MODELING ANNUAL BIKE SHARE RIDERSHIP AT HUBS WITH BIKE SHARE EXPANSION IN MIND

MODELING ANNUAL BIKE SHARE RIDERSHIP AT HUBS WITH BIKE SHARE EXPANSION IN MIND

By Geun Hyung (Jayden) Choi, B.Sc. (HONS.)

A THESIS

SUBMITTED TO THE SCHOOL OF EARTH, ENVIRONMENT & SOCIETY

AND THE SCHOOL OF GRADUATE STUDIES

IN PARTIAL FULFILMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

MASTER OF SCIENCE

MASTER OF SCIENCE (2020) McMaster University

(School of Earth, Environment & Society) Hamilton, Ontario, Canada

TITLE:	Modeling annual bike share ridership at hubs with bike share expansion in mind
AUTHOR:	Geun Hyung (Jayden) Choi, B.Sc. (McMaster University)
SUPERVIOR:	Dr. Darren M. Scott
NUMBER OF PAGES:	ix, 75

Abstract

Public bike share systems have been recognized as an effective way to promote active and sustainable public transportation. With the health benefits of bike share becoming better understood, North American cities have continued to invest in cycling infrastructure and impose new policies to not only encourage the usage of bike share systems but also expand their operations to new cities. The city of Hamilton, Ontario, implemented its own bike share system in March 2015. Using the system's global positioning system (GPS) data for annually aggregated trip departures, arrivals, and totals in 2017, this research explores various environment factors that have an impact on users' bike share usage at hub level. Nine predictive linear regression models were developed for three different scenarios depending on the type of hubs and members for trip departures, arrivals, and totals. In terms of variance explained across the core service area, the models suggested the main factors that attract users were distance to McMaster University and the number of racks available at hubs. Furthermore, the working population and distance to the Central Business District and the closest bike lane in the immediate vicinity (200 m buffer) also played important roles as contributing factors. Based on the primary predictors, this research takes one step further and estimates potential trips at candidate sites to inform future expansion of public bike share system. The candidate locations were created on appropriate land uses by applying a continuous surface of regularly shaped cells, a hexagonal tessellation, on the area of interest. The estimated potential usage at candidate sites demonstrated that the east part of the city should be targeted for future bike share expansion.

Keywords: Active travel; public bike share system; hub usage; ridership; sustainable transportation; travel behavior

Acknowledgements

I would like to express my deep and sincere gratitude to several individuals who provided enormous support and contributions throughout my academic career. First and foremost, I would like to thank Professor Darren M. Scott (Master's supervisor) for his constant patience, guidance, and support for the past two years of my master's program. His invaluable teaching of GIS fundamentals and encouragement since my undergraduate thesis has been a stepping stone in my journey to not only pursue my graduate research studies in geography and spatial analysis, but also contribute to the development of sustainable transportation systems around the world. I would also like to thank Pat DeLuca, Dr. Antonio Paez, and Dr. Niko Yiannakoulias for teaching me essential software that were used extensively in this thesis, such as ArcMap and RStudio. I am extending my special thanks to staff at the City of Hamilton and Hamilton Bike Share for providing quality data that made my study possible.

I would like to thank my TransLAB mates: Nosheen Alamgir, Christina Borowiec, Matthew Brown, Samira Hamiditehrani, Michele Tsang, and Raj Ubhi for making time spent in our lab enjoyable. Completing this thesis would have been difficult without their continuous assistance and valuable insights on various methods and interpretation of results. Especially, my special thanks go to Matthew Brown for helping me broaden my English skills, deliver a quality version of thesis, and for being my lab partner until late at night!

Lastly, this thesis would not have been possible without my friends and family who

provided me with their unconditional love and unwavering support during all the ups and downs throughout the course of undergraduate and graduate studies. I could not have done this without you!

Table of Contents

List of Tables

List of Figures

Glossary

Variables	Abbreviation	Description
Social Environment		*(In 200 m buffer from each hub)*
Population	Pop	Number of people aged from 15 to 64 living in residential areas
Employment	Emp	Number of employees in employment areas
Built Environment		*(In 200 m buffer from each hub)*
Major intersections	T_MAJINT	Number of major intersections
Bus stops	T_BSTOPS	Number of bus stops
Bus routes	T_BROUTES	Number of bus routes
Hubs	T_HUBS	Number of hubs
Hub racks	T_RACKS	Number of racks available at hubs
Length of major roads	LEN_MAJRD	Length (km) of major roads
Length of minor roads	LEN_MINRD	Length (km) of minor roads
Length of bike lanes	LEN_BLANES	Length (km) of bike lanes
Length of trails	LEN_TRAILS	Length (km) of trails
Length of bus routes	LEN_BROUTES	Length (km) of bus routes
Proximity		*(Distance measures from each hub)*
Distance to McMaster	DIS_MAC	Distance (km) to McMaster University
Distance to CBD	DIS_CBD	Distance (km) to central business district
Distance to hub	DIS_HUB	Distance (km) to closest hub
Distance to bike lanes	DIS_BL	Distance (km) to closest bike lanes
Accessibility		*(Measure of access to all hubs in the system)*
Population 15-64 (linear decay)	AcsPop	Hub accessibility based on population within 200 buffers with linear decay
Employment 15-64 (linear decay)	AcsEmp	Hub accessibility based on employment within 200 buffers with linear decay
Population & Employment 15-64 (linear decay)	AcsPopnEmp	Hub accessibility based on population and employment within 200 buffers with linear decay

Population 15-64 (estimated decay)	PopDecay	Hub accessibility based on population within 200 buffers with estimated negative exponential distance decay
Employment 15-64 (estimated decay)	EmpDecay	Hub accessibility based on employment within 200 buffers with estimated negative exponential distance decay
Population & Employment 15-64 (estimated decay)	PopnEmpDecay	Hub accessibility based on population and employment within 200 buffers with negative exponential estimated distance decay

1 Introduction

1.1 Research Problem

Public bike share systems provide users with short-term bike rental services with multiple bike hubs distributed over service areas. Over multiple evolutions since the 1960s, recent years have witnessed worldwide prevalence of these systems and have identified them as a convenient source of active transport by numerous users (Fishman, 2016). Initially available in only a small sample of cities across the world, bike share has now positioned its reputation as an alternative sustainable mode of transportation in diverse urban communities. As of 2016, El-Assi et al. (2017) stated that more than 800 cities around the globe have implemented their own public bike share system, and these systems are both growing (in users and fleet size) and becoming more reliable. A study by Faghih-Imani & Eluru (2015) investigated the benefits of introducing bike shares in a city and revealed that having an active public bike share system encouraged individuals to become more environmentally conscious. Furthermore, those who make use of bike share contribute to building healthy cities by minimizing the impact of automobile emissions and reducing congestion and fuel use (Hampshire & Marla, 2012). Supporting the idea of launching and expanding public bike share systems universally, Shaheen and her research team (2013) determined the major benefits of utilizing bike share systems as follows: flexible mobility, health benefits, convenient access to multimodal transport connections, and individual financial savings. Nevertheless, insufficient analysis and investigation on the factors that influence the usage of bike share systems that experience a rapid expansion can cause setbacks and failure due to the lack of supply control (Wu & Lei, 2019).

With the continuous support and recognition by numerous users, public bike share systems have become more prevalent in many urban communities (Zhang et al., 2017). Accordingly, a considerable number of cities are planning or actively expanding their bike share systems to serve more customers (Liu et al., 2017). Furthermore, many cities have been investing in supportive cycling infrastructure, such as bike paths and bike lanes, while imposing new policies to increase the usage of bike share systems (El-Assi et al., 2017; Heinen et al., 2010). As a result, the volume of scientific studies on public bike share has also risen with the apparent increase in the demand for bike usage. It is nevertheless striking that, despite increasing policy and academic interest in public bike share systems, insufficient attention has been paid to users' preferences in making use of a bike share as a primary mode of transport. For instance, many studies have established that there is a strong positive relationship between active travel and development of bike infrastructure (El-Assi et al., 2017; Ogilvie et al., 2007; Pucher et al., 2011). However, Song and her research team (2017) argued that the expansion of bike infrastructure alone might not suffice to endorse active travel, nor promote cycling as a primary mode of transportation. Therefore, it is essential to investigate the travel behavior of bike share users and various factors that attract more members to specific hubs in an effort to maximize ridership (Song et al., 2017). Besides the determinants of user preferences, or environment characteristics around a bike share hub, some researchers contended that a key to the success of a bike share system and further expansion is to explore the location of hubs and their relationship to bike share usage (Lin & Yang ,2011; Liu et al., 2017). Especially for bike share system expansion, a study by Liu and his colleagues (2017) discussed the importance of bike usage

prediction for expansion areas as it can help bike share system designers approximate the number of new users and additional cost for a larger system.

1.2 Research Objectives

To contribute to current research on bike shares, this thesis uses public bike share system data from Hamilton, Ontario's bike share system, Hamilton Bike Share (HBS), which was launched officially on March 22, 2015. The GPS (Global Positioning System) tracking device on each bike enables the collection of large-scale riding trajectory data, which allows researchers to investigate user's travel behavior in various ways. Benefits of GPS-equipped bikes include, but are not limited to, a reduction of bike theft instances and collection of ridership (such as departures and arrivals) data at each hub and cycling routes across a city (Chen et al., 2020). In this research, HBS ridership data for the year 2017 (both departure and arrival trips at hub level) were analyzed to develop predictive models for potential hub usage that can aid in future expansion of the system. Various environment factors (socio-demographics, hub attributes, built environment, and bike infrastructure around hubs), accessibility measures, and proximity to the nearest hub and bike lane were examined to understand the relationship between surrounding features near hubs and annual hub usage. The developed predictive models for hub usage were estimated using linear regression to determine the characteristics that influence ridership and attract bike share users in 2017. Based on the primary predictors, this study takes one step further and demonstrates a GIS-based approach to predict potential trips for representative candidate locations to inform future expansion of the public bike share system in the community. The results from this analysis provide a glimpse into cyclist behavior in the usage of a bike

share system, which provides policymakers and academic researchers with insights on the determinants that impact usage at specific hubs. Meanwhile, urban planners can examine the outcomes from this study and apply them to their communities to improve their own public bike share systems when predicting and planning a new location for a bicycle hub to maximize potential ridership.

1.3 Thesis Outline

There are 6 chapters comprising this thesis including this introduction. Chapter 2 reviews recent bike literature that investigates various factors influencing the usage of a public bike share system, and summarizes approaches employed for bike share network expansion. Chapter 3 reviews the study area for context, the data sources, ridership at the hub level, and predictor variables. Chapter 4 discusses the research methods applied to the data, such as creation of multiple Ordinary Least Squares (OLS) regression models, selection of appropriate variables for the analysis, and the GIS-based approach for offering insights into system expansion. Research results are presented in Chapter 5 with predictive model specification and interpretation of the results. This chapter also depicts representative candidate sites with potential trips estimated to inform the future expansion of bike share system. Chapter 6 encapsulates the key findings and limitations of this thesis, and recommendations for future areas of research.

2 Background

A number of scholars have explored recently the usage of bike share services to improve accessibility of urban transit. In general, those studies differ mainly in three aspects: temporal aggregation of demand, spatial unit of analysis (e.g., defined zones, dissemination areas, census blocks, or core service area), and type of bike share (e.g., docked or free-floating) (Guidon et al., 2020). With the availability of open data, such as bike hub-based data or trip level data, the impact of spatial (i.e., neighborhood) characteristics on bike share usage have been investigated to understand ridership at the hub level. By analyzing these spatial features around hubs further, it is possible to not only expand bike share networks in the vicinity (Zhang et al., 2017), but also predict expansion to a new city (Guidon et al., 2020) to serve more users. Since bike share designers across the world are often challenged to make a strategic decision when operating a new bike share service in a city or further expanding the system to new areas, some studies highlight the importance of knowledge about factors affecting hub usage and predicting bike demand accurately (Guidon et al., 2020; Liu et al., 2017; Noland et al., 2016). Conditions and challenges were also present in their analyses. For example, in order to investigate spatial characteristics, the ridership data from several days to weeks, months, seasons, or years must be aggregated to capture a sufficient number of observations (Guidon et al., 2020; Noland et al., 2016). Furthermore, historical bike transition records are not available for the expansion area, while bike demand at hubs tend to have huge variances across the city (Liu et al., 2017). Thus, this thesis reviewed assorted bike share literature that examined the influence of various environmental features on usage and approaches taken to

determine an appropriate expansion strategy for new areas. It turned out that many studies included similar predictor variables, indicating that their effect on hub usage can be compared.

2.1 Factors Influencing Bike Share Usage

Understanding and determining various factors that influence ridership is one of the most frequently examined topics in the bike share literature. As the experiences of bike share services in other cities have risen, there have been numerous attempts to define the major factors that affect service use. As an example, Heinen et al. (2010) revealed the primary contributing factors (built environment, natural environment, socio-economic, and psychological factors) by surveying previous bike literature. They described the built environment in three categories: urban form, infrastructure, and facilities at work; and the natural environment in two: landscape and weather conditions. Other contributing factors included socio-economic and psychological factors (attitudes, social norms, and habits), and other factors related to utility theory (cost, travel time, effort, and safety). However, Heinen et al. (2010) identified these factors by reviewing and comparing studies that used previous generations of bike share system where ridership data was collected either through human observation, or GPS trackers. Over the past decade, several studies found that the demand at the hub level is strongly associated with built environment characteristics near a hub (Daddio, 2012; El-Assi et al., 2017; Faghih-Imani et al., 2014; Hampshire & Marla, 2012; Rixey, 2013, Scott & Ciuro, 2019; Tran et al., 2015; Wang et al., 2016; Zhang et al., 2017). Scott & Ciuro (2019) examined the impact of various spatial variables on daily ridership at Hamilton, Ontario's bike share hubs using the first year of operation data.

Using multilevel models developed to estimate daily trip departures and arrivals, this paper suggested that weather conditions (temperature and precipitation), temporal variables (daylight hours, weekdays, holidays and university terms), and proximity to popular locations (McMaster University and downtown Hamilton) were strongly associated with ridership. On the other hand, hub attributes within 200 m buffers of hubs were markedly insignificant, indicating that the built infrastructure had little to no influence on daily ridership. Even though population density is often considered an essential independent variable when locating hubs, the results from their research suggested that population does not affect daily departures or arrivals (Scott & Ciuro, 2019). This finding is quite unlike many other ridership studies, where they discovered population near hubs is statistically significant (Daddio, 2012; Efthymiou et al., 2012; El-Assi et al., 2017; Hampshire & Marla, 2012; Tran et al., 2015; Zhang et al., 2017). Regarding this outcome, Scott & Ciuro (2019) argued that the users of the HBS system are usually regular members, such as employees and students in the vicinity of hubs. Because the data was collected from the Canadian Census, where the population is based on 'usual place of residence', such users may not have been necessarily captured in the population variable for their study. In addition, they revealed that the location of bike share hubs plays an important role. For instance, the distance effects estimated in their research suggested that daily bike share usage declined the farther away hubs were from popular locations, demonstrating the presence of a distance-decay effect (Scott & Ciuro, 2019).

Various researchers have examined the built environment factors that affect bike share usage through ridership data at hub-level along using a series of multiple regression

models, spatial analyses, and GIS techniques (El-Assi et al., 2017; Faghih-Imani et al., 2014; Rixey, 2013; Wang et al., 2016). For instance, a study by Rixey (2013) scrutinized the effects of demographic and built environment characteristics on bike share usage through various regression models and analyses. Using monthly aggregated 2010 – 2011 ridership data for three operational systems in the United States (Capital Bikeshare in Washington, D.C.; Nice Ride MN in Minneapolis-Saint Paul, Minnesota; and Denver B-Cycle in Denver, Colorado), they identified the following spatial variables as having statistically significant correlations with ridership at hub level: population and retail job density; bike, walk, and transit commuters; median income; education; presence of bikeways; non-white population; days of precipitation; and proximity to a network of other public bike share hubs. Rixey (2013) performed a regression analysis using the natural log of the number of monthly rentals by hub during the system's first operating season as the dependent variable. Here, the non-white population and days of precipitation variables were negatively associated with ridership, whereas proximity to a greater number of other hubs delineated a stronger positive correlation with ridership. This result suggested that accessibility to a comprehensive network of bike share hubs plays an important role when maximizing ridership. Similarly, El-Assi et al. (2017) revealed significant effects of road network configuration (intersection density and spatial dispersion of stations) and bike infrastructure (bike lane, paths, etc.) on daily bike sharing usage. However, instead of taking into account the total trips that occurred at hubs, El-Assi and his research team estimated distributed lag models to predict trip departures, arrivals, and hub-to-hub trips in an effort to analyze the influence of factors on departure and arrival flows at the hub level.

They made use of year-round historical trip data and developed a hub pair (origin-destination) regression model. As a result, the empirical models exhibited a negative association between distance and ridership, indicating that usage at each hub decreased as distance increased. A positive correlation was discovered between hubs located near university campuses, transit stations, and the downtown core area. In addition, lag models for trip generation and attraction to predict the trips at each hub suggested that temperature was positively correlated with bike share trip activity. Higher temperatures, lower humidity levels, and smaller amounts of ground snow showed a positive relationship with bike ridership (El-Assi et al., 2017). Supporting El-Assi et al. (2017)'s findings, a study by Faghih-Imani et al. (2014) proposed that increased bicycle flow and usage were associated with increased bicycle facilities near a hub and fewer intersections with major roads under good weather conditions. To investigate these factors postulated to affect bike share ridership in Montréal, Canada, minute-by-minute readings of bicycle availability at all BIXI public bike share hubs were collected from April to August 2012. For the ease of applying the developed methodology and findings to other regions, Faghih-Imani et al. (2014) examined the arrival and departure frequencies at the station level using a multilevel estimation approach via statistical modelling. The results from the models for arrival and departure rates were then evaluated using the data from May 2013 and found to be reasonably accurate to the observed rates. Although Wang et al. (2016) adopted a different approach (log-linear and negative binomial regression models) to identify significant determinants that influence ridership for the Nice Ride Minnesota bike share system in Minneapolis-Saint Paul, Minnesota, their research found similar results. Data for 13

independent variables, including indicators of economic activity (retail facilities and job accessibility), neighborhood socio-demographics (age, race, etc.), the built environment, and transportation infrastructure, were obtained from a variety of sources such as 2010 Census. The developed log-linear and binomial regression models affirmed that the hubs located near campuses, the central business district (CBD), water bodies, and parks were highly correlated with average daily hub usage (Wang et al., 2016). However, proximity to other bike share hubs showed a negative association with bike usage, contradicting previous studies. Wang et al. (2016, p. 8) and his research team concluded that "station oversaturation may present a challenge for system management. Relocation, monitoring of station use, and re-estimation of models may provide insights into the optimal number of stations, which can make the overall system more efficient." The findings from various studies that explored the determinants of bike share usage can potentially influence decision makers to upgrade active transport infrastructure or expand the system for potential bike hubs in new areas.

2.2 Locating Bike Share Hubs

With the increasing amount of literature available on public bike share systems around the globe, many researchers and preliminary studies argue that one of the critical keys to the success of such systems is the location and distribution of bike hubs (Garcia-Palomares et al., 2012; Lin & Yang, 2011; Park & Sohn, 2017). Generally, most public bike share systems can easily be found in city centers or higher-density areas (Garcia-Palomares et al., 2012). This is because the size and configuration of the city are fundamental indicators when implementing a new public bike share system. Once the

coverage area is determined, the location of hubs is examined to satisfy potential demand. As discussed above, factors ranging from the built environment to others related to the natural environment, weather, socio-demographics, psychology, and utility theory play an essential part in the decision to select optimal locations for bike share hubs. Faghih-Imani et al. (2014) suggest that the proximity between hubs should be taken into consideration for the convenience of users prior to choosing bike share hub locations. For instance, the BIXI public bike share system in Montréal deployed a hub every 250 – 300 m throughout the central city, allowing for easy access (Faghih-Imani et al., 2014). The same was the case for the Hamilton Bike Share system. Consequently, BIXI users were able to efficiently rent and return a bicycle at their convenience. Nevertheless, Shu et al. (2013) remarked that the over-coverage of hubs in one area could result in unfavorable outcomes due to the maintenance costs caused by the clustering of hubs in one area.

Various methodologies have been suggested to optimize the location of hubs in bike share systems and analyze the spatial distribution of potential ridership. Geographic Information Systems (GIS) have been at the forefront of this effort as an effective support tool (Garcia-Palomares et al., 2012; Kabak et al., 2018; Noland et al., 2016; Park & Sohn, 2017; Zhang et al., 2019). Many studies have sought to configure an initial bike share system using GIS techniques. Garcia-Palomares et al. (2012) demonstrated the benefits of adopting location-allocation models to a proposal for the optimal location of bike share hubs. Location-allocation analysis is an effective tool that identifies optimal facility location(s) in a way that serves the postulated usage pattern most efficiently. To calculate the spatial distribution of the potential demand for cycling trips, Garcia-Palomares and his

research team (2012) made use of two of the most used location-allocation modeling approaches (minimizing impedance and maximizing coverage). They examined the optimal location for bike share hubs, explored main characteristics near hubs, and measured the accessibility of each hub in the city center of Madrid, Spain. For their proposed method, the following statistical information was manipulated: a 2010 street network from the Madrid Regional Statistics Office, transport zones from a 2004 mobility survey, the 2010 population and number of jobs available at the building level from Cartociudad-National Geographical Institute, and stations and stops in the public transport network from 2011. When the minimize-impedance solution was selected as a location-allocation modeling approach, they found that this approach considers spatial equity as it generates a relatively uniform coverage based on the distance minimized for supply and demand. The maximize-coverage solution suggested a more interesting outcome for their study region in terms of efficiency. This is because the maximize-coverage solution maximized the potential demand within a specific radius (200 m) from the hubs, which can help identify potential areas for new hubs (Garcia-Palomares et al., 2012). However, their use of location-allocation modeling included some drawbacks: First, the bike system they employed was aimed to serve the local population on workdays, indicating that recreational or tourist bike share systems may require a different approach. Second, the incompetency in capturing certain places in the city, such as large parks, that have neither population nor jobs and yet still attract a considerable number of trips, was advised. Although location-allocation modeling might suffice for optimizing bike share hub locations for the initial configuration of a system, a significant drawback is that there is no ridership data, and

therefore, understanding how the system has performed is challenging. In addition, they might be missing out on other relevant neighborhood factors (e.g. number of bike lanes in the vicinity of hubs or number of racks available at each hub), which could be certainly a multi-criteria problem. Addressing these problems, Garcia-Palomares et al. (2012) suggested that employing a predictive model that considers environment factors around hubs is essential to determine the exact location of potential hubs for future work.

Once a city has implemented a bike share system and the system gains in popularity as an increasingly convenient source of transport, the city may consider expanding the existing service area to surrounding regions. Expanding the bike share service not only broadens the original users' capability to reach new areas but also attracts new users in the expanded areas. A study conducted by Noland and his colleges (2016) examined the determinants of bike share hub usage by estimating a series of Bayesian regression models of trip generation at hubs in an effort to forecast the trips generated at new hubs. Various factors were examined, including bike infrastructure (bike parking racks and bike routes), population and employment, different types of land uses (such as recreational, parking, residential, and other land uses), and access to other transit services. Estimates were developed for different months (February, July and November of 2014), weekdays and weekends, and type of user (members vs. non-members). The models' results suggested that hubs located near busy subway stations, as well as areas with more population and employment, tend to predict greater usage of the system by members, while bike lanes and paths were associated with more non-member bike share trips. Based on the factors analyzed in 2014, their study attempted to forecast trip generation at new stations opened

in 2015. Although the inferential models provided some insights for decision makers as to what factors could influence the success of the system, the outcome suggested a large variation in predictive power, and the models did not perform well when forecasting trip generation at the hub level. They concluded that the main culprit for overestimating usage at new hubs was due to insufficient time for the newly opened hubs to grow in 2015 (Noland et al., 2016).

Similarly, Guidon et al. (2020) estimated and assessed linear and spatial regression models, along with random forests, for bike share demand and predicted usage in a different city (e.g., in the case of an expansion). Using booking data from bike share "Smide" in Zurich, Switzerland, they developed multi-factor bike predictive models to predict the number of bookings for Berne, Switzerland, and the predictions were validated with the data from Berne. The trip data was the number of bookings which consisted of origin and destination coordinates, and a timestamp (costs 5 CHF – the official currency of Switzerland – and usage is charged pro rata on a per-minute basis) for each trip. The trips were aggregated to a 300 m raster covering the service areas for both cities in the study period of 2018, and used as the dependent variables for regression analysis and random forests. It is important to note that the Smide was newly introduced in the city of Berne, and therefore, only three months-long of trip data (from October to December) were collected – 120,472 trips in Zurich for full year data and 2,973 in Berne for three months of data. The results suggested that social environment variables (population and employment) and recreational variables (bars and restaurants) were essential contributing predictors that have a direct effect on bike usage. However, centrality measures (distance

to the main train station and the boundary of the service area) appeared to be less affective and should be included if a research goal is to make predictions for the same city, but should be omitted if the goal is to make predictions for a different city. The models demonstrated a reasonable performance in non-central areas; however, the central areas depicted large overprediction of usage for the new city. They described the overprediction in central areas was due to higher number of main train station users in Zurich than Berne, and that Zurich is a bigger city with more social activity in the center. In essence, the service area covers a larger share of the market of trip distances in Zurich, which allows for longer trips, than in Berne. Their overprediction in a smaller service area and the smaller size of the city also suggested that there could be other drivers of usage that were not considered in their study (i.e., Berne has fewer visitors and tourists than Zurich). Another explanation was that since the data from Berne was from the three months after the introduction of the system, the number of bike users and bookings require more time to grow first. This explanation was also considered by Noland et al. (2016), which also overpredicted usage. At this point, it should be noted that not only factors affecting demand should be considered, but also the time component of ridership data (e.g. days, weeks, months, seasons, or years) at hubs should be taken into account when forecasting usage at potential hubs in new expansion areas.

Overall, previous studies have identified population, employment, proximity to major centers (e.g. main subway station, university, museums etc.), and bike infrastructure as important variables to explain potential usage at hub level in the expanded areas. The effect of social environment variables (population and employment) was generally deemed

to show positive relationship with bike share usage. However, it should be noted that its magnitude may vary because of differences in trip purposes and user base (Scott & Ciuro, 2019). The multi-factor bike usage prediction models developed by previous studies could present important indications of which predictors should be considered when estimating potential usage in the expansion areas. Nonetheless, the question remains whether using short period of data to represent considerable variation across the year and predict hub usage in new areas can give reliable indications for decision makers. To bridge this research gap, this thesis makes use of an annual bike hub usage data to explore relevant factors that influence ridership at hub level and create a predictive model based on actual travel behavior by users to identify candidate locations in a new region and inform future expansion of public bike share system.

3 Data

3.1 Study Area

The city of Hamilton is a mid-sized Canadian city with a population of 536,917 in 2016 (Statistics Canada, 2017). The city launched its public bike share system, Hamilton Bike Share, in March 2015. There were only 110 hubs across the city when Hamilton's bike share was initially launched, but due to continued growth and support by numerous users, the bike share expanded its coverage through additional hubs. As of 2019, there were a total of 132 hubs and 825 bikes in operation in the service area (Hamilton Bike Share, 2019). Considering equity for all citizen access and increase in public bike share demand in the city, HBS implemented the Everyone Rides Initiative (ERI) program and deployed

12 new hubs in 2017. The ERI program provided ERI members with not only new hubs in low-income areas in the city, but also cycle training sessions, language translation services and subsidized memberships. Figure 3.1 shows the distribution of hubs in the core service area, where regular[1] hubs are distinguished from ERI hubs. The bike share also includes a small strip of service area along Van Wagner's Beach, but was not included in the analysis since it is not contiguous with the core service area. The regular hubs show a relatively clustered pattern near major centers – namely, the central business district area and McMaster University, while the ERI hubs are distributed across the residential neighborhoods in the eastern part of the core service area.

3.2 Dependent Variable

HBS bikes are GPS-equipped, meaning that the availability of bike at hubs and cycling routes can be identified in real time. For example, it is possible to inform users of the rental or parking availability at hubs (either start or end), while scholars can make use of this data for various geographic analysis (such as route choice analysis) (Lu et al., 2018). For this study, annual trip departures, arrivals, and totals were aggregated using bicycle data for all hubs in service from January 1, 2017, to December 31, 2017. However, as mentioned previously, the hubs near Lake Ontario (Van Wagner's) were not taken into consideration as these hubs are not contiguous with the core service area. To obtain the number of trip records, a series of data processing steps was conducted with the following criteria:

[1] Term "regular" was used for non-ERI hubs and members

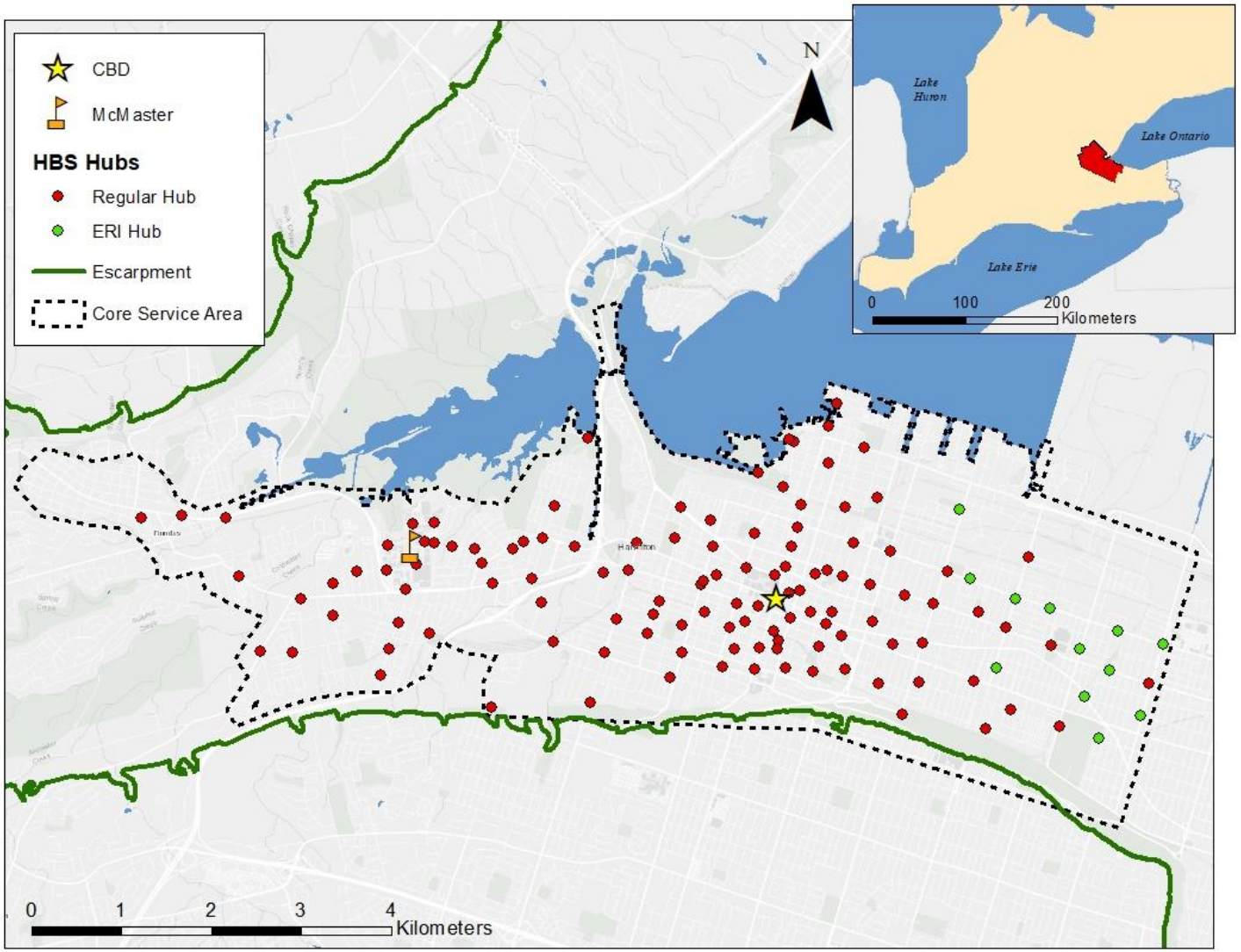

Figure 3.1: Study area of Hamilton Bike Share, Ontario

- Distance travelled was between 0.1 and 100 kilometers
- Trip duration was between 1 and 480 minutes
- Speed was between 1 and 35 kilometers per hour

As a result, 14,096 trips with missing distances or trip durations, or extreme values of the above attributes were eliminated, and the total number of 385,041 trip data were collected. Nonetheless, when the ridership data was explored in detail, there were 101,502 trips that did not contain both departing/arriving hub information, or only had either information recorded. For instance, 16,582 trips failed to collect both departing/arriving hub information, 69,357 trips missed departing hub information, and 48,477 trips missed arriving hub information. To cope with missing data pertaining to the departures and arrivals, and thus totals, each trip with missing information was identified and associated with the nearest hub by using Euclidean distance, which calculates the shortest distance from one location to the other. In this process, trips without departure and/or arrival hub information located outside of a 100 m buffer around hubs were considered as irregularities and removed from further analysis. A 100 m buffer was chosen as a threshold considering how trip records were tightly clustered around the hubs (Figure 3.2). In this figure, a sample of hubs in the core service area are shown. The 100 m buffers for each of these hubs depict a considerable amount of trips included. Yet, it is possible to detect some trips being scattered across the service area – generally the departing trips being more spread out than arrival trips. This outcome could have been caused by a lag between activating a bike and acquiring a GPS signal after a user starts riding with it. Another explanation could be due

to transformation of GPS trajectories into actual traces that user took (Lu et al., 2018). To describe the distribution of total departures and arrivals that were being captured within the 100 m distance from the center of buffers, Figure 3.3 was created. The gradual trendline of departure trips illustrates the necessity of a longer distance to capture the trip records compared to the steeper trendline of arrival trips, implying that the distribution of departures is more dispersed across the service area than arrivals.

Once all the 2017 ridership data were defined with their original and destination hub information, this study developed nine predictive models, where three main combinations based on hubs and users estimated departures, arrivals, and totals for hubs (Table 3.1). The first main model incorporates the total number of departures, arrivals and totals recorded at the regular hubs by regular bike share members only (regular hubs + regular members). Model 2 builds upon the Model 1 by considering the ERI hubs in service, but limits usage to trips made by regular members only (all hubs + regular members). Lastly, Model 3 consists of Model 2 and sums up the hub usage by all members, including both regular members and ERI members (all hubs + all members). Using these nine predictive models, the impact of the ERI program on bike share usage and the influence of current ERI hub locations, as well as the usage by ERI members versus regular members, can be explored thoroughly. Accordingly, nine dependent variables were generated through a series of data processing steps. In essence, the number of departure,

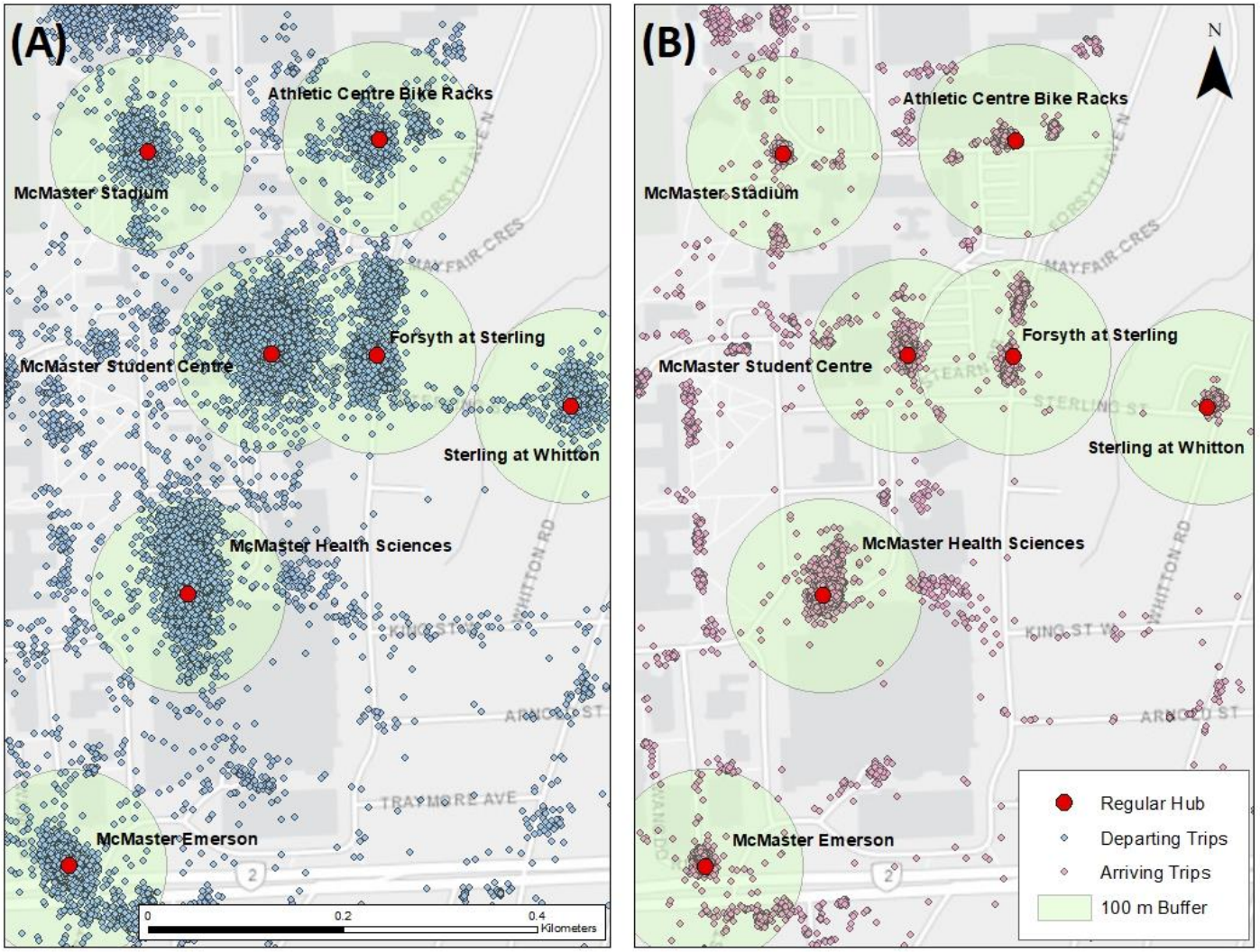

Figure 3.2: Example of total trip departures (A) and arrivals (B) within 100 m buffers around hubs in 2017

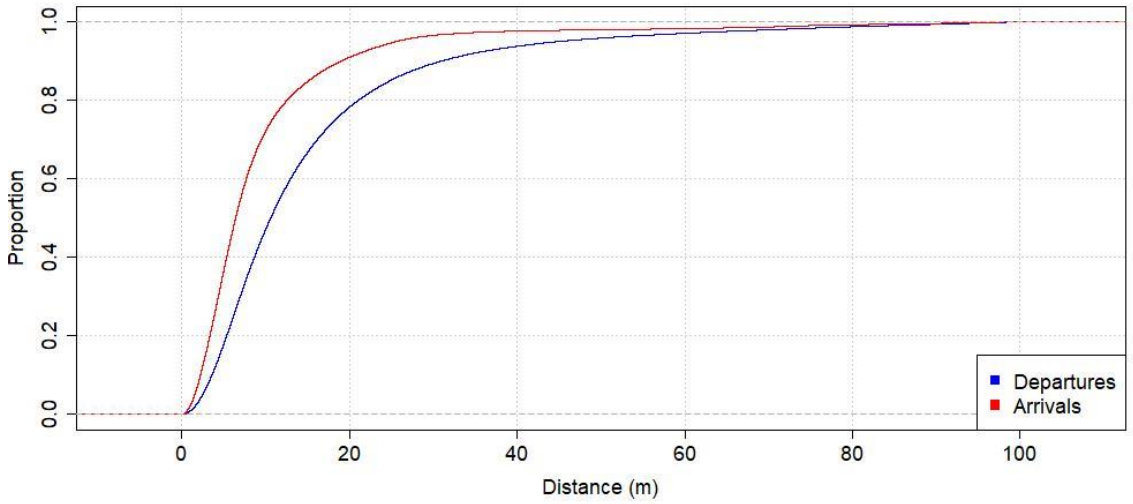

Figure 3.3: Cumulative proportion of total departures and arrivals captured within 100 m distance from center of buffers

arrival, and total trips before and after applying appropriate data processing (e.g., ERI members vs. regular members or ERI hubs vs. regular hubs) was investigated for the development of predictive models (Table 3.1). During this procedure, there were some mismatches found in hub names between the 2017 ridership data and the 2019 hub GIS dataset obtained from the City of Hamilton's open data portal. Since the 2019 hub dataset contained the updated version of the hub names, the mismatches between the two datasets were amended based on the 2019 hub data. For example, the hub name '40 Oxford' in the ridership data was corrected to 'Oxford at York', while temporary hubs launched each year for one-day events (i.e., 'Locke St Festival – Locke at Canada' or 'SUPERHUB') were disregarded and the associated data were removed. Along with the hub name updates, this study made a minor change on the following hubs in the 2019 hub dataset: 'Cootes Drive Dundas' and 'King at Millers'. This is because these hubs were replaced with a central

"virtual zone" where users can lock their bikes anywhere within this confined zone for free beginning in late 2017 (Campbell, 2017). However, considering how this arrangement was made in late 2017 and the number of trips made from the zone was less than 50, these two hubs were preserved. In other words, instead of creating a new study area to capture the trips that occurred in the virtual zone, the locations of Cootes Drive Dundas and King at Milers hubs were maintained, and 100 m buffers were used to obtain the ridership data around them.

The final trip counts for the nine predictive models were obtained and are displayed in Table 3.1. As expected, the first main model (regular hubs + regular members) for departures, arrivals, and totals contained the smallest number of trips, while Model 3 (all hubs + all members) contained the most. In this table, it can be noticed that a different number of departures, arrivals, and totals were obtained depending on whether or not trips were associated with the ERI program (ERI hubs and members). The arrival trips for Models 1 and 2 show a higher number in comparison to the departure trips, whereas Model 3 demonstrates the opposite. Explanations of such outcomes include, but are not limited to: (1): the GPS startup issue caused by delayed GPS signal, where the departing trips were recorded beyond 100 m buffers; (2) departure trips beginning at regular hubs in Model 2 (all hubs + regular members), but completing the trips at ERI hubs; and 3) the additional trips taken by ERI members in Model 3 or the tendency of ERI members to park their bikes outside of a hub, instead of returning it to a bike rack, at their convenience by paying an out-of-hub penalty (Hamilton Bike Share, 2019). It is important to note that the dependent variables in this analysis are the natural logarithms of the number of trips per hub in 2017.

Table 3.1: Data processing to acquire number of departures, arrivals, and totals for Models 1 to 3

	Model 1 (regular hubs + regular members)				**Model 2 (all hubs + regular members)**				**Model 3 (all hubs + all members)**			
	Departures		**Arrivals**		**Departures**		**Arrivals**		**Departures**		**Arrivals**	
	***With* hub info.**	***Without* hub info.**	***With* hub info.**	**W*ithout* hub info.**	***With* hub info.**	***Without* hub info.**	***With* hub info.**	**W*ithout* hub info.**	***With* hub info.**	***Without* hub info.**	***With* hub info.**	**W*ithout* hub info.**
Initial trips	283,789	101,252	283,789	101,252	283,789	101,252	283,789	101,252	283,789	101,252	283,789	101,252
Eliminated ERI members	267,634	73,846	267,634	73,846	267,634	73,846	267,634	73,846	-	-	-	-
Eliminated ERI-related trips that *start* at ERI hubs	265,465	73,522	265,465	73,522	-	-	-	-	-	-	-	-
Eliminated ERI-related trips that *end* at ERI hubs	263,626	72,775	263,626	72,775	-	-	-	-	-	-	-	-
Eliminated trips located outside 100 m buffer	-	46,334	-	49,396	-	47,761	-	50,580	-	62,803	-	61,973
After Name Update	309,890		312,952		315,358		318,179		346,553		345,725	
Eliminated recreational-purpose trips[2]	308,643		311,699		314,105		316,926		345,265		344,440	
Total departures & arrivals	308,643		311,699		314,105		316,926		345,265		344,440	
Total trips	620,342				631,031				689,705			

[2] Trips recorded in the vicinity of Van Wagner's were not only contiguous with the core service area, but were also found to be recreational-associated trips by visitors (Echo, 2018).

The natural log transformation was selected, rather than direct use of annual ridership data, in order to help linearize the variables, to improve the continuity of a discrete count variables, and to ensure that the estimated OLS regression models to be constrained to positive values (Osborne, 2002; Rixey 2013). This step is essential to avoid the violation of the independence assumption of traditional linear regression.

3.3 Independent Variables

Table 3.2 summarizes the independent variables developed for this study. A 200 m buffer around hubs was used to capture various spatial variables in their vicinities (i.e., socio-demographics, hub attributes, and environment factors). Since distances between each hub were approximately 300 to 600 m apart, 200 m was considered adequate as the walking distance to the closest hub (Scott & Ciuro, 2019). Similarly, a 250 m buffer around each hub was used for the BIXI public bike share system in Montréal in Faghih-Imani et al. (2014)'s study examining the spatial determinants affecting ridership at hubs. However, a 250 m buffer distance was found to be excessive for Hamilton's bike share as too many hubs overlapped, which can cause a problem when capturing trips and spatial variables within the buffer. Besides, using a 200 m buffer can minimize the number of proximate hubs within a buffer (Scott & Ciuro, 2019). The socio-demographic variables (population and employment) in this study were derived from 2016 Canadian Census data, allocated to appropriate land-use polygons to create refined variables for each hub, while the data for built environment variables were collected from the city of Hamilton's open data portal.

3.3.1 Social Environment

The social environment variables used in this analysis consisted of the total population in residential areas and employees aged from 15 to 64 within a 200 m buffer from each bike share hub. In this process, this study made extensive use of ArcGIS Pro, which is a desktop geographic information system (GIS) software product developed by Esri. Considering the inadequacy of assuming population to be equally distributed throughout a dissemination area (DA), the lowest level of geography for census data, people residing in each DA were first allocated to residential areas. Subsequently, 200 m buffers around hubs were intersected with the residential areas to create a cross-tabulation of population from DAs based on the proportion of residential area inside the buffers. The cross-tabulation was then aggregated to the hub level. The same process was used to obtain the working population in each buffer, except employment areas were used instead of residential areas. Moreover, since the number of workers within DAs is not a standard census product, individual census records were aggregated by their 'place of work' at the DA level. The residential areas and employment areas were derived from Hamilton's parcel data.

3.3.2 Built Environment

Many studies investigated numerous built environment factors on bike share usage using ridership data at the hub level (e.g., Daddio, 2012; El-Assi et al., 2017; Faghih-Imani et al., 2014; Hampshire & Marla, 2012; Rixey, 2013, Tran et al., 2015; Wang et al., 2016; Zhang et al., 2017). In this study, the following additional variables were created to explore built environment factors and transportation infrastructure around hubs: the number of

major intersection, lengths of major roads, minor roads, bike lanes, trails and bus routes, the number of HSR bus stops and bus routes, and the number of hubs and racks available in the service area. The data to construct these variables were obtained from Hamilton's open data portal. In addition, the distance from each hub to major centers (i.e., McMaster University and CBD), the nearest hub, and bike lanes were derived based on the Euclidian distance (refer to Table 3.2 for further information).

3.3.3 Accessibility

Three different variables were created to measure each hub's accessibility with respect to the rest of the system: one based on population; another, employment; and a third with the two combined. These accessibility measures were derived from social environment variables, where 200 m buffers were used around each of the HBS hubs. The population within the buffers was chosen to represent the number of people near the hub that might utilize the bike share service, while the employment within the buffers was chosen to represent the attractiveness of an area near a hub. In addition, a variable that combined the population and employment was also created and tested in the model to compare the results at the end. Inspired by Scott & Horner (2008), this study used two different functions to construct the accessibility variables. In addition to linear distance decay, a negative exponential distance decay impedance function was used to calculate one set of gravity-based accessibility measures. This method presumes that opportunities are complementary, and a cost is travel time or distance to the destination, which should be minimized and kept within a threshold (Saghapour et al., 2017; Vale et al., 2015). Gravity-

based accessibility measures typically take the form of the following expression, introduced by Hansen (1959):

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

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). 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. Rather than choosing an arbitrary value for β, a value was calculated using the unique hub-to-hub travel distances according to the following model:

$$I_k = \alpha \exp(-\beta t_k) \tag{2}$$

where I_k is the number of trips for the distance category k, and t_k is the trip distance in 100 m increments for category k. The β value computed in this model was 0.000692, as trip counts drastically decrease as the distance of the trip increases. By applying linear and estimated negative exponential distance decay functions, how spatial interaction declines with increasing distance can be quantitatively estimated – decrease in ridership as the distance to a nearby hub increases.

Table 3.2: Independent variables: definitions and descriptive statistics

Variables	Description	Mean	S.D.	Min.	Max.
Social Environment	*(In 200 m buffer from each hub)*				
Population	Number of people aged from 15 to 64 living in residential areas ($\times 10^{-3}$)	0.45	0.40	0	2.18
Employment	Number of employees in employment areas ($\times 10^{-3}$)	0.64	0.92	0	4.99
Built Environment	*(In 200 m buffer from each hub)*				
Major intersections	Number of major intersections	0.45	0.82	0	5
Bus stops	Number of bus stops	3.91	3.82	0	25
Bus routes	Number of bus routes	3.57	5.08	0	23
Hubs	Number of hubs	1.26	0.52	1	3
Hub racks	Number of racks available at hubs	10.42	3.77	5	30
Length of major roads	Length (km) of major roads	0.43	0.34	0	1.50
Length of minor roads	Length (km) of minor roads	1.25	0.56	0	2.51
Length of bike lanes	Length (km) of bike lanes	0.47	0.34	0	1.33
Length of trails	Length (km) of trails	0.22	0.37	0	2.12
Length of bus routes	Length (km) of bus routes	1.55	2.40	0	15.11
Proximity	*(Distance measures from each hub)*				
Distance to McMaster	Distance (km) to McMaster University	3.84	2.12	0.15	8.44
Distance to CBD	Distance (km) to central business district	2.24	1.58	0.05	7.12
Distance to hub	Distance (km) to closest hub	0.30	0.14	0.05	1.01
Distance to bike lanes	Distance (km) to closest bike lanes	0.10	0.15	< 0.01	0.84
Accessibility	*(Measure of access to all hubs in the system)*				
Population 15-64 (linear decay)	Hub accessibility based on population within 200 buffers with linear decay ($\times 10^{-3}$)	0.04	39.68	0	0.21

Employment 15-64 (linear decay)	Hub accessibility based on employment within 200 buffers with linear decay ($\times 10^{-3}$)	0.06	85.43	0	0.36
Population & Employment 15-64 (linear decay)	Hub accessibility based on population and employment within 200 buffers with linear decay ($\times 10^{-3}$)	0.09	99.43	0	0.36
Population 15-64 (estimated decay)	Hub accessibility based on population within 200 buffers with estimated negative exponential distance decay ($\times 10^{-3}$)	14.35	15.68	0	84.74
Employment 15-64 (estimated decay)	Hub accessibility based on employment within 200 buffers with estimated negative exponential distance decay ($\times 10^{-3}$)	21.46	32.54	0	141.82
Population & Employment 15-64 (estimated decay)	Hub accessibility based on population and employment within 200 buffers with negative exponential estimated distance decay ($\times 10^{-3}$)	35.49	38.20	0	151.76

4 Methodology

4.1 Assumptions of Multiple Linear Regression

Once all the dependent and independent variables were developed and organized, a series of multiple OLS regression models were estimated. For this thesis, three sets of predictive models were assessed. Within each set, models were estimated for trip departures, arrivals, and totals (departures and arrivals combined). The multiple OLS regression model takes the following form:

$$y_j = \beta_0 + \beta_1 x_1 + ... + \beta_n x_n + \varepsilon \qquad j = 1, ..., J \qquad n = 1, ..., N \qquad (3)$$

where y_j is observed values (in this case, the number of departures, arrivals, or totals for hub j) for all three models, β_0 is the intercept, β_n are the coefficients for independent variables, and x_n are observed independent variables for each hub. ε is a random error term assumed to follow a normal distribution with a mean of 0.

When estimating a multiple OLS regression model, several assumptions must be met (Osborne & Water, 2002):

- No or little multicollinearity in the data
- Normal distribution of residuals
- Linearity between the independent and dependent variables
- Homoscedasticity – freedom from extreme values

If any of the assumptions listed above are violated, the predictions and scientific insights yielded by the analysis can be biased or misleading. Hence, prior to analyzing multiple OLS regression models for various ridership variables, this study explored the relationship between the independent variables to not only detect any multicollinearity present between variables but also determine the variables that predict the best fit in each model. A high multicollinear value present between the explanatory variables can be an issue because it indicates that the variables are highly linearly correlated, tends to predict identical results, and undermines the capability of the variables to perform effectively in an OLS regression equation (Osborne & Water, 2002). To visualize which variables were positively or negatively correlated, a Pearson correlation matrix was created and is displayed in Figure 4.1. Most of the hub environment characteristics were found to be positively correlated. For instance, built environment variables such as the number of bus stops and the length of bus routes in the 200 m buffers from bike share hubs depicted a high-level of positive correlation with each other. Also, as expected, all accessibility variables showed significant positive correlation to their initial socio-demographic variables (e.g., population and employment in the vicinity of hubs). For instance, accessibility variables that are associated with the population group, such as “Pop”, “AcsPop”, and “PopDecay” in the matrix were strongly inter-correlated. The same phenomenon could also be expected from the employment group as well. An intriguing discovery for accessibility variables that combined two socio-demographic variables (e.g. “AcsPopnEmp” and “PopnEmpDecay”) can be noted in Figure 4.1 – they show twice as strong a relationship with the original employment variable as the population variable. This association suggests that these

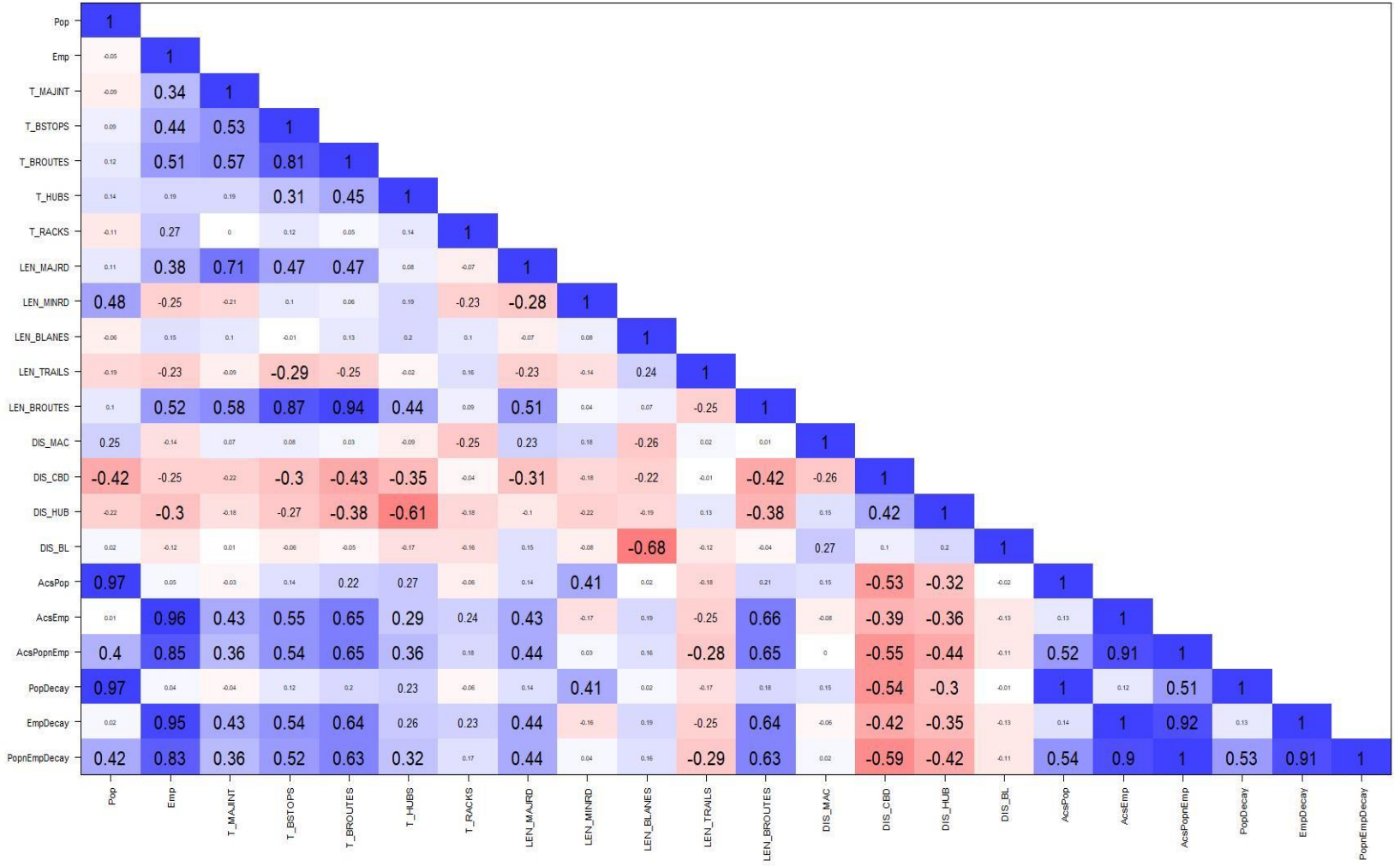

Figure 4.1: Pearson correlation matrix for all independent variables

heterogeneous accessibility variables were more likely derived from the employment variable, and the employment variable would have a greater influence in the regression analysis compared to the population variable. A negative relationship was found between the distance variables (to McMaster University and to the CBD) and most of the built environment variables as well as the accessibility variables. Specifically, the distance to major centers showed a strong inverse correlation with the accessibility variables.

As mentioned previously, highly correlated input variables (whether positively or negatively correlated) in a regression model can be problematic due to their tendency to estimate the same result. To identify important features that affect ridership at hubs, numerous OLS regression models were estimated and compared. The first estimation involved social environment variables with the built environment variables only. Then, accessibility variables were added into the model based on commonalities in how they were created. For instance, population and employment with a linear distance decay function were estimated together, while population and employment with the negative exponential distance decay function were tested together in a different regression model. For the accessibility variables that included both population and employment, a new regression model was estimated to plot them separately. Identifying diverse combinations of socio-demographic and built environment variables with accessibility variables, and estimating their effects on the ridership in each regression model was found to be challenging – it can be time-consuming and a researcher may omit a feature for the best combination of variables for a regression analysis. Therefore, the study employed Exploratory Factor Analysis (EFA) to categorize and group highly correlated variables to not only delve into

the multicollinearity in detail, but also identify a set of explanatory attributes that are more relevant and influential in regression estimations.

4.1.1 Model Specification

Exploratory Factor Analysis is a statistical technique that analyzes the latent relational structure among a set of variables and reduces a large number of variables to a smaller set (i.e., factors). In other words, instead of having to consider many variables that may be trivial, EFA categorizes input variables into meaningful groupings and describe the variance by a few summary factors. This process enables regression algorithms to operate faster and more effectively as the key variables that have significant correlation can be determined from a more compact result (Rummel, 1970). Using five summary factors, namely a five-factor solution, the explanatory variables were categorized in Figure 4.2 – the four-factor solution under-extracted components from the dataset, while the six-factor solution over-extracted. By way of explanation, the numbers displayed on top of each arrow demonstrate how strongly the variables are associated with each factor (1 = strongly correlated, 0 = no correlation, -1 = inversely correlated). MR represents a summary factor, where a solid arrow indicates a positive relationship and a dashed arrow a negative relationship. It should be noted that some variables showed a reasonable relationship with more than one factor (e.g., the number of major intersections was associated with both factors 3 and 4, and the distance to CBD with factors 2 and 5), but EFA classified these variables into groups with stronger correlation in each factor. To justify the performance of EFA and reliability of its outcome, MacCallum et al. (1999, 2001) asserted that factor loadings should explain at least 60% of the total cumulative variance. Here, the 5-factor

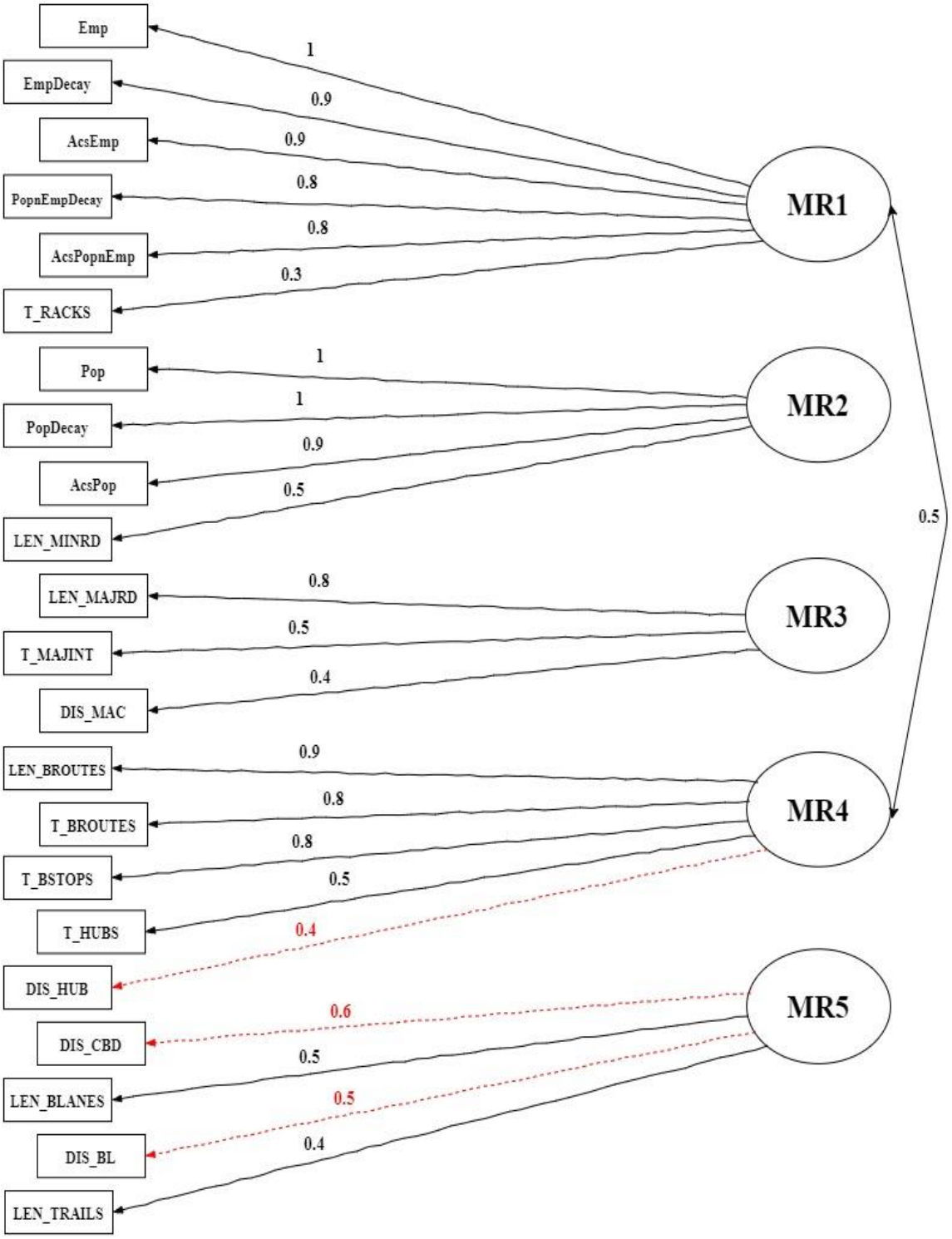

Figure 4.2: A graphic representation of a five-factor solution for explanatory variables

solution suggested that approximately 65% of the total variance in the dataset was explained cumulatively. In Figure 4.2, the first factor (MR1) explained the most variance in the dataset among other factors, accounting for 20.5% of the total variance, and it appeared to be primarily defined by the employment-related variables. The second factor (MR2) explained 16.8% of the total variance, where the variables concerned with population had high positive loadings. Interestingly, the population-related category (MR2) included the length of minor roads in its group. As this variable indicates the roads with less traffic that tend to be established in a residential neighborhood, the longer the length of minor roads in the 200 m distance from a hub suggests a greater population in the vicinity. The proportion of variance explained by the fourth factor (MR4) was approximately 14%, where the majority of the built environment variables were categorized. Especially public transit-related variables, such as length and number of bus routes, and the number of bus stops within the 200 m buffer seemed to have significant loadings on this factor. Lastly, the proportion of the variance explained by the third and the fifth (MR3 and MR5, respectively) combined was less than 13%. The third factor appeared to organize heavy-traffic related variables, while the fifth factor classified bike infrastructure-related variables. A noticeable relationship between summary factors 1 (MR1) and 4 (MR4) may propose that employees prone to make use of public transit to and from their workplace.

The findings from EFA supported the results suggested by the Pearson correlation matrix in terms of intensity of correlation between variables. However, the EFA technique enabled not only the summary factors (MR1 – MR5) to classify the variables that were

associated with each other for easier interpretation of the multicollinearity present in the dataset, but also regression examinations to perform efficiently by helping to identify the explanatory variables. Through various approaches employed to examine the independent variables that have significant influence on the ridership, the following variables were collected for robust predictive regression models: population and employment, the number of racks available at each hub, distance to McMaster University and CBD, and proximity to the nearest hub and bike lane, and length of bike lanes and bus routes within the vicinity of hubs. In this research, three sets of models for departure, arrival, and total trips were estimated using OLS regression in an effort to investigate the model's ability to predict usage for different configurations of the system. In essence, Model 1 (regular hubs + regular members) demonstrates the behavior of HBS prior to the launch of ERI program to the bike share network, Model 2 (all hubs + regular members) describes the behavior of the regular members with additional hubs added to the existing system, and Model 3 (all hubs + all members) considers the complete system which explains the effect of the ERI program. Identifying potential locations of additional hubs at representative candidate sites incorporated the complete state of HBS (e.g., all hub usage by all members) as new hubs would be added to the existing system. Thus, Model 3 was employed as the basic building block to estimate potential hub usage on the east side of the core service area to inform future expansion of bike share system.

4.2 Potential Location of Bike Hubs

After the primary environment features that influence ridership in the nine predictive models were identified, this study explored these characteristics further to

predict the potential number of trips at representative candidate sites to offer insights into future expansion of Hamilton's bike share system. In this study, the candidate locations were created on appropriate land uses by applying a continuous surface of regularly shaped cells, also known as a tessellation, on the area of interest. Larsen et al. (2013) also made use of a continuous surface, where they superimposed 300 m grid cells over the study region to determine the location of potential cycling infrastructures based on the number of cycling trips on travelled links throughout the central city of Montreal, Québec, Canada. Larsen and his research team's (2013) theory behind the grid cell approach was that the estimated number of observed and potential bicycle trips were aggregated by each grid cell, which would display the total number of estimated bicycle trips passing through the grid cell. This way, they were able to reveal the cells that more cycling trips pass through and determine potential locations for future infrastructure investments (Larsen et al., 2013). In a like manner, this study presents a GIS-based tessellation approach (e.g., the tiling of a plane built with one or more geometric shapes with no overlaps or gaps) to approximate candidate locations of future hubs in Hamilton, where high-priority cells based on predicted ridership represent the areas that should be prioritized for one or more additional hubs. However, rather than relying on the census geographies created mainly for administrative purposes (such as census geographies for the Canadian Census), this study made extensive use of GIS and Python to develop custom geographies that more accurately fit the study area. Concerning the purpose of this analysis, tessellations were developed to cover the Hamilton beneath the escarpment. To alleviate the issues of irregularly shaped geography of the study area, hexagons were chosen to create evenly spaced regularly-shaped cells for

further observation and experiment. Although there are other geometry shapes (e.g., triangles, squares, or diamonds) that could be considered to comprise the tessellations, the following suggestions outline how hexagons may be a better fit for this analysis (Birch et al., 2007):

- Hexagons can diminish sampling bias caused by edge effects of the grid shape, which is related to the low perimeter-to-area ratio of its shape. In other words, the circularity of each hexagon grid can represent curves in the irregularly shaped geography of the study area more naturally than squares or triangles. A circle has the lowest ratio among various geometries, indicating the capability of reducing sampling bias the best, but cannot tessellate to form a continuous grid. A hexagon is the most circular-shaped polygon that can form an evenly spaced grid for the given area of interest

- Any point inside a hexagon is closer to the centroid of its shape than any point in triangles or squares using the same area. This is because of the nature of more acute angles of the triangle (60°) or square (90°) versus the hexagon (120°). This benefits the researchers who weigh the centroid of each grid significantly for various reasons (i.e., assigning variables and placing a representative location within a grid)

- Hexagon grids can result in less distortion when operating across a large area due to the curvature of the earth compared to the geometry of a triangle or a square

- Lastly, detecting the neighbor grid is considerably straightforward using hexagons.

Having a fishnet (square) grid as an example, the distance of neighbor centroids in the Rook's Case (right/left/above/below) differs from the distance in the Queen's Case as the diagonal neighbors are located farther away. Meanwhile, the centroid of each neighbor for hexagon grids is equidistant with the identical edge or length of contact on each side

Considering the use of 200 m buffers to capture the surrounding features around each HBS hub in the previous part of this thesis, the area of an individual hexagon that makes up the tessellation was designed to 104,000 m^2 (200 m distance from the center of the hexagon to a side). Once the proper scope of each cell area was computed, a centroid was created for each cell to represent a candidate site for a potential bike share hub. Subsequently, buffers with the equivalent distance as the hexagonal grid (200 m) for each candidate location were developed to describe the environment features. Although the hexagonal-shaped tessellation was only constructed to shape a continuous surface over the study area, this study conducted an analysis with both hexagons and equivalent 200 m buffers to identify potential hub locations, and to compare the results in the end (Figures 4.4a, 4.4b). To entirely focus on the east of the core service area for bike share expansion, the following areas were eliminated consecutively, and the number of candidate sites was determined, respectively. For example, followed by the removal of water bodies present in the city, the unrelated geographies produced during the development of hexagon tessellation outside of the study area were disregarded. As a result, the number of candidates was reduced from 1,934 to 653. Then, the candidates on inappropriate land uses were identified using the land-use dataset obtained for the City of Hamilton, which was

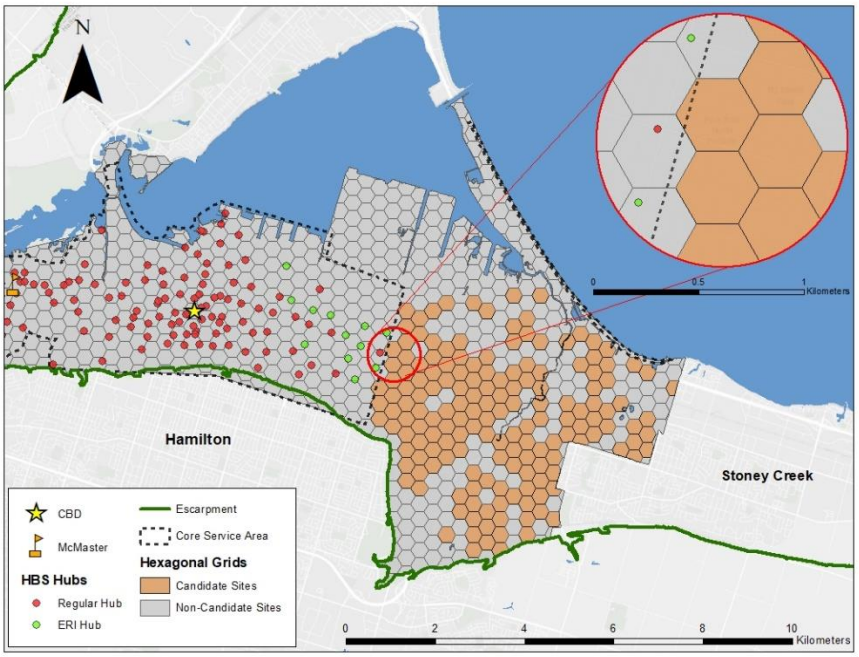

Figure 4.3 (a): Candidate locations on hexagonal grids created over the potential area of bike share expansion for both departure and arrivals trips

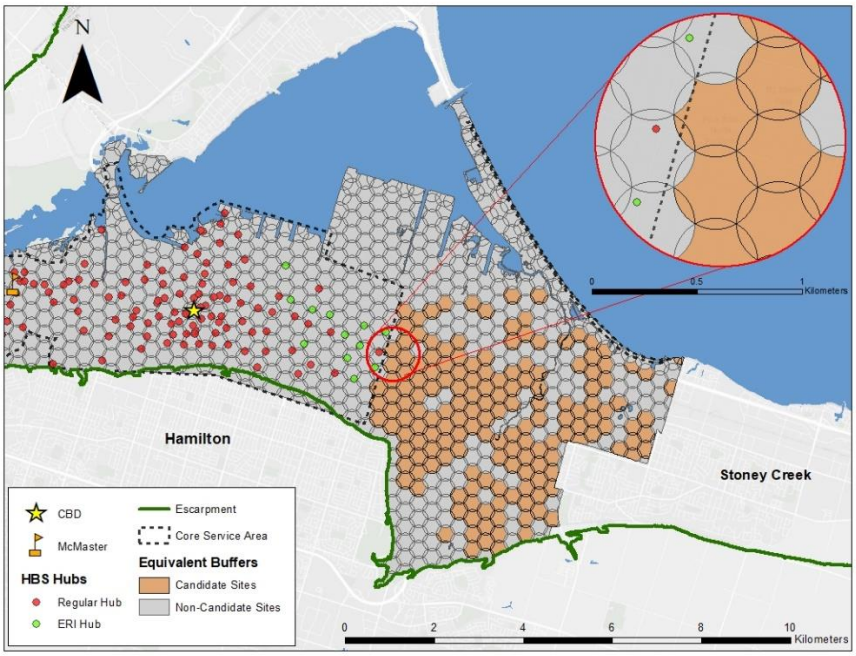

Figure 4.3 (b): Candidate locations on 200 m equivalent buffers created over the potential area of bike share expansion for both departure and arrivals trips

provided by the City of Hamilton Department of Planning and Economic Development, GIS – Planning and Analysis. The types of land uses that were relevant and sufficient to determine the potential location of new hubs were residential, commercial, institutional, and office, while the non-candidate areas included industrial, agriculture, vacant lots, open spaces, and miscellaneous. The number of candidate sites after selecting only appropriate land uses shortened to 288. Ultimately, this thesis examined the candidate sites outside of the active core service area to propose a potential area for future expansion of Hamilton's bike share. Correspondingly, the final count of 164 candidates remained as the potential sites for new hubs in the study area. When exploring the visualizations of hexagonal grids and equivalent 200 m buffers, the hexagons formed a continuous surface throughout the given area (Figure 4.4 (a)), where buffers overlapped with each other (Figure 4.4 (b)). In fact, as previously stated, buffers in Figure 4.4 (b) demonstrate how the intersected areas by the edge of its shape form a hexagon-shaped geometry in the center.

5 Results

Before examining and interpreting the results of the OLS regression analyses, the outcomes in each of the nine predictive models were found to violate some fundamental assumptions of regression analysis that were discussed in section 4.1. For instance, residual distributions depicted negative skewness (left-skewed) due to extreme outliers in the dataset, which can cause biased results. To cope with the outliers in the dataset for each predictive model, this study used Cook's distance, a multivariate model approach that investigates unusual combinations of model variables (Prabhakaran, 2016). Alternatively, the univariate method, which searches for data points with extreme values on only one

variable, could be considered. However, declaring observations outliers based on just one variable may lead to unrealistic inferences (Prabhakaran, 2016). As one of the critical multivariate approaches, Cook's distance is beneficial for this study as it computes the impact exerted by each observation on the predicted outcome in a given regression model (Jayakumar & Sulthan, 2014). Applying Cook's distance in the 2017 hub dataset, two hubs (George Street and James at Vine) were removed from Model 1 (regular hubs + regular members) and Model 3 (all hubs + all members), while one hub (George Street) was removed from Model 2 (all hubs + regular members) for all ridership. Consequently, the distributions of residuals for all three models became normally distributed, which is one of the underlying assumptions for regression analysis (Figure 5.1). In fact, the hubs removed as outliers recorded constantly the lowermost trips in the ridership data. Figure 5.1 illustrates the distributions of residuals before and after removing the outliers (George Street and James at Vine hubs) from Model 3. Instead of displaying the improvement of residual distributions after eliminating outliers for all nine models, departure trips from Model 3 were chosen for consistency. In this figure, the peaks of trendlines show a significant difference in terms of height and width, where the peak is more centered once the outlier was removed. This is an indication of model's significant improvement in terms of generating a more consistent and less biased results. The linearity between the independent and dependent variables was confirmed once the normality of the residual histogram was stabilized by removing outliers from the predictive models. As another key assumption of regression analysis, multicollinearity between explanatory variables in

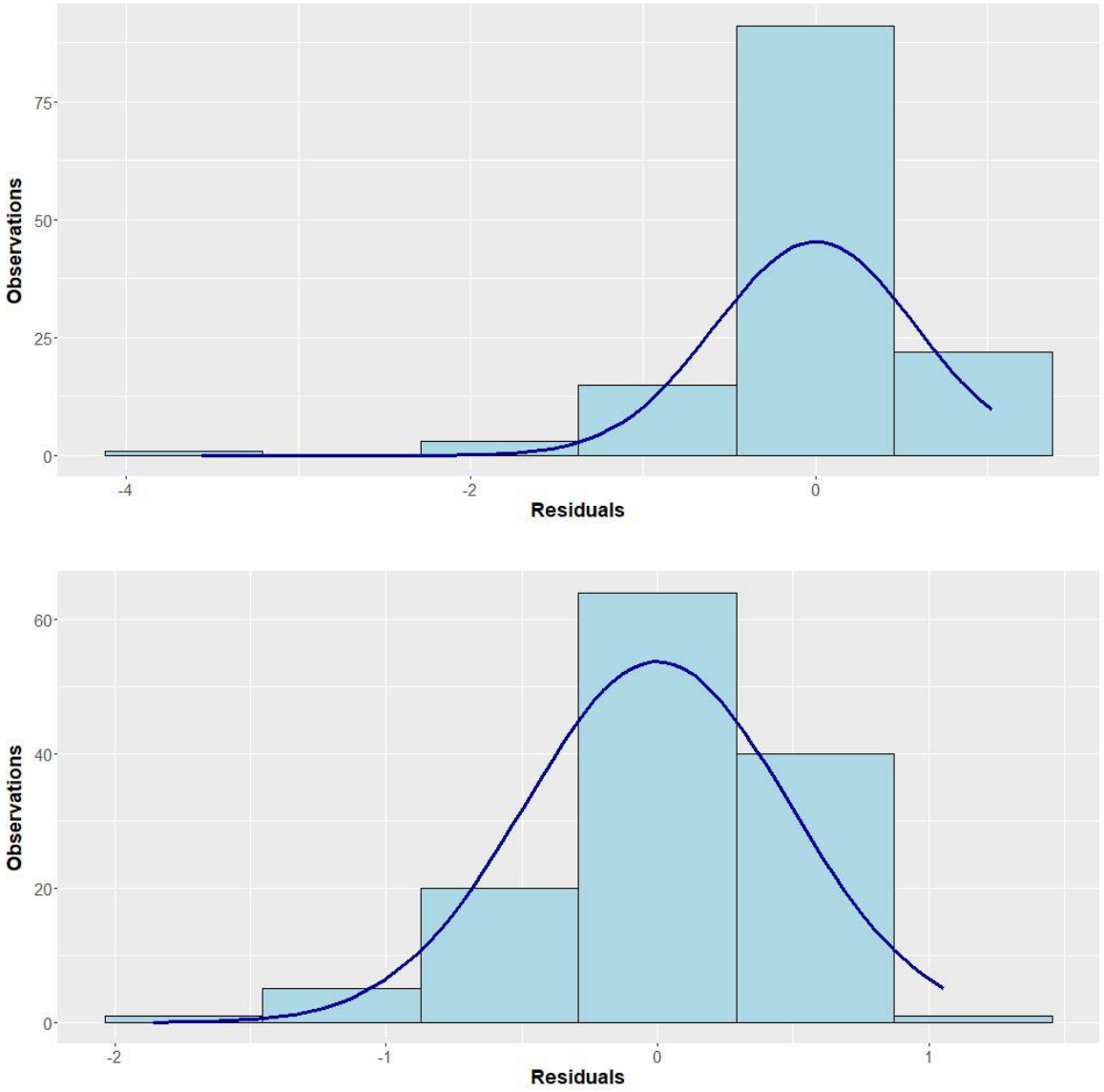

Figure 5.1: Distributions of residual before (top) and after (bottom) removing outliers for departure trips from Model 3 (all hubs + all members)

each OLS model was explored. A high multicollinearity value causes a problem when interpreting the outputs from the analysis because it implies that two or more variables are strongly linearly related that they tend to predict the identical phenomenon. On this account, this study estimated various combinations of variables based on correlations to

not only identify the most effective factors but also assure that there were no or only acceptable correlation present in the models. In addition, Variance Inflation Factor (VIF) was also evaluated to quantitatively display the value of multicollinearity between predictor variables, which were confirmed under 2.5 – VIF greater than or equal to 5 is problematic as it can increase the variance of the regression coefficients (Kock, & Lynn, 2012). To delve into any spatial autocorrelation in the dataset, Moran's I spatial analysis was investigated for each model. This spatial analysis tests if the dataset is spatially clustered or randomly dispersed throughout the space by adding spatial weights into a regression analysis (Stieve, 2012). Although the ridership dataset appears to be relatively clustered due to the locations of Hamilton's bike share hubs (which were initially deployed based on population density), Moran's I analysis on residuals proposed a random distribution of the predicted trips, and therefore, there is no autocorrelation and spatial regression analysis is not required.

5.1 Regression Model Results

Taking the key elements for regression analysis into consideration, Tables 5.1 – 5.3 present the results for the three sets of multiple OLS regression models for three different types of users and hubs. Through the nine predictive models created for different configurations of the system, the effect of 12 additional hubs on the existing bike share network was explored by Model 1 (regular hubs + regular members) and Model 2 (all hubs

Table 5.1: Summary statistics table for Model 1 estimation results without outliers

Dep. Variable: ln (trips + 1) / Variables	Model 1 (regular hubs + regular members)					
	A: Departures		B: Arrivals		C: Totals	
	Coe.	*t* statistic	Coe.	*t* statistic	Coe.	*t* statistic
Intercept	8.218	33.973***	8.120	31.656***	8.866	36.602***
Social Environment						
Population ($\times 10^{-3}$)	0.101	0.936	-0.111	-0.970	<0.001	0.006
Employment ($\times 10^{-3}$)	0.118	2.381*	0.122	2.324*	0.112	2.405*
Built Environment						
ERI hub (dummy)	-	-	-	-	-	-
Length of bike lanes ($\times 10^{-3}$)	0.007	0.055	-0.050	-0.375	-0.017	-0.136
Length of bus routes ($\times 10^{-3}$)	0.015	0.762	0.015	0.692	0.015	0.744
Hub racks	0.038	3.506***	0.044	3.847***	0.041	3.804***
Distance to McMaster ($\times 10^{-3}$)	-0.229	-7.703***	-0.176	-5.594***	-0.205	-6.897***
Distance to CBD ($\times 10^{-3}$)	-0.083	-1.899*	-0.056	-1.210	-0.073	-1.661*
Distance to hub ($\times 10^{-3}$)	-0.080	-0.264	-0.375	-1.171	-0.193	-0.637
Distance to bike lanes ($\times 10^{-3}$)	-0.458	-2.933**	-0.425	-2.566*	-0.434	-2.772**
Trips	308,643		311,699		620,342	
Adjusted R^2	0.599		0.533		0.579	

Significance levels: '***' = p < 0.0001, '**' = p < 0.001, '*' = p < 0.01, '*' = p < 0.05, ' ' = p < 1
All coefficients and t-values are rounded up to the third decimal place.

Table 5.2: Summary statistics table for Model 2 estimation results without outliers

Dep. Variable: ln (trips + 1) / Variables	Model 2 (all hubs + regular members)					
	A: Departures		B: Arrivals		C: Totals	
	Coe.	*t* statistic	Coe.	*t* statistic	Coe.	*t* statistic
Intercept	8.270	27.968***	8.225	26.761***	8.938	30.294***
Social Environment						
Population ($\times 10^{-3}$)	0.112	0.923	-0.098	-0.773	0.013	0.108
Employment ($\times 10^{-3}$)	0.109	1.952*	0.111	1.903*	0.109	1.955*
Built Environment						
ERI hub (dummy)	-0.860	-4.521***	-0.955	-4.832***	-0.909	-4.788***
Length of bike lanes ($\times 10^{-3}$)	-0.171	-0.974	-0.238	-1.306	-0.201	-1.147
Length of bus routes ($\times 10^{-3}$)	0.010	0.461	0.010	0.436	0.010	0.465
Hub racks	0.042	3.507***	0.047	3.803***	0.045	3.748***
Distance to McMaster ($\times 10^{-3}$)	-0.217	-7.734***	-0.175	-6.000***	-0.197	-7.048***
Distance to CBD ($\times 10^{-3}$)	-0.100	-2.568*	-0.077	-1.917*	-0.090	-2.318*
Distance to hub ($\times 10^{-3}$)	-0.264	-0.739	-0.537	-1.445	-0.367	-1.027
Distance to bike lanes ($\times 10^{-3}$)	-0.838	-2.134*	-0.821	-2.011*	-0.831	-2.120*
Trips	314,105		316,926		631,031	
Adjusted R^2	0.722		0.688		0.714	

Significance levels: '***' = p < 0.0001, '**' = p < 0.001, '*' = p < 0.01, '*' = p < 0.05, ' ' = p < 1
All coefficients and t-values are rounded up to the third decimal place

Table 5.3: Summary statistics table for Model 3 estimation results without outliers

Dep. Variable: ln (trips + 1) / Variables	Model 3 (all hubs + all members)					
	A: Departures		B: Arrivals		C: Totals	
	Coe.	*t* statistics	Coe.	*t* statistics	Coe.	*t* statistics
Intercept	8.303	32.394***	8.242	31.144***	8.963	35.232***
Social Environment						
Population ($\times 10^{-3}$)	0.091	0.806	-0.094	-0.812	0.004	0.032
Employment ($\times 10^{-3}$)	0.147	2.835**	0.146	2.732**	0.146	2.832**
Built Environment						
ERI hub (dummy)	-0.833	-4.681***	-0.897	-4.881***	-0.864	-4.894***
Length of bike lanes ($\times 10^{-3}$)	0.037	0.296	-0.010	-0.073	0.019	0.150
Length of bus routes ($\times 10^{-3}$)	0.002	0.108	0.002	0.106	0.002	0.113
Hub racks	0.036	3.228**	0.042	3.598***	0.039	3.515***
Distance to McMaster ($\times 10^{-3}$)	-0.186	-7.087***	-0.151	-5.564***	-0.170	-6.522***
Distance to CBD ($\times 10^{-3}$)	-0.094	-2.391*	-0.073	-1.806$^{\cdot}$	-0.086	-2.187*
Distance to hub ($\times 10^{-3}$)	-0.414	-1.252	-0.700	-2.051*	-0.525	-1.601
Distance to bike lanes ($\times 10^{-3}$)	-0.422	-2.685**	-0.365	-2.248*	-0.389	-2.490*
Trips	345,265		344,440		689,705	
Adjusted R^2	0.738		0.707		0.732	

Significance levels: '***' = p < 0.0001, '**' = p < 0.001, '*' = p < 0.01, '$^{\cdot}$' = p < 0.05, ' ' = p < 1
All coefficients and t-values are rounded up to the third decimal place

+ regular members), while the introduction of the ERI program on HBS hub usage was examined by Model 3 (all hubs + all members). The explanatory variables, except for the number of racks available, were scaled by dividing by 1,000 to improve coefficient interpretation. Independent variables that are not listed in the tables were not considered for further analysis due to insignificant contribution to the models' outcome or violation of key assumptions of multivariate linear regression examination. For instance, accessibility variables were not included in the final predictive models as these variables not only demonstrated a high collinearity with the social environment variables (population and employment), but also found less significant through EFA and regression examinations. A dummy variable "ERI Hub" was also created and employed in the analysis in order to distinguish between ERI hubs and regular hubs. Figure 5.2 illustrates how the addition of the ERI dummy variable further improved the distribution of the residual histogram from Figure 5.1, indicating that more reliable and precise regression outputs would be produced from the predictive models.

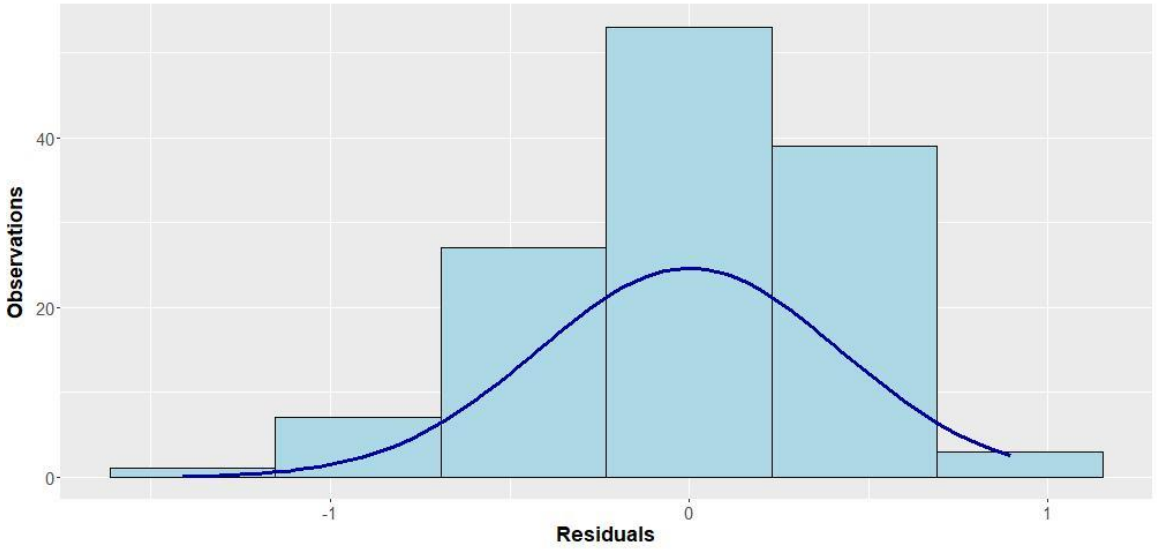

Figure 5.2: Distribution of residuals with the addition of the ERI dummy variable for departure trips for Model 3

The effect of each explanatory variable in terms of predicting ridership in Tables 5.1 – 5.3 suggests that the variables are largely consistent across the models. For example, the primary contributing factors, such as hub racks and distance to the McMaster campus, were found consistently significant, whereas population and length of bike lanes and bus routes in the vicinity of hubs had no impact on hub usage throughout the models. In order to explore the performance of each main model, the adjusted R-squared values were investigated. The provided adjusted R-squared values in the tables explained the capability of models to define the variance of the ridership (number of departures, arrivals, and total trips) with the given independent variables. Although multiple R-squared values for each model could be considered to inspect the model's performance and consistency, the proposed multiple R-squared values could be misleading due to their tendencies to increase by addition of new variables. In contrast, an adjusted R-squared value only increases or

decreases depending on the validity of the new variable to the prediction of the model. When scrutinizing the behavior of the main predictive models with adjusted R-squared values, a consistent improvement in adjusted R-squared values were observed from approximately 60% of variance explained by Model 1 to 73% by Model 3. Considering consistent statistical significance of explanatory variables between the models, and capability of Model 3 to better explain the variance in the dataset compared to Models 2 and 1, the remainder of the thesis interpreted the OLS regression results for Model 3 (Table 5.3).

5.1.1 Social Environment Characteristics

Table 5.3 exhibited contrasting results for the two socio-demographic variables (population and employment), where the working people aged from 15 to 64 captured within the 200 m buffers from each hub suggested a stronger impact on ridership than population. Such an outcome is a compelling result considering how population density was an important factor in determining initial HBS hub locations. Moreover, the insignificant population variable in this study differs from the findings of many other articles on bike share systems that suggest the critical role of population in their models to determine the usage at the hub level (e.g., Daddio, 2012; Efthymiou et al., 2012; El-Assi et al., 2017; Hampshire & Marla, 2012; Tran et al., 2015; Zhang et al., 2017). Possible reasons for insignificant population variable include, but are not limited to, the student population near McMaster University that was not accounted for through the 2016 Canadian Census (Scott & Ciuro, 2019), and a policy that bike share members can pick up and drop off bikes at any location in the service area by paying an out-of-hub fee.

Nevertheless, the positive coefficients of the population variable imply that the number of departures, arrivals, and total trips tend to increase as population increases within 200 m of a hub. On the contrary, the significant role of employment proposes that many working populations can be considered as a surrogate for the attractiveness of a destination.

5.1.2 Built Environment Characteristics

Two bike infrastructure variables were statistically insignificant in the regression models – length of bus routes and length of bike lanes. However, for length of bus routes, its positive coefficient indicated that if there were longer bus routes in the vicinity of hubs, then hub usage would also increase. Another interpretation that one could take from this result is that some HBS users also make use of the bikes as part of a multimodal trip. With respect to the relationship between ridership and presence of bike lanes near hubs, interesting results can be drawn – length of bike lanes in the vicinity of hubs was found insignificant, whereas distance to nearest bike lane was found important. This suggests that bike share users consider the presence of bike lanes close to hubs as attractive characteristics, but are not concerned with the length in the neighborhood of the hubs. Yet, retaining these insignificant variables contributed to improvement of the model's performance in general for all three models in Table 5.3. There are several predictor variables that were found statistically significant across Model 3 (A, B, and C): the number of racks available at each hub, the hub's distance to McMaster University, and the ERI hub dummy variable. This indicates that regardless of the purpose of the bike usage, or whether a user chose to rent or return, these variables always play a significant role in influencing ridership at hubs. The amount of racks available at each of HBS hub could demonstrate

that there is a greater likelihood of finding a bike with increasing racks. Furthermore, it could reflect the fact that the racks were scaled to account for users' demand – hubs with greater number of racks were placed in the areas of high demand. The effective relationship between the distance to McMaster University and ridership reveals that the university is both an origin and destination for trips; less trips were recorded as the distance to McMaster University increased. In fact, the top two hub locations with the highest trip counts were on McMaster University's campus. Similarly, the ERI dummy variable created to distinguish between regular and ERI hubs depicted a strong negative correlation with the ridership in the regression analysis, indicating that ERI hubs attract fewer trips compared to the regular hubs. This could explain the early state and progression of new hubs (Model 2) and the system (Model 3) to the existing bike share network (Model 1). The variables that showed a reasonably strong impact in the regression examination for Model 3 were the distance to CBD and the nearest hub. The influence of the distance to CBD variable on bike usage was also found to be influential in Scott & Ciuro (2019)'s findings, where they found that the number of daily trips recorded at each hub in 2015 decreased as the distance from CBD increased. In table 5.3, this variable displayed a more significant importance in Model 3 A (departure trips) than B (arrival trips), which could suggest that bike share users tend to have CBD as a departure point. In contrast, the distance to the nearest hub was insignificant across the models except for in model B. However, its small impact on arrival trips proposes that bike users' concern with the distance between the hubs when returning the bikes after use. This relationship may be linked to the outcome of the rack availability, where a stronger influence was found for arrival trips compared to departures in Table 5.3,

because users would desire to be able to return their bikes successfully without having to travel to a different hub location as the hub is out of space for parking.

In terms of improvement of the models' performance, the importance of applying the log-transformation on dependent variables can be addressed and quantitatively justified by comparing the predicted values from regression models and the actual number of trips taken at each hub. This is because a model may under- or over-estimate the outcome and cause an irrational result without the log-transformation. Taking departure trips from Model 3 for instance, the minimum value for the estimated predicted trips before the log-transformation applied was a negative value (approximately -449 trips), which is unreasonable for the trip counts at hubs. On the other hand, when the natural logarithm transformation was applied on ridership, positive minimum and maximum predicted values were obtained (approximately 233 and 10,103 trips, respectively). The average difference between the number of actual trips and predicted trips was around 200, which would be acceptable considering the total number of departure trips was nearly 350,000.

5.2 Predicting Usage for Potential Hub Expansion

Using the regression results for Model 3 (all hubs + all members), this research estimated potential trips at 164 candidate locations to inform future expansion of the public bike share system in Hamilton. Model 3 was chosen for this analysis because the new expanded system would be based on the behavior of the complete state of the existing bike share network. However, as the negative coefficients of the ERI dummy variable suggested the negative impact of the early progression of new hubs, potential hubs in the expanded

area could also experience less ridership until they become familiar as regular hubs. Taking this outcome from regression analyses for predictive models into account, this study considered a best-case scenario where the outcome reflects usage after some time has passed. In other words, newly added hubs in the expanded areas were evaluated as regular hubs and the bike share network would be expanded the same way the initial HBS was originally set up. Accordingly, there were some changes conducted to several variables for further analysis. For instance, a value of zero was given to ERI dummy variables because the potential hubs would behave as regular hubs, meaning that the hubs do not need to be distinguished considering the equity of citizens in the city nor target low-income populations. Furthermore, the policy sensitive variables that also required adjustments were the availability of the number of racks at each hub and the closest distance measured between hubs within the core service area. To assign a reasonable number of racks to each candidate, the mean value of hubs across the service area was computed and a constant value of ten was obtained. This is because HBS hubs have a diverse number of accessible racks dependent on their size (e.g., a hub with a greater demand for bikes expects a higher number of racks). Lastly, instead of measuring distances from a hub to the nearest existing hub in the core service area, a distance from a candidate to the nearest candidate site was calculated to demonstrate predictive ability of the models for expansion of the system. Initially, this variable was created to describe that the farther away from a hub in the existing system, there are fewer trips predicted for that hub. However, as this study aims for the gradual expansion of the system eastward, the distance between candidate locations was investigated. In order to analyze a reasonable distance between the candidates, the

initial configuration of HBS, in the range of 300 – 600 m between hubs (Scott & Ciuro, 2019), was taken into consideration. Respectively, the shortest distance between candidates measured 341 m, the longest distance was 729 m, and the mean was 354 m. In fact, a constant distance of 347 m was obtained between the top ten candidate locations, which supports the idea of acknowledging the initial configuration of HBS in this research.

Potential ridership was computed at representative candidate sites using the two different custom geographies (hexagons and equivalent 200 m buffers). Despite the fact that the centroids of the continuous surface were used to create the 200 m equivalent buffers for the given area of interest, this study assessed potential trips for both approaches to compare the results at the end. Here, the explanatory variables adopted in Model 3 (all hubs + all members), as well as the appropriate policy sensitive variables described above, were calculated for both approaches. Consequently, intriguing results were derived, where paired *t*-tests suggested that the mean of the differences between the two approaches was around three for departure trips ($p < 0.0001$) and one for arrival trips ($p < 0.0001$). Furthermore, the top ten candidate sites identified by both custom geographies were practically the same – only differences were the order of the ranks for both models, and the last candidate for departure trips (Table 5.4).

Table 5.4: Comparison of potential departure and arrival trips estimated by hexagons and buffers at top ten candidate sites

Rank	Trips by Hexagons				Trips by Buffers			
	Grid ID	**Departures**	**Grid ID**	**Arrivals**	**Grid ID**	**Departures**	**Grid ID**	**Arrivals**
1	AR-23	660	AQ-27	807	AR-23	670	AR-25	802
2	AQ-27	656	AQ-26	807	AQ-27	663	AQ-26	802
3	AQ-26	648	AR-25	804	AQ-26	656	AQ-27	802
4	AR-24	635	AR-23	803	AR-24	646	AR-23	801
5	AR-25	633	AR-24	801	AR-25	641	AR-24	795
6	AS-23	626	AR-26	785	AS-23	634	AR-26	784
7	AQ-25	620	AQ-25	782	AQ-25	625	AQ-25	779
8	AR-26	609	AS-23	769	AR-26	615	AS-22	769
9	AS-24	598	AS-22	766	AS-24	605	AS-23	766
10	AT-24	578	AS-24	754	AS-22	585	AS-24	751

Note. Grid ID = the unique identification codes for the top ten candidate sites provided when the hexagonal tessellations were created

When examining the rounded departure and arrival trips in Table 5.4, a constant decrease in potential ridership from the first candidate site to the last can be observed. However, the departure model shows a smaller number of trips estimated in general than the arrivals, which could have been caused by the stronger effects of distance to McMaster University and the CBD for the departure model. For departures, there were eight candidate sites with over 600 trips predicted in the east side of the core service area. The 9th and 10th candidates were found slightly farther away from the core service area with less than 600 trips estimated. Similarly, the arrival model predicted over 800 trips for the first five candidate sites, followed by 750+ trips for the subsequent ranks. Nevertheless, the results for both models suggested that the candidate sites located further away from the core service area had a fewer number of usages computed. Considering the redundancy of adopting the 200 m equivalent buffer approach and the advantages of using hexagonal grids (e.g., the tessellation approximates the area around the grids as a potential location for a new hub), this study used the tessellation technique to illustrate the candidate sites of potential hubs for future expansion of the bike share system in the city (Figure 5.3 (a) & (b)).

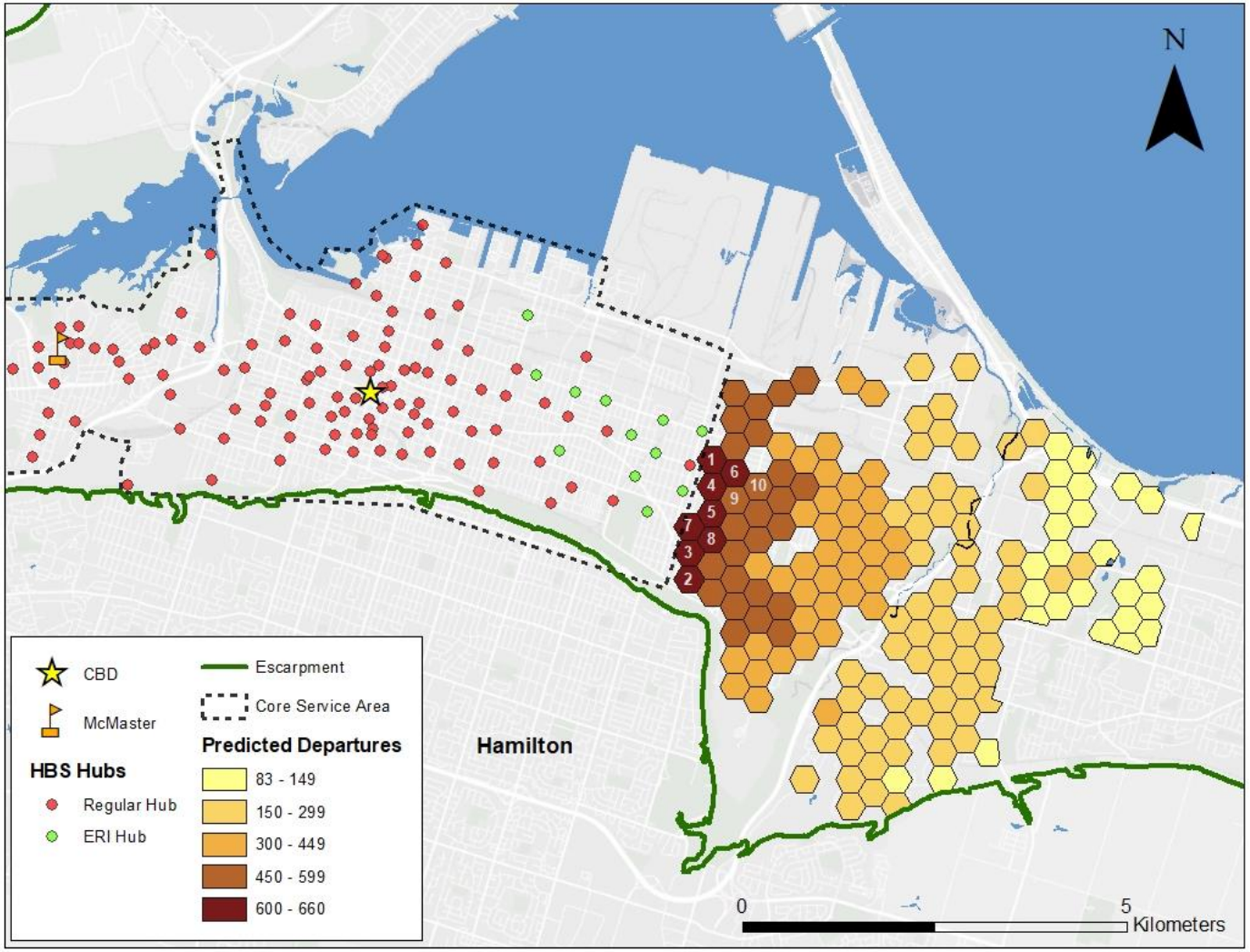

Figure 5.3 (a): Top ten candidate locations for departure model

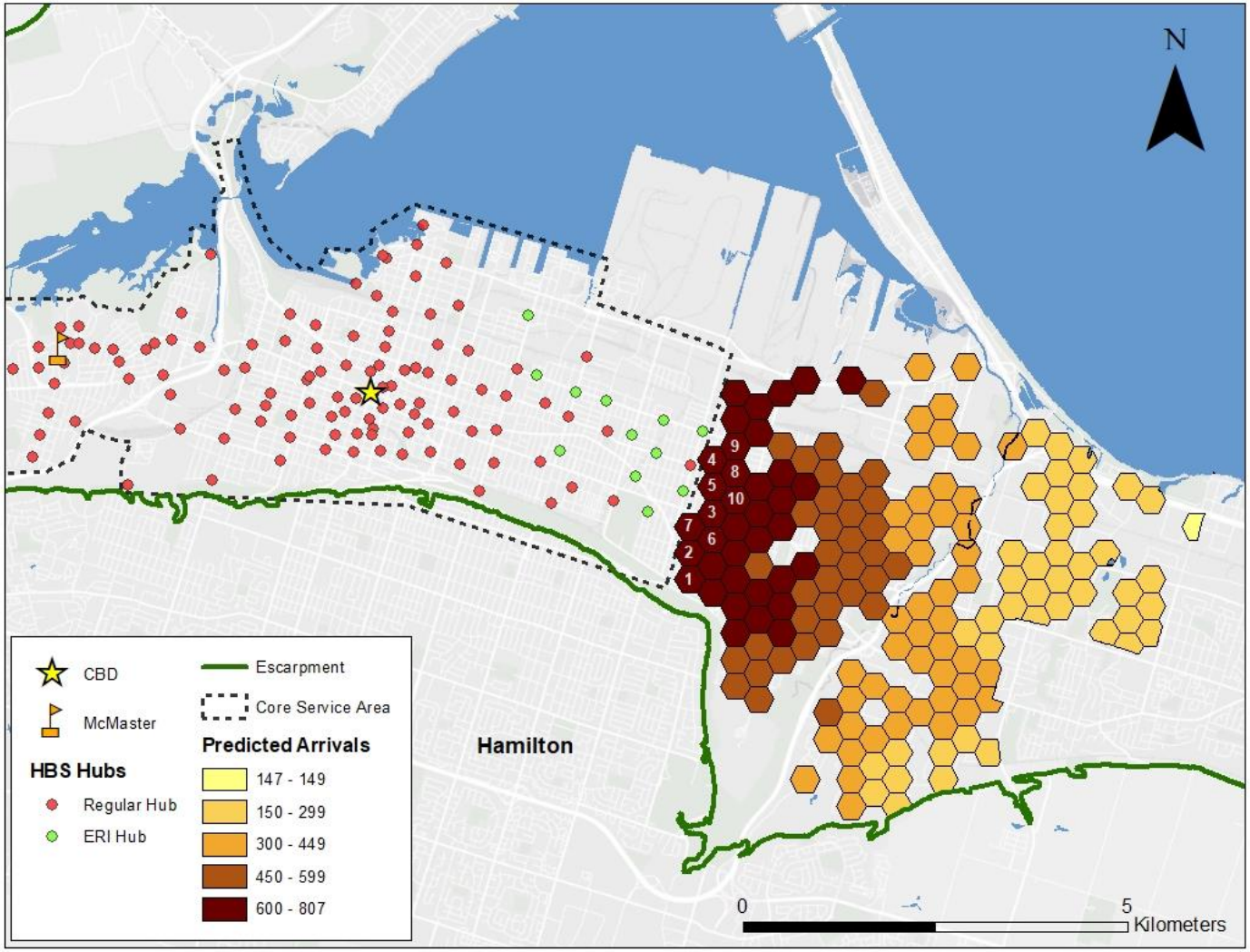

Figure 5.3 (b): Top ten candidate locations for arrival model

6 Conclusion

6.1 Summary of Findings

This thesis used annually aggregated bike share GPS data to explore the determinants that have an impact on users' bike share usage at hub level and estimated potential trips at representative candidate sites to inform the future expansion of the public bike share system in Hamilton, Ontario. Annual trip departures, arrivals, and totals were aggregated using HBS bicycle data for all hubs in the core service area from January 1, 2017, to December 31, 2017. By estimating diverse combinations of regression models along with EFA for the given dataset, the relationship between ridership and assorted variables were examined to create predictive bicycle usage models. Three sets of main predictive linear regression models were developed for three different scenarios depending on the type of hubs and members to investigate the model's ability to predict ERI usage for different configurations of the system – Model 1 (regular hub + regular members) incorporated the total ridership counted at each regular hub by regular members; Model 2 (all hubs + regular members) built upon Model 1 by incorporating the ERI hubs in service, but usage by only regular members was considered; and lastly; Model 3 (all hubs + regular members) consisted of Model 2 with the hub usage by all members (both regular and ERI members). Through the nine predictive models, it was possible to analyze the impact of various factors on ridership for the existing hubs in the core service area, while the effect of the ERI program was discerned. In essence, Model 1 demonstrates the behavior of HBS prior to the launch of ERI program to the bike share network, Model 2 describes the behavior of the regular members with additional hubs in the existing system, and Model 3

considers the complete system which explains the effect of ERI program. To understand the primary environment factors that influence the ridership, numerous environment variables (socio-demographic, built environment, and accessibility) within a 200 m buffer from a hub were explored based on the purpose of trips - starting, ending, or both combined. As a result, unlike the findings of many other bike share studies that suggest population density variable as a significant variable (e.g., Daddio, 2012; Efthymiou et al., 2012; El-Assi et al., 2017; Hampshire & Marla, 2012; Tran et al., 2015; Zhang et al., 2017), it did not play a significant role in influencing the hub usage in this study. Instead, the following three variables were found to be consistently significant throughout the models: number of racks available at a hub and distance to McMaster University (ERI hub dummy variable was found significant in Models 2 and 3). The positive relationship between the number of racks and the hub usage indicated that the availability of the bikes does have a notable impact on the decision of an individual to select a specific hub - a greater number of racks implies the likelihood of greater bike availability. It should be noted that racks at hubs were scaled to accommodate demand in the areas. The negative coefficient between the distance to McMaster University and ridership demonstrates that being far away from the campus negatively influences usage at a hub, suggesting the importance of the university in attracting and generating trips. In fact, the top two hub locations with the most trips recorded were on the McMaster campus. The negatively correlated ERI dummy variable implied that there were less departing and attracting trips at ERI hubs compared to the regular hubs. This dummy variable could also demonstrate the early state and progression of new hubs (Model 2) and the system (Model 3) to the existing bike share network (Model

1). Interestingly, an environment characteristic that showed a strong relationship with departure trips across the models was distance to the CBD. For example, this variable was significant for Models 1A, 2A, and 3A, which indicates that hub usage increases close to a major source of activity. The distance to the closest bike lane measured was found effective throughout the predictive models, especially for 1A, 1C, and 3A, while employment variable was found significant in Models 1 and 3 (A, B, and C). With regards to the strong impact of distance to nearest bike lane on bike usage, one could conclude that bike share users consider the presence of bike lanes near hubs as a significant feature that influence the hub usage. The working population variable was only slightly effective in Models 1 and 3, which explains the high number of employees starting or ending their trips near the regular hubs, and how the presence of the ERI hubs did not have a large impact on their usage. When the performance of each of the three models was reviewed for departures, arrivals and totals, Model 3 was found to predict the best fit for ridership followed by Model 2 and Model 1 by observing adjusted R-squared values.

Considering the fact that the expansion of HBS takes into account the complete state of the existing bike share network, the number of potential trips at 164 candidate sites were predicted using the Model 3 regression equation. When adopting and calculating the predictor variables used in Model 3 for the custom geographies to predict the potential trips at representative candidate sites, some variables were explored in detail and modified to demonstrate predictive ability of the models for expansion of the system. In this process, an important assumption was constructed, where the potential hubs in the expanded area were presumed to behave as regular hubs in the existing service area that bike users are

familiar with the new hubs. This is due to the fact that the strong negative relationship of the ERI dummy with the ridership suggests the negative impact of the early progression of new hubs, and that potential hubs in the expanded area could also experience less ridership until they become familiar as regular hubs. Accordingly, a value of zero was set for the ERI dummy variable and ten (mean value of racks at HBS hubs) for the number of racks available. Instead of using the distance to nearest hub variable, this research measured the distances between candidate sites by reflecting the initial configuration of HBS where the distance between hubs were in range of 300 – 600 m. Subsequently, the top ten closest candidate locations were found approximately 347 m apart from each other, supporting the idea of concerning the initial network of HBS. As the number of potential ridership were estimated using custom geographies (hexagonal tessellations and 200 m equivalent buffers), the outcome produced by the buffer approach was found superfluous, and thus, this study proceeded the analysis using the tessellation technique only. Through a series of estimation and evaluation, the top ten potential sites for additional hubs and the expansion of the ground coverage were placed by the east side of the core service area, where the potential hubs located further away had a fewer number of usage computed.

6.2 Limitations and Future Research

Despite the important findings from a series of regression models, the investigation included several limitations when exploring the relationship between bike hub usage and surrounding characteristics for the existing hubs in the core service area. For instance, the total variance explained by each variable in the regression examinations (approximately 60% for Model 1, 70% for Model 2, and 73% for Model 3) was adequate to demonstrate

its consistency as a trip prediction model, but future work could consider employing new variables to improve the model's performance further. With regards to the explanatory variables created in the vicinity of HBS hubs, one could argue that socio-demographic variables should have taken the different age groups into consideration. In other words, hypothetically, if a population variable is mainly composed of age groups that are less likely to use bike share in contrast to the age groups that are more likely to use them, the significance of the variable as well as the behavior of the models could have produced different results. However, Winters and her research team (2019) stated that only a nuanced difference was found in terms of bike usage by different age groups from 16 to 34 and 35 to 54 in Vancouver, Canada. In addition, the variables that were altered for the candidate sites in the expanded areas were policy sensitive, indicating that other researchers could conduct an experiment with different number of racks or distance between the candidate sites to observe if they produce better prediction. However, the strength of these effect could be incomparable since every study is different in design of their research in terms of modeling approach, data collection and aggregation, and predictor variables.

A potential improvement one can propose includes applying a different form of transformation, instead of a natural log transformation on the dependent variables, which might be more suitable for the given dataset as well as the area of study. Another direction that future work could consider is determining the potential location of bike share hubs at a different zone level. Compared to the method that manipulates the geographies of the area of interest, and investigates the surrounding environments of each candidate site, the street network representing the centerline of the street can be analyzed for more expansive

examination. Among diverse types of roads, those that are classified as major roads in the city can be taken into account, even though candidate sites nearby highways might be improper as a potential hub due to limited environment features that could attract bike users. Limitations also persist when utilizing the multi-factor usage prediction models for the existing core service area to predict potential hub usage in the expanded areas. This implies that the behavior in the new region would equate and react uniformly as the behavior in the existing area, which could bring large prediction error (Liu et al., 2017). Furthermore, there might be other important attracters affecting the hub usage, which could not be considered in this study. Examples include hotspots of social activity (e.g., commercial areas in downtown Stoney Creek) that cannot be easily identified and obtained from open-data sources. Ultimately, it should be noted that this project deliberated the results of various analyses in the context of one bike share system with an annual hub usage data. Considering other studies are concerned with modeling shorter time frames (e.g., daily or monthly), the year data used for this study differs by the time frame and presents important and significant results. Future research can take a step further by incorporating multiple years of operational GPS data to validate the findings of incremental growth over time, or compare predictions with different cities to compare the impact of environment characteristics on ridership and the location of potential bike hubs to promote sustainable public transportation.

References

Birch, C. P. D., Oom, S. P., & Beecham, J. A. (2007). Rectangular and hexagonal grids used for observation, experiment and simulation in ecology. *Ecological Modelling*, 206, 347-359.

Campbell, C. (2017). Two Dundas bike share racks replaced by virtual hub downtown. [online] HamiltonNews.com. Available at: https://www.hamiltonnews.com/news-story/7497636-two-dundas-bike-share-racks-replaced-by-virtual-hub-downtown/ [Accessed 21 July. 2020].

Chen, Z., van Lierop, D., & Ettema, D. (2020). Dockless bike-sharing systems: What are the implications? *Transport Reviews*, 40, 333-353.

Daddio, D. W. (2012). Maximizing bicycle sharing: An empirical analysis of capital bikeshare usage. Master's Thesis at University of North Carolina at Chapel Hill. https://doi.org/10.17615/qv32-b860

Echo, B. (2018). Van Wagners Hub – Hamilton Bike Share (SoBi) – Bicycle Rentals on Waymarking.com. [online] Available at: https://www.waymarking.com/waymarks/WMXQJK_Van_Wagners_Hub_Hamilton_Bike_Share_SoBi/ [Accessed 21 July. 2020].

Efthymiou, D., Antoniou, C., & Tyrinopoulos, Y. (2012). Spatially aware model for optimal site selection: Method and application in a Greek mobility center. *Transportation Research Record*, 2276, 146–155.

El-Assi, W., Mahmoud, S. M., & Habib, K. N. (2017). Effects of built environment and weather on bike sharing demand: A station level analysis of commercial bike sharing in Toronto. *Transportation,* 44, 589-613.

Faghih-Imani, A., Eluru, N., El-Geneidy, A. M., Rabbat, M., & Haq, U. (2014). How land-use and urban form impact bicycle flows: Evidence from the bicycle-sharing system (BIXI) in Montréal. *Journal of Transport Geography*, 41, 306-314.

Faghih-Imani, A., & Eluru, N. (2015). Analysing bicycle-sharing system user destination choice preferences: Chicago's Divvy system. *Journal of Transport Geography*, 44, 53-64.

Fishman, E. (2016). Bikeshare: A review of recent literature. *Transport Reviews*, 36, 92-113.

García-Palomares, J. C., Gutiérrez, J., & Latorre, M. (2012). Optimizing the location of stations in bike-sharing programs: A GIS approach. *Applied Geography*, 35, 235–246.

Guidon, S., Reck, D. J., & Axhausen, K. (2020). Expanding a(n) (electric) bicycle-sharing system to a new city: Prediction of demand with spatial regression and random forests. *Journal of Transport Geography*, 84, 102692.

Hampshire, R. C., & Marla, L. (2012). An analysis of bike sharing usage: Explaining trip generation and attraction from observed demand. [online] Available at: https://nacto.org/wp-content/uploads/2012/02/An-Analysis-of-Bike-Sharing-

Usage-Explaining-Trip-Generation-and-Attraction-from-Observed-Demand-Hampshire-et-al-12-2099.pdf/ [Accessed 21 July. 2020].

Hansen, W. G. (1959). How accessibility shapes land use. J*ournal of the American Institute of Planners*, 25, 73–76.

Heinen, E., van Wee, B., & Maat, K. (2010). Commuting by bicycle: An overview of the literature. *Transport Reviews*, 30, 59–96.

Jayakumar, G. S. D. S., & Sulthan, A. (2014). Exact distribution of Cook's distance and identification of influential observations. *Hacettepe Journal of Mathematics and Statistics*, 44, 1–1.

Kabak, M., Erbaş, M., Çetinkaya, C., & Özceylan, E. (2018). A GIS-based MCDM approach for the evaluation of bike-share stations. *Journal of Cleaner Production*, 201, 49-60.

Kock, N., & Lynn, G. (2012). Lateral collinearity and misleading results in variance-based SEM: An illustration and recommendations. *Journal of The Association for Information Systems*, 13, 546-580.

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, 299-317.

Lin, J.-R., & Yang, T.-H. (2011). Strategic design of public bicycle sharing systems with service level constraints. *Transportation Research Part E: Logistics and Transportation Review,* 47, 284–294.

Liu, J., Sun, L., Li, Q., Ming, J., Liu, Y., & Xiong, H. (2017). Functional zone based hierarchical demand prediction for bike system expansion. *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 957-966. https://doi.org/10.1145/3097983.3098180

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.

MacCallum, R. C., Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods*, 4, 84-99.

MacCallum, R. C., Widaman, K. F., Preacher, K. J., & Hong, S. (2001). Sample size in factor analysis: The role of model error. *Multivariate Behavioral Research*, 36, 611-637.

Noland, R. B., Smart, M. J., & Guo, Z. (2016). Bikeshare trip generation in New York City. *Transportation Research Part A: Policy and Practice*, 94, 164-181.

Ogilvie, D., Foster, C., Rothnie, H., Cavill, N., Hamilton, V., Fitzsimons, C. F., & Mutrie, N. (2007). Interventions to promote walking: Systematic review. *British Medical Journal*, 334, 1204.

Osborne, J. (2002). Notes on the use of data transformations. *Practical Assessment, Research and Evaluation*, 8, 6. https://doi.org/10.7275/4vng-5608

Osborne, J. W., & Waters, Elaine. (2002). Four assumptions of multiple regression that researchers should always test. *Practical Assessment, Research & Evaluation*. 8, 2. https://doi.org/10.7275/r222-hv23

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.

Prabhakaran, S. (2016). Outlier treatment with R | Multivariate outliers. [online] R-statistics.co. Available at: http://r-statistics.co/Outlier-Treatment-With-R.html/ [Accessed 21 July. 2020].

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, 451–475.

Rixey, R. A. (2013). Station-level forecasting of bike sharing ridership: Station network effects in three U.S. systems. *Transportation Research Record: Journal of the Transportation Research Board*, 2387, 46-55. https://doi.org/10.3141/2387-06

Rummel, R.J. (1970). Applied factor analysis. *Evanston, IL: Northwestern University Press*.

Saghapour, T., Moridpour, S., & Thompson, R. G. (2017). Measuring cycling accessibility in metropolitan areas. *International Journal of Sustainable Transportation*, 11, 381–394.

Scott, D. M., & 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.

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.

Shaheen, S., Martin, E., & Cohen, A. (2013). Public bikesharing and modal shift behaviour: A comparative study of early bikesharing systems in North America. *International Journal of Transportation*, 1, 35-54.

Shu, J., Chou, M. C., Liu, Q., Teo, C., & Wang, I. (2013). Models for effective deployment and redistribution of bicycles within public bicycle-sharing systems. *Operations Research*, 61, 1346-1359.

Hamilton Bike Share (2019). SoBi Hamilton Number of Hubs. [online] Available at: https://hamilton.socialbicycles.com/ [Accessed 21 July. 2020].

Song, Y., Preston, J., & Ogilvie, D. (2017). New walking and cycling infrastructure and modal shift in the UK: A quasi-experimental panel study. *Transportation Research Part A: Policy and Practice*, 95, 320-333.

Statistics Canada. (2017). Census profile, 2016 Census. [online] Available at: https://www12.statcan.gc.ca/ [Accessed 21 July. 2020].

Stieve, T. (2012). Moran's I and spatial regression. [online] Available at: https://drive.uqu.edu.sa/_/aeelfarouk/files/Morans-I-and-Spatial-Regression.pdf/ [Accessed 21 July. 2020].

Tran, T. D., Ovtracht, N., & d'Arcier, B. F. (2015). Modeling bike sharing system using built environment factors. *Procedia CIRP*, 30, 293–298. https://doi.org/10.1016/j.procir.2015.02.156

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*.

Wang, X., Lindsey, G., Schoner, J. E., & Harrison, A. (2016). Modeling bike share station activity: Effects of nearby businesses and jobs on trips to and from stations. *Journal of Urban Planning and Development*, 142, 04015001.

Winters, M., Hosford, K., & Javaheri, S. (2019). Who are the 'super-users' of public bike share? An analysis of public bike share members in Vancouver, BC. *Preventive Medicine Reports*, 15, 100946.

Wu, S., & Lei, X. (2019). The analysis of the influencing factors on the problems of bike-sharing system in China. *Institute of Electrical and Electronics Engineers Access*, 7, 104000-104010.

Zhang, Y., Thomas, T., Brussel, M., & van Maarseveen, M. (2017). Exploring the impact of built environment factors on the use of public bikes at bike stations: Case study in Zhongshan, China. *Journal of Transport Geography*, 58, 59–70.

Zhang, Y., Lin, D., & Mi, Z. (2019). Electric fence planning for dockless bike-sharing services. *Journal of Cleaner Production*, 206, 383-393.