016/j.scs.2017.07.001

McKenzie, G. (2019). Spatiotemporal comparative analysis of scooter-share and bike-share usage patterns in Washington, D.C. *Journal of Transport Geography*, *78*, 19–28. https://doi.org/10.1016/j.jtrangeo.2019.05.007

Meddin, R., & DeMaio, P. (2019). *Bike share map*. https://bikesharemap.com/

Menghini, G., Carrasco, N., Schüssler, N., & Axhausen, K. W. (2010). Route choice of cyclists in Zurich. *Transportation Research Part A: Policy and Practice*, *44*(9), 754–765. https://doi.org/10.1016/j.tra.2010.07.008

Midgley, P. (2011). *Bicycle-sharing schemes: Enhancing sustainable mobility in urban areas* (pp. 1–12). United Nations, Department of Economic and Social Affairs.

Miller, T. L. (2017). *leaps: Regression subset selection*. https://CRAN.R-project.org/package=leaps

Miranda-Moreno, L. F., & Nosal, T. (2011). Weather or not to cycle: Temporal trends and impact of weather on cycling in an urban environment. *Transportation Research Record: Journal of the Transportation Research Board*, *2247*(1), 42–52. https://doi.org/10.3141/2247-06

Moran, P. A. P. (1948). The interpretation of statistical maps. *Journal of the Royal Statistical Society. Series B (Methodological)*, *10*(2), 243–251.

Muhs, C. D., & Clifton, K. J. (2015). Do characteristics of walkable environments support bicycling? Toward a definition of bicycle-supported development. *Journal of Transport and Land Use*. https://doi.org/10.5198/jtlu.2015.727

Novak, D. C., & Sullivan, J. L. (2014). A link-focused methodology for evaluating accessibility to emergency services. *Decision Support Systems*, *57*, 309–319. https://doi.org/10.1016/j.dss.2013.09.015

O'Connor, J. P., & Brown, T. D. (2010). Riding with the sharks: Serious leisure cyclist's perceptions of sharing the road with motorists. *Journal of Science and Medicine in Sport*, *13*(1), 53–58. https://doi.org/10.1016/j.jsams.2008.11.003

Open Hamilton. (2018). https://open.hamilton.ca/

Otero, I., Nieuwenhuijsen, M. J., & Rojas-Rueda, D. (2018). Health impacts of bike sharing systems in Europe. *Environment International*, *115*, 387–394. https://doi.org/10.1016/j.envint.2018.04.014

Paez, A. (2019). Using spatial filters and exploratory data analysis to enhance regression models of spatial data. *Geographical Analysis*, *51*(3), 314–338. https://doi.org/10.1111/gean.12180

Park, C., & Sohn, S. Y. (2017). An optimization approach for the placement of bicycle-sharing stations to reduce short car trips: An application to the city of Seoul. *Transportation Research Part A: Policy and Practice*, *105*, 154–166. https://doi.org/10.1016/j.tra.2017.08.019

Parkes, S. D., Marsden, G., Shaheen, S. A., & Cohen, A. P. (2013). Understanding the diffusion of public bikesharing systems: Evidence from Europe and North America. *Journal of Transport Geography*, *31*, 94–103. https://doi.org/10.1016/j.jtrangeo.2013.06.003

Pebesma, E. (2018). Simple features for R: Standardized support for spatial vector data. *The R Journal*, *10*(1), 439–446. https://doi.org/10.32614/RJ-2018-009

Pucher, J., & Buehler, R. (2008). Making cycling irresistible: Lessons from the Netherlands, Denmark and Germany. *Transport Reviews*, *28*(4), 495–528. https://doi.org/10.1080/01441640701806612

Pucher, J., Buehler, R., Bassett, D. R., & Dannenberg, A. L. (2010). Walking and cycling to health: A comparative analysis of city, state, and international data. *American Journal of Public Health*, *100*(10), 1986–1992. https://doi.org/10.2105/AJPH.2009.189324

Pucher, J., Buehler, R., & Seinen, M. (2011). Bicycling renaissance in North America? An update and re-appraisal of cycling trends and policies. *Transportation Research Part A: Policy and Practice*, *45*(6), 451–475. https://doi.org/10.1016/j.tra.2011.03.001

Pucher, J., Garrard, J., & Greaves, S. (2011). Cycling down under: A comparative analysis of bicycling trends and policies in Sydney and Melbourne. *Journal of Transport Geography*, *19*(2), 332–345. https://doi.org/10.1016/j.jtrangeo.2010.02.007

Réquia, W. J., Koutrakis, P., & Roig, H. L. (2015). Spatial distribution of vehicle emission inventories in the Federal District, Brazil. *Atmospheric Environment*, *112*, 32–39. https://doi.org/10.1016/j.atmosenv.2015.04.029

Romanillos, G., Austwick, M. Z., Ettema, D., & De Kruijf, J. (2016). Big data and cycling. *Transport Reviews*, *36*(1), 114–133. https://doi.org/10.1080/01441647.2015.1084067

Ryu, S., Chen, A., Su, J., & Choi, K. (2018). Two-stage bicycle traffic assignment model. *Journal of Transportation Engineering, Part A: Systems*, *144*(2), 04017079. https://doi.org/10.1061/JTEPBS.0000108

Saberi, M., Ghamami, M., Gu, Y., Shojaei, M. H. (Sam), & Fishman, E. (2018). Understanding the impacts of a public transit disruption on bicycle sharing mobility patterns: A case of Tube strike in London. *Journal of Transport Geography*, *66*, 154–166. https://doi.org/10.1016/j.jtrangeo.2017.11.018

Saghapour, T., Moridpour, S., & Thompson, R. G. (2017). Measuring cycling accessibility in metropolitan areas. *International Journal of Sustainable Transportation*, *11*(5), 381–394. https://doi.org/10.1080/15568318.2016.1262927

Sarlas, G., & Axhausen, K. W. (2016). Exploring spatial methods for prediction of traffic volumes. *ETH Zurich*. https://doi.org/10.3929/ethz-b-000116988

Scott, D., & Ciuro, C. (2019). What factors influence bike share ridership? An investigation of Hamilton, Ontario's bike share hubs. *Travel Behaviour and Society*, *16*, 50–58. https://doi.org/10.1016/j.tbs.2019.04.003

Scott, D., & Horner, M. (2008). Examining the role of urban form in shaping people's accessibility to opportunities: An exploratory spatial data analysis. *Journal of Transport and Land Use*, *1*(2). https://doi.org/10.5198/jtlu.v1i2.25

Shaheen, S., Cohen, A., & Martin, E. (2013). Public bikesharing in North America: Early operator understanding and emerging trends. *Transportation Research Record: Journal of the Transportation Research Board*, *2387*, 83–92. https://doi.org/10.3141/2387-10

Shimbel, A. (1953). Structural parameters of communication networks. *The Bulletin of Mathematical Biophysics*, *15*(4), 501–507. https://doi.org/10.1007/BF02476438

Slezakova, K., Castro, D., Delerue–Matos, C., Alvim–Ferraz, M. da C., Morais, S., & Pereira, M. do C. (2013). Impact of vehicular traffic emissions on particulate-bound PAHs: Levels and associated health risks. *Atmospheric Research*, *127*, 141–147. https://doi.org/10.1016/j.atmosres.2012.06.009

SoBi Hamilton FAQ. (2019). https://hamilton.socialbicycles.com/

Soriguera, F., & Jiménez-Meroño, E. (2020). A continuous approximation model for the optimal design of public bike-sharing systems. *Sustainable Cities and Society*, *52*, 101826. https://doi.org/10.1016/j.scs.2019.101826

Statistics Canada. (2017). *Census profile, 2016 Census*. https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/prof/index.cfm?Lang=E

Teschke, K., Reynolds, C., Ries, F. J., Gouge, B., & Winters, M. (2012). Bicycling: Health risk or benefit? *UBC Medical Journal*, *3*, 6–11.

Teschke, Kay, & Winters, M. (2014). *Letter to Editor*. https://cyclingincities-spph.sites.olt.ubc.ca/files/2014/06/Graves-AJPH-as-submitted.pdf

Tessum, C. W., Hill, J. D., & Marshall, J. D. (2014). Life cycle air quality impacts of conventional and alternative light-duty transportation in the United States. *Proceedings of the National Academy of Sciences*, *111*(52), 18490–18495. https://doi.org/10.1073/pnas.1406853111

Tilahun, N. Y., Levinson, D. M., & Krizek, K. J. (2007). Trails, lanes, or traffic: Valuing bicycle facilities with an adaptive stated preference survey. *Transportation Research Part A: Policy and Practice*, *41*(4), 287–301. https://doi.org/10.1016/j.tra.2006.09.007

Tin Tin, S., Woodward, A., Robinson, E., & Ameratunga, S. (2012). Temporal, seasonal and weather effects on cycle volume: An ecological study. *Environmental Health*, *11*(1). https://doi.org/10.1186/1476-069X-11-12

Tobler, W. R. (1970). A computer movie simulating urban growth in the Detroit region. *Economic Geography*, *46*, 234. https://doi.org/10.2307/143141

Ton, D., Duives, D. C., Cats, O., Hoogendoorn-Lanser, S., & Hoogendoorn, S. P. (2019). Cycling or walking? Determinants of mode choice in the Netherlands. *Transportation Research Part A: Policy and Practice*, *123*, 7–23. https://doi.org/10.1016/j.tra.2018.08.023

Vale, D. S., Saraiva, M., & Pereira, M. (2015). Active accessibility: A review of operational measures of walking and cycling accessibility. *Journal of Transport and Land Use*. https://doi.org/10.5198/jtlu.2015.593

Venables, W. N., & Ripley, B. D. (2002). *Modern Applied Statistics with S* (Fourth). Springer. http://www.stats.ox.ac.uk/pub/MASS4

Vize, S. (2017, July 27). *'Everyone Rides Initiative' adds 12 new SoBi bike hubs in Hamilton*. https://www.chch.com/everyone-rides-initiative-adds-12-new-sobi-bike-hubs-hamilton/

Wei, X., Luo, S., & Nie, Y. (Marco). (2019). Diffusion behavior in a docked bike-sharing system. *Transportation Research Part C: Emerging Technologies*, *107*, 510–524. https://doi.org/10.1016/j.trc.2019.08.018

Welch, T. F., Gehrke, S. R., & Wang, F. (2016). Long-term impact of network access to bike facilities and public transit stations on housing sales prices in Portland, Oregon. *Journal of Transport Geography*, *54*, 264–272. https://doi.org/10.1016/j.jtrangeo.2016.06.016

Whitfield, G., Ussery, E., Riordan, B., & Wendel, A. (2016). Association between user-generated commuting data and population-representative active commuting surveillance data—Four cities, 2014-2015. *MMWR. Morbidity and Mortality Weekly Report*, *65*, 959–962. https://doi.org/10.15585/mmwr.mm6536a4

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org

Wickham, H. (2017). *tidyverse: Easily install and load the "tidyverse."* https://CRAN.R-project.org/package=tidyverse

Wickham, H., François, R., Henry, L., & Müller, K. (2019). *dplyr: A grammar of data manipulation*. https://CRAN.R-project.org/package=dplyr

Wing, M. K. C. from J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., Cooper, T., Mayer, Z., Kenkel, B., Team, the R. C., Benesty, M., Lescarbeau, R., Ziem, A., Scrucca, L., Tang, Y., Candan, C., & Hunt, T. (2019). *caret: Classification and regression training*. https://CRAN.R-project.org/package=caret

Winters, M., Teschke, K., Brauer, M., & Fuller, D. (2016). Bike Score®: Associations between urban bikeability and cycling behavior in 24 cities. *International Journal of Behavioral Nutrition and Physical Activity*, *13*(1), 18. https://doi.org/10.1186/s12966-016-0339-0

Wu, X., Lu, Y., Lin, Y., & Yang, Y. (2019). Measuring the destination accessibility of cycling transfer trips in metro station areas: A big data approach. *International Journal of Environmental Research and Public Health*, *16*(15). https://doi.org/10.3390/ijerph16152641

Xie, F., & Levinson, D. (2007). Measuring the structure of road networks. *Geographical Analysis*, *39*(3), 336–356. https://doi.org/10.1111/j.1538-4632.2007.00707.x

Zhao, P., & Li, S. (2017). Bicycle-metro integration in a growing city: The determinants of cycling as a transfer mode in metro station areas in Beijing. *Transportation Research Part A: Policy and Practice*, *99*, 46–60. https://doi.org/10.1016/j.tra.2017.03.003