

Cycling in Toronto: Route Choice Behavior and Implications to Infrastructure Planning

by

Siyuan Li

A thesis
presented to the University of Waterloo
in fulfilment of the
thesis requirement for the degree of
Master of Applied Science
in
Civil Engineering

Waterloo, Ontario, Canada, 2017

© Siyuan Li 2017

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

This research investigates the route choice behavior of cyclists in the City of Toronto using data collected from a smartphone application deployed to a large number of cyclists in the City. A total 4,556 cyclists registered for this study and logged over 30,000 commuting trips and 9,600 recreational trips over a study period of 9 months. The routes of individual cycling trips were estimated by a map-matching algorithm using second-by-second GPS readings of each trip and Toronto's cycling road network. Personal information such as age, gender and residence, work or school place was collected from the participants on a voluntary basis. The collected cycling trip data were used to estimate path-size logit route choice models – variant of multinomial logit model for both commuting and recreational trips with various modeling options and combinations of candidate factors. The estimations of the models were evaluated using various performance measures and statistical tests, resulting in findings and conclusions on the optimal modeling structure, the factors that had statistical significant effects on cyclists' routing decisions and the magnitude of these effects.

The modeling results revealed the critical importance of cycling facilities such as bicycle lanes, multiuse pathways and trails on cyclists' route choice decisions. It was shown that directness as measured by travel distance is the most important factor considered by commuting cyclists in making their route choices. It was also found that cyclists prefer cycling along major streets than local streets and do not mind traveling along transit routes. Furthermore, they tend to choose routes with more bicycle facilities especially dedicated off-street facilities. Comparing to recreational trips, the routes chosen for commuting were in general closer to the routes of minimum distance and energy consumption. In contrast, for recreational trips, cyclists were less concerned about the directness or the degree of challenges of the routes. For these trips, cyclists appeared to place safety at a higher priority instead of time as they showed a higher preference to dedicated bike facilities such as bike lanes and off-street bike paths than on-street mixed facility. Weather and personal attributes were not found to be statistical significant in affecting cyclists' route choices. These along with other findings from this thesis research have provided valuable information for Toronto's ongoing effort on bicycle network planning. The results could also be used to enhance route-finding tools available to cyclists for improved cycling experience.

Acknowledgements

I would like to thank Dr. Liping Fu for his supervision of my master's degree. His guidance and support greatly contribute to this research, and his deep knowledge and attitude towards science influence me in many aspects.

This research was supported through a project funded by Ontario Centers of Excellence (OCE), Brisk Synergies, and the Natural Sciences and Engineering Research Council of Canada (NSERC). I would like to acknowledge Brisk Synergies for their support in data collection.

I greatly acknowledges Matthew Muresan for his support and advice in this research and in TRB paper. He also provided me great advice on both research and presentations. Finally, I would like to thank other iTSS group fellows for supporting and encouraging me during my study period.

Table of Contents

Author's Declaration.....	ii
Abstract	iii
Acknowledgements.....	iv
Table of Contents.....	iv
List of Figures	vii
List of Tables	viii
Chapter 1 Introduction	1
1.1 Introduction.....	1
1.2 Research Problems.....	2
1.3 Objectives and Scope of Work	6
1.4 Structure of the Thesis	6
Chapter 2 Literature Review	8
2.1 Data Collection	8
2.1.1 Stated Preference (SP) Surveys and Revealed Preference (RP) Surveys	8
2.1.2 Route Tracking Data with Global Positioning System (GPS)	10
2.2 Alternative Routes Generation.....	13
2.2.1 Shortest Path Algorithms	13
2.2.2 Route Choice Set Generation Methods.....	17
2.3 Route Choice Models.....	19
2.3.1 Probit Model	20
2.3.2 Logit Models.....	20
2.4 Factors Affecting Route Choice.....	24
2.4.1 Alternative Route Dependent Factors	24
2.4.2 Generic Trip Attributes	28
2.5 Summary of Previous Studies	30
Chapter 3 Data Description.....	33
3.1 Study Area	33
3.2 Data Sources	34
Chapter 4 Data Processing and Exploratory Data Analysis.....	39

4.1 Data Pre-processing	39
4.2 Map Matching	41
4.3 Exploratory Data Analysis	44
Chapter 5 Development of Route Choice Models	55
5.1 Discrete Choice Model - Framework.....	55
5.2 Attributes Affecting Route Choice Behavior.....	56
5.3 Model Estimation.....	61
5.3.1 Model Specification	61
5.3.2 Generation of Alternative Routes	63
5.3.3 Independence from Irrelevant Alternatives (IIA)	64
5.3.4 Path-Size Logit Models.....	64
5.4 Results and Discussions of Route Choice Models.....	66
5.4.1 Alternative Route Choice Set.....	66
5.4.2 Modeling Results	71
5.4.3 Evaluation of Competing Model Structures	75
5.4.4 Effects of Related Factors.....	77
Chapter 6 Conclusions and Future Research	88
6.1 Summary of Work and Findings.....	88
6.2 Contributions and Future Research.....	90
References	92
Appendix.....	99
I Tools and Functions Used in Data Pre-Processing and Map-matching	99
II Statistic Summary for Commute Trips and Recreational Trips	101
III Map-matching Arbitrary Observation (PL/pgSQL Script)	102
IV Trip Generation (PL/pgSQL Script)	104
V Dijkstra's Shortest Path Generation with Binary Heap (C# Codes)	108
VI Tables of Correlations among Candidate Factors for Commute and Recreational Trips	114

List of Figures

Figure 3-1 Cycling Network in the Study Area	33
Figure 3-2 An Example of GPS Observations	34
Figure 3-3 An Example of User Profile Data	35
Figure 3-4 Map of Parallel Parking Area and Bus Stops.....	36
Figure 4-1 Pre-Processing of Network	39
Figure 4-2 Map-matching Issues in Intersections and Parallel Streets	42
Figure 4-3 Tool for Map-matching Processing in ArcGIS	43
Figure 4-4 Buffers for Links (left), Intersections (middle) and Parallel Streets (right).....	44
Figure 4-5 An Example of Routes with Gaps	44
Figure 4-6 Distribution of Trips by Months	45
Figure 4-7 Distribution of Trips by Purposes	46
Figure 4-8 The Number of Commuting Trips VS. Recreational Trips.....	46
Figure 4-9 Distribution of Trips by Distance.....	47
Figure 4-10 Distribution of Gender, Income, Age, Winter Cycling, Rider History and Comfort Level in Personal Characteristics.....	49
Figure 4-11 Distribution of Types of Edges	51
Figure 4-12 Distribution of Different Bicycle Facility Types	54
Figure 4-13 Distribution of Bicycle Facility Types Except No Facility.....	54
Figure 5-1 An Example of Observed Route and Alternatives	68

List of Tables

Table 2-1 Summary of Shortest Path Algorithms	16
Table 2-2 Summary of Route Choice Set Generation Methods	18
Table 3-1 Summary of Data Sources	37
Table 4-1 Description of Road Types	50
Table 4-2 Description of Bicycle Facility Types	52
Table 5-1 Attributes of Route Choice Models	59
Table 5-2 Detailed Information of the Sample Trip	68
Table 5-3 Overlapping Percentages for Alternative Routes and Observed Routes for Commuting and Recreational Trips	71
Table 5-4 Commuting Trip Route Choice Model Estimations for Different Scenarios	72
Table 5-5 Recreational Trips Route Choice Model Estimations for Different Scenarios	73
Table 5-6 Model Estimations with Different Market Segmentations by Trip Length.....	74
Table 5-7 Marginal Rates of Substitution with Respect to the Natural Log of Distance	76
Table 5-8 Model Estimations with Different Market Segmentations by Age	80
Table 5-9 Model Estimations with Different Market Segmentations by Gender	81
Table 5-10 Model Estimations with Different Market Segmentations by Rider History	82
Table 5-11 Model Estimations with Different Market Segmentations by Weather	83
Table 5-12 Model Estimations with Different Market Segmentations by Month	84

Chapter 1 Introduction

1.1 Introduction

Increased travel demand due to population growth creates substantial challenges in many regions worldwide. It leads to an increase in road traffic, which in turn causes more congestion, collisions, and noise and environment pollution. The Greater Toronto Area (GTA) in Ontario, Canada is one of Ontario's fastest growing regions with its projected population for 2041 to reach over 40% increase of its present population (Ontario Ministry of Finance, 2016). Previous studies have estimated that congestion in Toronto already costs commuters \$3.3 billion in lost time and additional vehicle maintenance, and the wider economy an additional \$2.7 billion (Delcan and Lura Consulting, 2013). In addition to lost time, automobile travel increases both noise and air pollution (OECD, 2014). Transportation-related emissions account for nearly 24% of total GHG emissions, of which 80% are caused by road traffic specifically. Over the past decade, these emissions have grown by 12%, and are a key contributor to the formation of smog and negative health effects (Environment Canada, 2013).

Regional and municipal governments have therefore been taking initiatives to encourage the use of alternative transportation modes. Active modes of transportation, such as walking and cycling, have been promoted as sustainable alternatives to travel by automobiles as their use has been shown to reduce congestion, air and noise pollution, parking needs, economy costs (e.g. infrastructure costs and fossil fuels), safety costs, and energy consumption (Nelson & Allen, 1997; Usyukov, 2013; Vandenbulcke et al., 2009). In addition to positive effects on the transportation system, cycling and walking have benefits on personal health, the city environment and social well-being (Nelson & Allen, 1997). Cycling is also frequently faster than other modes including automobiles for short commuting trips (Usyukov, 2013). Therefore, cycling has become a common mode choice for trips in many countries and regions around the world. For example, in Pelotas, a medium-sized city in Southern Brazil, 17% of workers use bicycles to get to work (Bacchieri et al., 2010). Bicycles are the most frequently used traveling mode among men accounting to 27% of all transportation (Bacchieri et al., 2005). However, cycling in Canada is not as popular as in European or American countries. Past studies have found that only 2% of work trips in Canada were completed by cycling as compared with 28% in Netherlands (Pucher & Dijkstra, 2003). Instead of for commuting, cycling is often used for recreational and tourist trips (Chen & Chen, 2010). A recent nationwide survey in the United States found that 27% of people aged over 16 had cycled in the past 30 days, with recreation being the most common purpose.

In recent decades, many municipal and provincial governments in Canada have recognized the importance of cycling and are taking steps to promote cycling. For example, the Region of Waterloo's 2004 Master Plan proposes the construction of over 700 km cycling infrastructure and an investment of \$33 million in the next 20 years aiming at increasing bicycle travel by 2% (Region of Waterloo, 2004). Programs like this will largely increase the number of cycling trips especially in urban areas. For example, recent studies done by the City of Hamilton, Ontario shows a 30% growth in cycling as a commute mode between 1996 and 2000, which has outpaced the population growth (Yiannakoulis et al., 2012). Similarly, Toronto, the fifth largest city in

North America, with a population of over 2.6 million, has seen a substantial growth in cycling over the past five years. The number of cycling trips made by residents increased by 33% from 2001 to 2006 (City of Toronto, 2016).

As a result of increasing cycling demand and recognition of the importance of cycling in promoting active transportation, many municipalities have started to place a significant emphasis on improving their cycling infrastructure. For example, the City of Toronto is currently developing a 10-year's plan to revitalize its cycling network with a budget of \$73 million to build and upgrade over 1000 km of its cycling infrastructure (City of Toronto, 2010). Many cycling network improvement projects are being proposed. These projects are to be evaluated on eight evaluation criteria, including Connectivity, Coverage, Crossing Barriers, Current Demand, Potential Demand, Trip Generators, Safety, and Population and Employment Density. One of the most important requirements for estimating the benefits of these projects is a quantitative understanding of factors affecting travelers' preferences towards different modes as well as different bicycle facilities. This research is mainly motivated by the need to understand cyclists' route choice behavior and develop models that can be used to quantify the relative effects of various factors on cyclists' routing decisions.

1.2 Research Problems

The problem of determining how and why individual cyclists choose the particular traveling routes for the trips they have made is complex in nature. The complexity arises due to the challenges of collecting all relevant data pertaining to cyclists' route choice decisions, identifying the candidate routes for each trip that the cyclists could have considered before making the trip, and extracting cyclists' preference structure and the influencing factors based on their revealed choices. The following section provides an overview of the existing literature on each of these challenges.

Data Collection

In the literature, quantitative studies aimed at uncovering the factors affecting cyclists' behavior have taken two main approaches in collecting the data needed. The first approach, which is also the most popular one, involves surveying cyclists to uncover the reasons and factors that they consider when choosing routes. The second approach uses GPS-enabled devices to track cyclists' location during the trips in real-time, and has become popular in recent years as technology involved into scientific researches.

The survey-based approach has traditionally been employed by researchers to collect and analyze cyclist behavior. Two types of surveys could be applied: stated preference (SP) surveys and revealed preference (RP) surveys. SP surveys ask respondents to rank the factors that they considered affecting their route choice decisions (Yang & Mesbah, 2013) or ask them to choose from some hypothetical route choices controlled by experimenters (Broach et al., 2012; Stinson & Bhat, 2003; Tilahun et al., 2007). In contrast, RP surveys collect data on the route choices that trip-makers made in the past and the conditions under which these decisions were made (Howard & Burns, 2001; Hyodo et al., 2000). With the knowledge of chosen routes, mathematical models can then be developed to estimate cyclists' route choice behavior (Bacchieri et al., 2010; Stinson & Bhat, 2003). Comparing to RP surveys, SP surveys are relatively easier and less expensive to

perform (Broach et al., 2012; Stinson & Bhat, 2003). However, SP surveys often use simulated alternatives and decision-making scenarios; as a result, the findings from SP are considered less reliable. Despite their simplicity, survey-based methods are time consuming and the collected data are often incomplete as the respondents may forget important information related to their trips.

As GPS-enabled devices such as smartphones have become ubiquitous over the past years, they have provided a new opportunity for collecting detailed route choice data. By using GPS-enabled devices, the locations of cyclists can be collected automatically in an electronic format (latitude and longitude) (Bierlaire & Frejinger, 2008). Demographic information and detailed decision making processes can also be collected from the participants through a specific app (Hood et al., 2013) or through a follow-up web questionnaire (Broach et al., 2012). Collecting route tracing GPS data saves both time and money, and the route data collected are more accurate than those from traditional surveys. However, one of the main challenges of using GPS data is matching GPS points to the street network so that the actual travel routes in the form of sequential links can be identified. The common approach to this problem is snapping each point to the nearest link (Hood et al., 2011; Menghini et al., 2009; White et al., 2000). In dense networks, however, this method could lead to infeasible or non-unique solutions (Quddus et al., 2003; White et al., 2000). The second challenge is the high computational demand when a large number of trips need to be processed.

Alternative Route Set Generation

The second challenge that needs to be addressed for route choice modeling is how to generate the set of possible routes that were considered by the trip makers before selecting a route to travel. In the literature, alternative routes for any given trip were usually generated through applications of a shortest path algorithm (SPA) over a pre-defined network assuming a specific routing criterion such as minimizing total travel distance, travel time or travel risk. In order to generate multiple routes, the shortest path algorithm is applied iteratively and after each iteration the cost attributes of some of the network links are modified to ensure minimum similarity between alternatives. Two approaches are commonly used for the step of modifying network attributes: stochastic and deterministic. In a stochastic approach, the travel cost of each link is assumed to follow a known probability distributions, and its realization is generated before each iteration of the shortest path search (Bliemer & Bovy, 2008; Bierlaire & Frejinger, 2005; Nielsen et al. 2000; Prato & Bekhor, 2007; Ramming, 2001). This method can generate a diverse set of alternative routes; however, it has no control on the level of similarity or reasonableness of the generated routes. A deterministic method, on the other hand, either penalizes or completely removes the links that are already on the routes identified in the previous iteration of the shortest path research to ensure sufficient dissimilarity between the identified alternative routes (Lim & Kim, 2005; Rieser-Schüssler et al., 2013). While this method has the advantage of being able to control the level of similarity within the route set, it may result in routes that are too unreasonable to be considered by the cyclists. Lastly, it should be noted that little past efforts have examined the reasonableness of these different path generation methods and their implications to the subsequent route choice analyses.

Route Choice Modeling

After the alternative routes for each of the individual trip are generated, the next step is to estimate route choice models so that the key factors that influence decisions can be determined accordingly. These models are usually based on random utility theory, where the preference of each alternative route is defined as a utility function (Bliemer & Bovy, 2008; Hood & Charlton, 2013; Hopkinson & Wardman, 1996; Stinson & Bhat, 2003; Menghini et al., 2009).

The most commonly employed model is the logit model, which assumes that the desirability of choosing a candidate route is controlled by a utility function consisting a deterministic part and a random error. The random errors estimating the unmeasurable factors are assumed identically and independently distributed (IID) and following the Gumbel Type I distribution (Ben-Akiva & Lerman, 1985). The key consequence of the IID requirement is that the choices modelled with a logit model must be perceived as significant and relevantly different from each other to the decision maker. One of the characteristics or limitations of the logit model is typically called the Independence of Irrelevant Alternatives (IIA) limitation. The alternative choice set generated must abide by this property, otherwise counter-intuitive results may appear. For modeling route choices, overlapping among alternative routes violates this property. In the literature, the C-logit model and Path-size logit model are frequently used to address this problem by employing a commonality factor or a path-size factor in the utility function to estimate how much overlapping exist among alternatives (Bliemer & Bovy, 2008; Cascetta et al., 1996; Hood & Charlton, 2013; Menghini et al., 2009). Previous studies confirmed that the simple logit model and the C-logit model are easily affected by the size of alternative route set, while the Path-size logit model is more robust with respect to the size of alternative route set (Prato & Bekhor, 2007; Ramming, 2001) outperforming other models according to log-likelihood measures (Prato & Bekhor, 2007; Ramming, 2001).

Route Choice Related Factors

Route choice models, once estimated, provide indications on the factors that affect cyclists' route choice behavior. Previous studies have identified many factors affecting route choice behavior; these factors can generally be classified into two categories: alternative dependent attributes (e.g. trip distance, travel time, network characteristics, etc.) and generic trip attributes (e.g. socio-demographic characteristics of the cyclists, trip purpose and weather conditions). Alternative dependent attributes include factors vary by routes, while generic trip attributes are the same for an entire trip and for all alternative routes in a trip instance, and only change between different trips.

Previous studies have shown that alternative dependent attributes are statistical significant in route choice decisions with trip distance and travel time routinely identified as the two most important factors in cyclists' route choice decision-making (Bovy & Bradley, 1985; Howard & Burns, 2001). These two types of factors have been shown as especially important to commuting cyclists, who have commonly been found to prefer the most direct route. Both trip distance and travel time have been shown to lower the probability of a particular route being selected (Aultman-Hall et al., 1997; Bovy & Bradley, 1985; Howard & Burns, 2001). Recreational cyclists, on the other hand, are generally not as sensitive as cyclist commuters to trip distance and travel time (Broach et al., 2012).

While factors like distance and travel time can be influenced by decision makers, the key factors of importance to researchers and planners are network attributes. Past research has found that cyclists are particularly sensitive to traffic volume, road types, bicycle facility types, speed limit, etc., which are related to their perceived risk. Furthermore, researchers have also shown that they are sensitive to the road grade, number of turns, scenery and on-street facilities (e.g. fountains, parking spots, maintenance equipment, etc.), which are related to their comfort level (Ladis et al., 1997; Parkin et al., 2007). Many studies have shown that high motor traffic volume and speed limit along the route have a negative effect on a route's desirability (Sorton & Walsh, 1994). In addition, road types significantly affect a route's desirability, and cyclists have been shown to prefer low traffic residential streets than busy major roads (Abraham et al., 2002; Aultman-hall & Kaltenecker, 1999; Hood et al., 2001; Parkin et al., 2007). Cyclists also shows obvious preference to roads with bicycle facilities (Sener et al., 2009; Stinson & Bhat, 2003; Tilahun et al., 2007); though, results on the importance of each of these factors varies between studies with no clear consensus. Some findings show a more preference for separated off-street bicycle facilities such as pathways than for on-street facilities such as bike lanes (Abraham et al., 2002 & Parkin et al., 2007), while other studies concluded that considering both safety and speed, bike lanes were the most preferred bicycle facility type followed by the separate pathway (Aultman-hall & Kaltenecker, 1999; Hood et al., 2001; Sener et al., 2009; Stinson & Bhat, 2003; Tilahun et al., 2007). Furthermore, previous studies also show the negative effect of on-street parking, among which parallel on-street parking was the least preferred (Sener et al., 2009; Stinson & Bhat, 2003; Tilahun et al., 2007).

Comfort level related network characteristics (e.g. gradient, number and type of turns, scenery, and on-route facilities) measure the comfort level cyclists may experience when traveling on the route. Number of turns has been shown to negatively affect a route's desirability, as cyclists always want to avoid delays (Hood et al., 2001; Parkin et al., 2007). In terms of the gradient, studies have found conflicting results. Menghini et al. (2009) and Hood et al. (2001) concluded that cyclists generally try to avoid slopes, while Sener et al. (2009) found moderate slope was preferred by cyclists to avoid boredom. Scenery and on-route facilities were found more sensitive to recreational trips than commute trips increasing the possibility of a particular route being chosen (Chen & Chen, 2010).

Finally, in the context of generic trip attributes, only socio-demographic characteristics have been mentioned in previous studies. Previous studies have shown contradicting results on the effects of various socio-demographic characteristics, with some determining that all socio-demographic factors have no significant effect on route choice decisions (Abraham et al., 2002; Hood et al., 2001), while others finding slight effects of some of the socio-demographic factors such as age, gender, household size and income (Antonakos, 1994; Hulse & Shafizadeh, 1996; Krizek, 2006; Sener et al., 2009). The magnitude of the effects of those variables are significantly dependent on the study area and data being used.

In summary, although many studies have been devoted to the subject of cyclist route choice behavior modeling, most of these studies have limited generalizability, and the results of their findings may not be transferrable due to their use of choice data with low resolutions, small sample sizes or incomplete attributes. Many questions pertaining to the route choice behavior of the cyclists in a specific region or city still need to be investigated using local data taking into location specific characteristics related to the cycling population, road network conditions, and

weather. Many questions still remain: what kind of road facilities and cycling environments are preferred by cyclists? What are the main factors that influence cyclists' route choice decisions when they make commuting or recreational trips? What are the substitutive relationship between the factors considered by cyclists?

Furthermore, there is a lack of systematic investigation on various methodological aspects of route choice modeling. For example, what criteria should be considered in generating alternative routes? How many alternatives should be considered? How sensitive are the route choice modeling outcomes to the route alternative generation method and the size of the choice set?

1.3 Objectives and Scope of Work

The goal of this research is to address some of the key questions raised in the previous section pertaining to the cyclist route choice behavior in the context of the City of Toronto. In particular, the research is to make use of a rich dataset including details on the trips made by a large group of cyclist participants to investigate the effects of various factors on cyclists' choice behavior. The specific objectives are as follows:

- To conduct a thorough review on the literature related to modeling and analysis of cyclist route choice behavior and synthesize the major findings and limitations of the past studies;
- To develop a methodology to process cycling GPS data and algorithms to perform map matching for recreation of the link-by-link travel routes of the individual cycling trips;
- To develop and evaluate models and algorithms for generating alternative routes for the individual trips (or origin-destination pairs) made by the participating cyclists;
- To develop route choice models that can be used to determine the effect of various attributes related to the underlying road network, the trip being made and the individual making the trip on route choice decisions;
- To investigate the sensitivity of route choice modeling results to various modeling options such as route generation method, size of candidate route set, and selection of attributes being included.

1.4 Structure of the Thesis

This thesis follows the following structure:

Chapter 1 introduces the background, motivation, research problem, objectives and scope of the work.

Chapter 2 describes previous studies done in cycling route choice behavior analysis and highlights existing gaps of knowledge on cyclists' route choice behavior;

Chapter 3 describes the study area and data used in the studies;

Chapter 4 introduces the data processing (map-matching and trip generation) and exploratory data analysis;

Chapter 5 describes multiple route choice models based on multiple model settings and combinations of factors for both commuting and recreational trips, and estimates the effects of various trip operational, network, socio-demographic and weather properties on cyclists' route choice behavior.

Chapter 6 summarizes the research result findings and discusses the limitations and areas for future research exploration.

Chapter 2 Literature Review

Cyclists' route choice analysis has received considerable attention in the past few decades, and many studies have been done to identify the factors that capture travelers' route choice behavior. Research on route choice behavior, however, faces four main challenges. The first is the ability to collect data on the travel patterns and preferences of users, which is also extremely time consuming (Hood & Charlton, 2013). Previous studies have frequently employed three different data collection methods: revealed preference surveys, stated preference surveys and direct route tracking with global positioning system (GPS) with their own benefits and shortcomings (Li et al., 2005).

The second challenge is particularly related to studies using revealed preference data. In these studies, only the routes that were chosen are known while the true alternative routes that had been considered by the travelers are unknown and can only be speculated. The alternative routes are needed for choice modeling to gain quantitative understanding of cyclists' preference structure related to different choice factors. The method to identify the alternative route set is one of the challenges when analyzing route choice behavior.

The third challenge is to specify the model that could be used to simulate cyclists' route choice behavior. Some researchers simply compared several alternative routes, and obtained empirical results. However, in order to have an improved understanding of cyclists' preferences, quantitative models should be used to estimate the route choices. Discrete choice models were the most commonly used, among which various forms of multinomial logit models were applied in the previous studies.

The last challenge is to identify the factors that may influence cyclists' route choices. Previous studies have shown that cyclists choose their routes based on a combination of multiple factors (Howard & Burns, 2001), which can be divided into three types: trip operational (e.g. distance and travel time), network (e.g. road surface conditions) and socio-demographic factors (e.g. gender, age and household income). This chapter summarizes the existing studies on route choice behavior analysis pertaining to each of the four challenges, and highlights knowledge gaps and problems in the existing literature.

2.1 Data Collection

Route choice analysis has been performed for many years (Bovy & Bradley, 1985). Among all the procedures, collecting data is always the most time-consuming step, and the quality of data affects the result directly. In previous studies, three kinds of data were used in route choice analysis including stated preference surveys, revealed preference surveys and route tracing records obtained from global positioning system (GPS) devices (Li et al., 2005).

2.1.1 Stated Preference (SP) Surveys and Revealed Preference (RP) Surveys

Previously, many studies relied on either stated or revealed preference surveys as the main method to collect user preference data. In these studies, researchers mailed or emailed surveys with a questionnaire covering the reasons for route choice behavior to sample cyclists, and

performed analysis based on their responses. The first SP surveys were designed asking respondents to rank the factors that they considered affecting their route choice decisions (Yang & Mesbah, 2013). Although the sample size of the stated preference survey can be large, the results cannot be used in quantitative models, and the accuracy of the result needs to be considered, for the respondents may subjectively rank the factors without any evidence. Additionally, the factors listed by researchers are usually not comprehensive (Yang & Mesbah, 2013). Therefore, many subsequent stated preference surveys used an adjusted format by asking respondents to choose from some hypothetical route choices controlled by experimenters to obtain quantitatively understanding of the route choice factors (Tilahun et al., 2007).

The advantages of stated preference surveys are obvious. They are easy to perform, and can be designed avoiding co-linear relationships among options because alternatives are controlled by researchers (Broach et al., 2012; Stinson & Bhat, 2003). Moreover, complex data processing steps such as map-matching and trip generation are not necessary for stated preference data, and model specifications is generally easier due to the clean data and limited alternatives (Broach et al., 2012). Despite these benefits, stated preference surveys can provide some impractical options to individuals, which cannot be done by approaches based on actual route records (Hopkinson & Wardman, 1996).

In stated preference surveys, hypothetical choices with several options are the key components (Abraham et al., 2002). Many previous studies tried to develop the hypothetical choices in a quantitative way to quantify the factors listed in the options. Hopkinson and Wardman (1996), Krizek (2006) and Tilahun et al. (2007) all developed quantitative stated preference surveys by including combinations of travel time and various other factors (e.g. bicycle facilities, number of red lights, etc.) in alternatives. The travel time offsets between options were considered as the relative preference of the other variable in the option. For example, Tilahun et al. (2007) asked respondents to choose from two options, one with bike lanes but with a travel time of 40 minutes, and the other one without bicycle facility but with a travel time of only 20 minutes. The 20-minute offset was considered as cyclists' preference to the bike lane. Another study by Abraham et al. (2002) represents all the bicycle facilities and road types by travel time durations, so that the preference of bicycle facilities and road types could be estimated in mathematical models (e.g. logit models).

The limitation of stated preference surveys is that both alternatives and choices in stated preference surveys are not based on actual routes. Some unreasonable choices may conflict with the actual choices, which are not sound enough for quantitative route choice models. In contrast, revealed preference survey address these limitations by asking respondents to describe the route they took recently from an origin to the destination. Some of them asked respondents to draw the route on the map (Aultman-hall & Kaltenecker, 1999).

Hyodo et al. (2000) collected 754 map-based routing records in Kurume and Utsunomiya, Japan. In this study, the researchers defined a concept called "cognitive length", which is the perception of a trip's length to a user. The cognitive length is calculated as a function of various network and individual factors (e.g. bicycle facilities, lane number, lane width, ect.), and it is assumed that users always choose the path with the minimum cognitive length. With the knowledge of the routes chosen by cyclists, they estimated the coefficients of factors in cognitive length that can maximize the overlapping between the chosen routes and the shortest cognitive length paths.

Many recent studies have taken advantage of geospatial information system (GIS), which have become commonly used in transportation analysis. For example, Aultman-Hall et al. (1997) and Howard and Burns (2001) asked respondents to draw their routes in GIS software (e.g. ArcGIS) based on a digitalized network. GIS provides an easy way to store, process, and analyze routes. A geographic information system (GIS) also provides functions for shortest path algorithm and routes comparison (Aultman-Hall et al., 1997; Howard & Burns, 2001).

The results of revealed preference surveys are based on the actual routes avoiding subjective answers in stated preference surveys, but collecting revealed preference data by surveys is extremely time consuming especially drawing routes, so sample sizes are limited. Furthermore, as revealed preference surveys collect data after the trip has been made, it is impossible for cyclists to recall all the details correctly, and some data may be missing unconsciously (Bacchieri et al., 2010). Some respondents were also found to make mistakes in map interpretations and missing directions when drawing routes on the map (Aultman-Hall et al., 1997). Therefore, the direction-related factors such as gradient and turn type cannot be accurately estimated (Aultman-Hall et al., 1997). Moreover, even if all the details were correctly recalled, the way to define alternative routes is still a big challenge, for the alternatives should follow cyclists' perceptions (Stinson & Bhat, 2003; Tilahun et al., 2007).

Yang and Mesbah (2013) combined the stated and revealed preference surveys together to analyze the effects of various factors including distance, travel time, safety, gradient, security facilities, and scenery on route choice behavior. The revealed preference survey asked cyclists to recall their actual routes and personal information, while in the stated preference survey, the respondents were asked to compare their selections with the pre-defined options, and ranked the attributions based on their perceptions. The analysis about the consistency of actual routes and their reported ranking showed similar results for factors such as distance and travel time. However, they found the safety and gradient were ranked as a high priority, but seemed less substantial in chosen routes. The researchers concluded this was because route choice decision is a result of multiple factors, and in the stated preference survey the interactions among factors were not considered. Their research highlighted the limitation of stated preference surveys.

Both stated and revealed preference surveys are time consuming due to the long waiting time for responses. Although surveys were conducted through the Internet in recent decade to increase the level of accessibility (Krizek, 2006; Sener et al., 2009), too much information in surveys caused overwhelming for respondents, which largely affected the quality of the answers (Sener et al., 2009). In addition, web-based surveys limit the scope of the studies into a coverage of cyclists having accessibility to the Internet (Sener et al., 2009). Bias still exists in sample selection.

2.1.2 Route Tracking Data with Global Positioning System (GPS)

Until recently, data for route choice analysis mostly come from revealed and stated preference surveys, because these methods do not require specialized devices. However, surveys are time consuming, and the sample size is often limited due to high costs. It is also challenging to contact interviewees afterwards. Researchers have therefore been pursuing the use of newer technologies to support data collection, such as GPS-enabled smartphones. These devices have become popular in recent decades, and many studies analyzing transportation route choice behavior have now been conducted based on GPS data instead of survey results (Bierlaire & Frejinger, 2008;

Broach et al., 2012; Hood et al., 2013; Menghini et al., 2009). By using GPS-enabled devices, the locations of cyclists can be collected automatically in the electronic format (Bierlaire & Frejinger, 2008). Demographic information and detailed decision making processes can be collected by surveys performed during route data collection in the application (Hood et al., 2013) or through a follow-up web questionnaire (Broach et al., 2012).

For example, Bierlaire and Frejinger (2008) collected GPS data to estimate the relative preferences of various types of roads in Switzerland by Path-size logit model (PSL), and Menghini et al. (2009), in addition, collected 11,000 GPS trip observations from over 2,000 cyclists in Zurich to identify the factors that can influence cyclists' route choice behavior.

GPS-enabled devices have revolutionized route choice analysis at previously un-imaginable scales by enabling cyclists to participate in easily. For example, Broach et al. (2012) performed their study in a large network in Portland, Oregon, USA. They collected 1,449 utilitarian GPS trip records and demographic information for 164 cyclists to estimate the attractiveness of bicycle facilities, traffic controls and route topology conditions (e.g. turn frequency) to cyclists.

Similarly, Hood et al. (2013) additionally collected 7,096 GPS trip observations from 952 volunteer cyclists in the City of San Francisco through a free smartphone application, CycleTracks, which can collect per-second GPS readings as well as trip purpose and individual properties at the same time. Although the scope of the data was limited to cyclists having smartphones, this drawback is much outweighed by the advantages of the easy data collection and accurate route records.

Collecting route tracing GPS data saves both time and money, and solves the problem of limited sample size. It also decreases the possibility of regional bias in sample selection. It is widely used for route choice analysis in transportation studies, including studies on both cyclists and drivers. For example, another recent study by Jan et al. (2000) collected more than 3,000 GPS trip records from 216 drivers for analyzing driving path choices in Lexington. They also calculated travel time and speed based on the GPS data. They found that GPS data along with the use of GIS platforms decreased the data collection times, and increased the quality of data. Some quantitative values such as travel time and speed were also easily calculated from GPS data, which could not be done by survey results.

As generic GPS records are collected automatically, trips with other traveling modes (except for cycling) could be collected as well due to the mis-usage of devices. One of the challenges of using GPS observations is separating trips by different traveling modes. Trips made by other modes (except for cycling) can be detected by the speed or acceleration. Hood et al. (2011) monitored speeds and a 95th percentile acceleration by fuzzy logit models to detect the mode changes.

Map-matching methods

Another big challenge of using GPS observations is snapping GPS point data to the street network. Previous studies proposed many map-matching methods. The basic solution of map-matching is snapping the nearest edge (Euclidean distance) to the observation (Menghini et al., 2009; White et al., 2000). For example, Hood et al. (2011) adopted this idea, and matched the

observations directly to the link with minimum Euclidean distance. Then, each trip was made up of a sequence of links.

White et al. (2000) additionally did an analysis on various map-matching methods, and found that using nearest edges correctly map-matched nearly 73% of the GPS observation in four routes. Although this method performed well in their studies, it is sensitive with the number of turns in the trip, and may cause errors when observations are closer to the wrong links, which often happened around intersections due to the low accuracy of GPS devices (White et al., 2000).

Some other algorithms improved the map-matching methods by creating a candidate link set based on the topology connectivity (Quddus et al., 2003; White et al., 2000). For example, Bierlaire and Frejinger (2008) connected GPS observations with network links by a term called Domain of Data Relevance. The GPS observations would be matched with a set of links, to whom the distance was within a pre-defined threshold. The possibility of each link was equally divided.

Marchal et al. (2004) developed an advanced map-matching method, which was adopted by Menghini, et al. (2009). They stored the N nearest edges (Euclidean distance) in the candidate route set with their distance as the score. The topology connectivity was considered in updating the link sequence of the trip, and the scores of the candidate route were calculated by the sum of all Euclidean distances between observations and network links, as shown in Equation 2-1.

$$\sum_{j=1}^p \sum_{i=1}^T d(Q_i, E_j) \delta_{ij}, \quad (2-1)$$

where $d(Q_i, E_j)$ is the distance between observation i (Q_i) and edge j (E_j), and $\delta_{ij} = 1$ if observation i is matched to edge j . The map-matching result of a trip was determined by the candidate route with the highest score. They concluded that the accuracy of this method (around 96%) was high with $N = 30$. However, due to the inaccuracies of GPS observations, some observations cannot be map-matched with any link connected to the previous link; these results were then considered as two trips by the researchers. However, instead of simply considering them as two trips, the gaps in the observations caused by signal lose or inaccuracies in GPS should be connected by advanced methods. Moreover, the accuracy of the map-matching largely depends on the value of N . A large N leads to higher accuracy, but lower speed.

Many previous studies have also considered map-matching as a statistical problem instead of a searching problem. Pyo et al. (2001) proposed multiple hypothesis technique (MHT) method, which keeps considering multiple high probability hypotheses (routes or paths) using Bayesian methods and likelihood probability. The final results were selected by Kalman filtering to avoid bias. Quddus et al. (2003), additionally, adopted Greenfeld's weighting scheme (2002) considering orientation, proximity and intersecting angles to select the best option in route or path candidates.

These statistical map-matching methods in previous studies were more accurate than just using the Euclidean distance, but they require large computational power for large network, and the complexity in implementation is high.

Various GIS-enabled platform such as ArcGIS, QGIS or other self-programmed scripts can be used to perform the map-matching (Broach et al., 2012). Instead of calculating the Euclidean distance by mathematical functions, Strauss, et al. (2015) used buffers to calculate distance in geospatial software (ArcGIS) to increase the efficiency. They created a buffer for each link by a pre-defined distance, and considered observations within the buffer was with a distance lower than the pre-defined maximum threshold.

Despite improvements has been applied to map-matching methods, errors still exist. Even the best map-matching algorithm still contained around 6% errors (Bierlaire & Frejinger, 2008). Nielsen and Jørgensen (2004) also observed that 90% of trips in their study had missing links in the trips.

2.2 Alternative Routes Generation

In the studies with revealed preference data, the second challenge is the definition of alternative routes. Analysis of choice behavior requires knowledge not only the chosen option but also the non-chosen options to gain quantitative understanding of preferences. Route generation is a topic related in transportation engineering, computer science and management system that spans many fields. This section describes a number of shortest path algorithms such as the Dijkstra's algorithm, the Bellman-Ford algorithm, the A* algorithm, etc., among which Dijkstra's shortest path algorithm (SPA) was the most commonly used core algorithm to generate the shortest path due to its benefits in implementation and computational efficiency (Skiena, 1990).

In order to meet the amount requirement of the alternative routes, alternative route set generation methods are based on the iterations of shortest path algorithms (SPA). This section also introduces two kinds of alternative route set generation methods: stochastic route set generation and deterministic route set generation.

2.2.1 Shortest Path Algorithms

The shortest path algorithm (SPA) is a significant topic in both transportation engineering and computer science. The objective of SPA is finding a path that can minimize or maximize the impedance cost from an origin to a destination, while the impedance cost can be a function of multiple criteria or single criterion. Many shortest path algorithms have been developed to perform this task in a quick and accurate manner (e.g. the Dijkstra's Algorithm, the Bellman-Ford algorithm, the A* algorithm, etc.), which are highlighted in this section.

Dijkstra's Algorithm

Dijkstra's algorithm includes an iterative process going through all the nodes and updating the labels to find the shortest path (Dijkstra, 1959). During the process, the algorithm keeps a record of two labels for each node i : $L(i)$ - path length label, which stores the generalized costs of the current path from node i back to the origin node r ; $P(i)$ - the predecessor index, which stores the preceding node to reach node i through the best path (Mehlhorn & Sanders, 2008). The nodes are maintained in a list called scan eligible node set (Q), tracking the candidate nodes that are going to be examined (Mehlhorn & Sanders, 2008).

Dijkstra's algorithm can be implemented in four steps, as shown below. Initially, each node's path length label (L) is assigned an infinite value, and the origin node is assigned a value of 0 as the start point. Iterations are then made from the origin node, and expand to the connected nodes. The cost of the examined node equals the sum of the previous label value and the generalized cost of the link between. If the new cost is smaller than the label of the node, the path length label (L) and P index are updated and the node is inserted into the scan eligible node set (Q). The process is performed for all the nodes adjacent to the examined node, and continued to the next node in Q with the lowest label value. The entire process continues until Q is empty or the destination node is reached (selected for next iteration).

- Initialisation: Set $L(r) = 0$; all the other $L = +\infty$; add the origin node r into the scan eligible node set: $Q = \{r\}$
- Node Selection: Select and remove the node i with the lowest label from Q
- Node Expansion: Scan each link emanating from node i . For each link $(i, j) \in A$
- If $L(i) + C_{ij} < L(j)$, then $L(j) = C_{ij} + L(i)$; $P(j) = i$; and add node j to Q
- Stopping Rule: If Q is empty, then stop; otherwise, go to Step 2 (Node Selection).

Many implementations of shortest path generation algorithms adopt this method, as it is easy to implement (Aultman-Hall et al., 1997; Usyskov, 2013). Additionally, many GIS programs (such as ArcGIS) contain tools that can perform the algorithm.

Many additional improvements were made to further increase the efficiency of Dijkstra's algorithm such as label-setting algorithm (LS), which includes LS with priority queues (e.g. binary heap) to store and scan the eligible node set with labels (Atkinson et al., 1986; Brodal, 1996; Fredman & Tarjan, 1987; Fu et al., 2006). Binary heap, the easiest implementation of a priority queue, is a heap-order tree following the form of a binary tree. Nodes in binary tree always have a parent node and two (left and right) children. In minimum binary heap, the value of the parent node is always smaller than the values of children nodes, and all left trees and right trees are also in the form of a binary tree. The first level node in the binary heap is the one with the smallest value, which is the next node going to be processed in Dijkstra's algorithm (Node Selection). The complexity of finding the node with the minimum value is $O(1)$ compared to $O(|n|^2)$ for a simple Dijkstra's algorithm, and the complexities of deleting a node and inserting a new node are both $O(\log n)$ (Atkinson et al., 1986).

Bellman-Ford Algorithm

Similarly with Dijkstra's algorithm, the Bellman-Ford algorithm goes through all nodes and updates the labels in iterations when smaller or better values are found (Bellman, 1958). The difference between Dijkstra's and Bellman-Ford's algorithm is that in Dijkstra's algorithm, only unprocessed nodes are being examined in Node Expansion Step, but the Bellman-Ford algorithm simply processes all nodes ($V-1$) (Bellman, 1958). This algorithm is more versatile and can solve single-source multi-target problems. In addition, the Bellman-Ford algorithm can process negative edge costs, which is not allowed in Dijkstra's algorithm (Skiena, 1990). However, the

Bellman-Ford algorithm has a higher time complexity ($O(VE)$), where V and E are the number of vertices and arcs. As a real network always contains a large number of nodes and links, the complexity of Bellman-Ford algorithm is often prohibitively high resulting in it being not often applied in large networks.

A Algorithm*

The A* algorithm is a pre-processed speed-up heuristic shortest path algorithm, which presents a more efficient way for path finding (Fu et al., 2006; Fuchs, 2010). The A* algorithm is a generalization of Dijkstra's algorithm that includes a potential function to constrain the shortest path search within a certain area, so that not all nodes between the origin and destination would be examined (Fu et al., 2006). Therefore, the goal of this algorithm is minimizing $f(n)$ defined as (Fuchs, 2010):

$$f(n) = g(n) + h(n) \quad (2-2)$$

where $g(n)$ is the actual distance from the origin node to node n , and $h(n)$ is the potential function representing the estimated distance from node n to the destination node. Small value of $h(n)$ leads to more nodes to be examined. If $h(n) = 0$, then the A* algorithm is the same as Dijkstra's algorithm.

A reasonable $h(n)$ can significantly increase the efficiency when compared to Dijkstra's algorithm. Zeng and Church (2009) compared A* algorithm with Dijkstra's algorithm in real network in Los Angeles and Santa Barbara. They defined potential function by Euclidean distance, and nearly half of the nodes were skipped during the examination. They also found that in large network, A* performed better than other algorithms with lower computational power and shorter processing time.

The actual performance of the A* algorithm in transportation analysis depends primarily on the quality of the potential function (Fu et al., 2006). The way to define the potential function is one of the biggest challenges. Both Goldberg and Harrelson (2005) and Rieser-Schüssler et al. (2013) pre-defined some topology nodes as landmarks, and used the distance or costs from the closest landmark node to the destination node as the potential function. The distances or costs from other nodes to each landmark are calculated during pre-processing stage. Both of them have found that A* landmarks routing algorithm had increased the efficiency (from 79% to 94%) when compared to Dijkstra's algorithm. They also found that increasing the number of landmarks (from 8 to 16) also improved the efficiency (from 10% to 40%). However, the selection of landmarks was a big challenge. In addition, the pre-processing of distances from landmarks to other nodes required large disk storage, especially for large-scaled city network. On very large networks, these storage requirements may be prohibitively large.

Shortest path finding is a combination topic related with management science, geography, transportation engineering and computer science. There are various shortest path algorithms existed in previous studies with their own benefits and limitations. Table 2-1 shows a brief summary of some commonly used shortest path algorithms in transportation studies.

Table 2-1 Summary of Shortest Path Algorithms

Shortest Path Algorithm	Description	Time Complexity	Reference Example
Bellman-Ford Algorithm	Update labels of all connected nodes based on the sum of costs from the origin to the node.	$O(VE)$	Bellman (1958)
A* Algorithm	Use a potential function to limit the number of processing nodes in Dijkstra's algorithm.	Worst case: $O(E)$	Zeng & Church (2008)
A* Landmarks	Include A* algorithm, landmarks and triangle inequality.	Worst case: $O(E)$	Goldberg and Harrelson (2005); Rieser-Schüssler et al. (2010)
Simple Dijkstra's Algorithm	Update labels of unprocessed nodes based on the sum of costs from the origin to the node.	$O(V ^2)$	Skiena (1990); Aultman-Hall et al. (1997)
Dijkstra's with Priority Queue (e.g. Binary Heap)	Dijkstra's algorithm but storing the unprocessed nodes into a priority queue such as Binary Heap.	$O(E \log V)$	Atkinson, et al. (1986)

*V and E are the number of vertices and arcs

In addition to the algorithms, the efficiency of shortest path generation depends on the database. Multiple read or write operations through databases or computer files would decrease the efficiency. Su et al. (2010) suggested storing the nodes and links in an indexed table in temporary memory following the topological structure to increase the efficiency of finding and updating nodes. This method is particularly suitable for developing web or desktop applications, as searching into the database always costs more time than searching the memory.

Another problem for SPAs is that turn penalties cannot be considered when generating routes (Lim & Kim, 2005). In order to overcome the limitation, Lim and Kim (2005) modified Dijkstra's algorithm to a link-based shortest path algorithm. In their algorithm, instead of updating the labels of nodes, the labels of links were updated. The new cost of the link was calculated by:

$$LEC(i, j) + TP[\text{link}(i, j), \text{link}(j, k)] + LC(j, k), \quad (2-3)$$

where $LEC(i, j)$ is the minimum path cost from origin to node j traveled from node i ; $LC(j, k)$ is the link cost from node j to node k ; $TP[\text{link}(i, j), \text{link}(j, k)]$ is the turn penalty from link (i, j) to link (j, k) . If the new cost of the link is smaller than the label of the link, $LEC(j, k)$, the label of the link between node j and node k would be updated. This link-based shortest path algorithm solves the limitation of traditional node-based shortest path algorithm; however, as the number of links is much larger than the number of nodes in a real network, the traditional node-based algorithm outperforms the link-based algorithm in efficiency.

2.2.2 Route Choice Set Generation Methods

As shortest path algorithms only produce one path each time, the route choice set is generated by iteratively applying the shortest path algorithm. Stochastic and deterministic route choice set generation methods are two main methods used for generating a large set of alternative routes in previous studies. Stochastic choice set generation pre-defines the link costs by a probability distribution, such as a normal distribution (Bliemer & Bovy, 2008; Ramming, 2001) or truncated normal distribution (Bierlaire & Frejinger, 2005; Nielsen et al. 2000; Prato & Bekhor, 2007), before each SPA iteration. All previous studies produced more than 20 alternative routes by stochastic choice set generation methods in a short time. For example, Ramming (2001) extracted 48 draws from a normal distribution in the iterations of the SPA, and Bierlaire and Frejinger (2005) extracted 20 draws from a truncated normal distribution to perform a traffic safety study in Swedish. The means and variances equaled to the link travel time recorded by sample trips.

Stochastic choice set generation is efficient in time and cost; however, due to insufficient variation, many alternative routes and their properties are similar, which in turn increases the time of generating a required number of unique alternatives. Moreover, the negative values in normal distribution make the alternative unreasonable (Prato, 2009).

Deterministic route set generation methods include k-shortest path, link elimination, link penalty and path labeling. Path labeling is the most commonly used method in route choice analysis, which repeats the SPA with different cost functions called labels. The label can be a single variable or a function of multiple variables such as travel time, distance, the number of left turns, congestion, etc.

Su et al. (2010) considered elevation, slope, air pollution and vegetation in the path labeling process to generate recommendation routes for cyclists. Howard and Burns (2001), additionally, used route labeling to generate optimized routes in distance, time (directness) and perceived safeness in ArcGIS Desktop, and compared optimized routes with observed routes.

K-shortest path is an extension of path labeling. Instead of setting different labels, the k-shortest path produces routes with the same label by either breadth search first algorithm (Eppstein, 1998), Eppstein's algorithm (Eppstein, 1998), Yen's algorithm (Yen, 1971) or other advanced routing algorithms that can produce multiple routes when traversing or searching the network graph.

Menghini et al. (2009) applied k-shortest path with distance as the generalized cost in a multi-agent transport simulation toolkit (MATSim), and generated 60 alternative paths for each origin-destination pair. 20 alternatives were then randomly selected for route choice analysis. However, the alternative routes generated were found to be not effective, as small variations existed in alternatives generated by k-shortest path.

Broach et al. (2012), in addition, performed a comparison of three options: stochastic route choice generation, k-shortest path generation and path labeling, with revealed cycling data. Stochastic route choice generation and k-shortest path both have the weakness of generating

paths that are overly similar, which may violate the independence from irrelevant alternatives (IIA) property required for the subsequent route choice modeling.

The link penalty method increases the link costs by a factor (link penalty factor) to ensure the diversity among alternative routes, while link elimination follows a certain order to remove the links on the identified paths to avoid large overlapping among alternatives. The order of deletion can be random or by the order of link appearance in the routes (Rieser-Schüssler et al., 2013).

Lim and Kim (2005) adopted the method of link penalty, and checked the sensitivity of link penalty factor (Oz) to the degree of overlapping between route k and route n ($Op_{k/n}^{rs}$) defined as:

$$Op_{k/n}^{rs} = \frac{ol_{k/n}^{rs} \text{ (overlapping length between route } n \text{ and } k\text{)}}{l_n^{rs} \text{ (length of route } n\text{)}}. \quad (2-4)$$

The threshold of overlapping ratio (Op) was set beforehand, with the link penalty factor (Oz) being defined as:

$$Oz = \left[\frac{1}{Op}\right]^\alpha. \quad (2-5)$$

The new link cost is equal to the previous link cost multiplied by the link penalty factor. Only the route, whose degree of overlapping is within the threshold, would be accepted as one of the alternatives. Finally, they found that threshold of overlapping ratio (Op) between 0.5 and 0.7 with a larger than 4 value of α , produced the highest number of unique alternatives based on a virtual network.

Link penalties and link elimination can reduce the level of overlapping, and their efficiency is higher than other methods. However, as the link costs are changed by the link penalty, and some links are even eliminated in link elimination, they may produce some unreasonable routes without any meaning.

Deterministic route set generation methods use meaningful costs to generate routes, and avoid overlapping; however, a large number of scenarios need to be developed as generalized costs and the efficiency of generating large number of routes is not as high as in stochastic generation. Bovy and Fiorenzo-Catalano (2007), thus, developed a method combining the benefits of both stochastic and deterministic path generations called doubly stochastic. In the algorithm, both the attributes and coefficients in the label function in path labeling are randomized. Hood and Charlton (2013) later adopted this method to generate alternative paths. Although it was determined that this method is more time efficient with less overlapping than link elimination, it is too complicated to be implemented. It also has higher possibility to generate unreasonable routes, such as routes with long detours, if distance is not included in the generalized cost.

Table 2-2 shows a summary of various stochastic and deterministic route choice set generation methods as well as their benefits and limitations.

Table 2-2 Summary of Route Choice Set Generation Methods

Route Set	Route Set	Description	Benefits	Limitations	References
-----------	-----------	-------------	----------	-------------	------------

Generation Type	Generation Methods				Examples
Stochastic	-	The costs of links follows probability functions like normal distribution or truncated normal distribution	Efficient; Easy implemented	Large overlapping; Unreasonable routes	Bierlaire and Frejinger (2008)
Deterministic	Route Labeling	Minimizing or maximizing the generalized cost by scenario	Reasonable routes;	Need multiple labels;	Howard and Burns (2001)
	Link Penalty	Involved a penalty factor for existed links to avoid duplicated selection	Avoid overlapping	No reasonable meaning of routes; Penalty factor is hard to define	LIM and KIM (2005)
	Link Elimination	Delete the existed links following roles to avoid overlapping	Avoid overlapping; Increase the efficiency	Miss some links; No reasonable meaning of routes; Rule of deletion is hard to define	Rieser-Sch ü ssler, et al. (2010)
	K-Shortest Path	Generating k shortest path by various algorithms	Generate multiple routes by one scenario	Large overlapping; Hard to meet the requirement number of alternatives	Menghini, et al. (2009)

2.3 Route Choice Models

Most of the route choice models in previous studies were based on discrete choice theory, specially, random utility theory, where the degree of preference of an individual toward a choice option can be captured by a utility function, and individuals always choose the options with the highest utility. The utility function contains a deterministic part (a function of attributes) and a random error, which is used for capturing unmeasurable variables such as the taste or mood of decision-makers, systematic or measurement errors, etc. The deterministic part of utility function

is easy to measure, while estimating the random error is challenging. Therefore, discrete choice models assume the random error following some statistical distributions. Based on the different distributions of the random error, two kinds of models have been proposed in the literature: Probit models and Logit models.

2.3.1 Probit Model

The multinomial probit model developed by Daganzo and Sheffi (1977) assumes a multivariate normal distribution for the random component in the utility function. Daganzo and Sheffi (1977) were the first to apply multinomial probit model to estimate traffic assignment. The probability of choosing one of the options (route r) is described by:

$$P_r = \int_{\varepsilon=-\infty}^{\varepsilon_r+V_r-V_1} \dots \int_{\varepsilon_r=-\infty}^{\infty} \dots \int_{\varepsilon_r=-\infty}^{\varepsilon_r+V_r-V_R} \varphi(\varepsilon) d_{\varepsilon_r} \dots d_{\varepsilon_1} \quad (2-6)$$

where V_r is the deterministic part in the utility function of route r , ε is the random error, and the density function ($\varphi(\varepsilon)$) is based on a vector of means and a covariance matrix. The random component cannot be expressed in a closed form, so the greatest challenge when employing a multinomial probit model is the specification of covariance matrix.

Yai et al. (1997) developed a multinomial probit model in route choice behavior analysis in Japan. They developed a structured covariance matrix related to measurable overlapping variables, such as the common length among the routes. They assumed the covariance of route utilities was proportional to overlap length, and the stochastic loading was performed by numerical integration.

As numerical integration and simulation are needed in multinomial probit models, the calculation of the probability is complicated. Although the Probit model was confirmed to be quite realistic (Yai et al., 1997), it is far more computational intensive. The costly computational effort makes the Probit model seldom found in practice.

2.3.2 Logit Models

Logit models were the most commonly used route choice models, in which the error term is identically and independently distributed (IID), and follows the Gumbel Type I distribution (Ben-Akiva & Lerman, 1985):

$$P(\xi_i < x) = e^{-e^{-x}} \quad (2-7)$$

Then, the probability of choosing option i (A_i) among all the alternatives (A) in a multinomial logit model (MNL) is defined as:

$$P(A_i) = \frac{e^{V_i}}{\sum_{A_j \in A} e^{V_j}} \quad (2-8)$$

where A is the alternative route choice set from a specific origin to a destination, and A_j is one of the routes in alternative route choice set.

Multinomial logit model

The simple multinomial logit model estimates the probability of choosing an option directly from the route utility without any network structure factors. The utility value is generally formulated as a linear function of predictor variables.

Hopkinson and Wardman (1996) applied a simple multinomial logit model with utility calculated as a linear function of various bicycle facilities. The final Rho-square was relative high, and all the coefficients were consistent with the survey results. Li et al. (2005) adopted a binary logit model to estimate the preference of cyclists on changing their morning commute routes. The utility function was related with individual characteristics, commute journey attributes (e.g. departure time), and primary route attribution (e.g. time, distance, speed, idle stop, etc.). Similarly, Stinson and Bhat (2003), applied a logit model to estimate the preferences of link-level variables (e.g. road-way class, parking, bicycle facility, terrain, etc.), and route-level variable (e.g. travel time) on commute route choice across U.S. However, in their research, data was collected by stated preference surveys, and only two options were available. The logit model was applied as a binary logit model form. Furthermore, Usyukov (2013) estimated the effects of route related factors (e.g. route length, auto speed, auto volume, grade, the existence of bike lanes) through a simple multinomial logit model in Region of Waterloo.

Although the MNL is easy to implemented, it has a key limitation known as the Independent from Irrelevant of Alternatives (IIA) problem. The IIA problem states that the ratio of the probabilities of choosing two alternatives in an MNL model for each individual is independent of the availability or attributes of any other alternatives. It requires all the alternatives to be discrete, but perfect distinct alternatives never exist in route analysis. Overlapping in alternatives may affect the model performance, and causes counter-intuitively increase in the probabilities of choosing the overlapped routes.

In order to avoid the IIA problem, several modifications have been made in previous studies. The idea of the modifications were adding a factor that can estimate the overlapping between alternatives. C-logit model and Path-size logit model were the most commonly used. Other complicated model forms also existed such as cross nested logit model, logit kernel, subnetwork, but due to the complexity in implementation and low time efficiency, they are beyond description, and not considered in this research.

C-logit model

The C-Logit model proposed by Cascetta et al. (1996) includes a commonality factor in the utility function to estimate overlapping among alternatives. There were four forms of commonality factors (CF_i^{rs}) proposed in the literature, among which Equation 2-9 and 2-10 were commonly applied. The other two forms have been shown in other studies to not provide adequate performance (Ramming, 2001).

$$CF_i^{rs} = \sum_{j \in C^{rs}} \left(\frac{L_{ij}^{rs}}{\sqrt{L_i^{rs} L_j^{rs}}} \right) \quad (\text{Bliemer \& Bovy, 2008}), \quad (2-9)$$

$$CF_i^{rs} = \ln \left[1 + \sum_{j \in C^{rs}} \left(\frac{L_{ij}^{rs}}{\sqrt{L_i^{rs} L_j^{rs}}} \right) * \left(\frac{L_i^{rs} - L_{ij}^{rs}}{L_j^{rs} - L_{ij}^{rs}} \right) \right] \quad (\text{Ramming, 2002}), \quad (2-10)$$

where C^{rs} is alternative route choice set from origin r to destination s ; L_{ij}^{rs} is the common length of route i and route j , L_i^{rs} is the length of route i , and L_j^{rs} is the length of route j .

The increased value of the commonality factor means a higher overlapping between route i and j . As the involvement of commonality factor is aimed at addressing the improper increase in the value of utility, the commonality factor should have a negative coefficient in the estimation result.

Cascetta et al. (1996) developed both simple multinomial logit model and C-logit model to estimate the preference of route related factors (e.g. travel time, monetary cost, etc.) for 1471 paths chosen by truck drivers on the Italian inter-city road network. All the commonality factors were statistically significant, and improved the Rho square values by ranging from 0.02 to 0.04 when compared to simple multinomial logit models.

Bliemer and Bovy (2008), in addition, compared the first form of C-logit model (Equation 2-9) with other route choice models (e.g. simple multinomial logit model, Path-size logit models, cross nested logit model, etc). They found that C-logit model performed as well as Path-size logit model, which were better than other models. It also successfully captured the small changes between two similar routes.

However, Ramming (2001) compared both forms of commonality factors, and found that none of them were statistically significant. When the Path-size factor was included, it even turned to be positive. He concluded that the performance of C-logit model was affected by the size of alternative route set. The C-logit model performs well with small number of alternatives, but is not suitable for large urban network.

Path-size logit model

Similar to the C-logit model, the Path-size logit (PSL) model introduces a path size factor into the utility function to estimate the level of overlapping, and has been widely used in various studies (e.g. Ben-Akiva & Bierlaire, 1999; Hood et al., 2013; Menghini et al., 2009; Ramming, 2001). There are two forms of the path size factor (PS_{in}) specified as:

$$PS_i^{rs} = \sum_{a \in \Gamma_i} \frac{l_a}{L_i^{rs}} * \frac{1}{\sum_{j \in C^{rs}} \left(\frac{L_i^{rs}}{L_j^{rs}} \right)^\gamma \delta_{aj}} \quad (2-11)$$

$$PS_i^{rs} = \sum_{a \in \Gamma_i} \frac{l_a}{L_i^{rs}} * \frac{1}{\sum_{j \in C^{rs}} \delta_{aj}} \quad (2-12)$$

where Γ_i is the set of links in alternative i , l_a is the length of link a , L_i^{rs} is the length of route i from origin r to destination s and δ_{aj} is 1 if route j includes link a and 0 if not (Frejinger and Blerlaire, 2007). γ is a positive scaling term that penalizes long routes. Although $\gamma > 0$ improved the route choice model fit (Hoogendoorn-Lanser et al., 2005), Frejinger and Blerlaire

(2007) found that $\gamma > 0$ would also result in over-estimating corrections and illogical path probabilities. If the alternative routes do not contain unreasonable detours, Broach et al. (2012) suggested γ being set to 0, as shown in Equation 2-12. With the path size factor, the probability of choosing option i in a Path-size logit model (PSL) is defined as:

$$P(A_i) = \frac{e^{V_i + b \cdot \ln(PS_i^{r,s})}}{\sum_{A_j \in A} e^{V_j + b \cdot \ln(PS_j^{r,s})}} \quad (2-13)$$

where A is the alternative route choice set from origin r to destination s , and A_i is one of the routes in alternative route choice set. The path size factor estimates the level of overlapping of a route in the alternative route choice set, which is between 0 and 1. It is opposite to the commonality factor in the C-logit model. A higher value indicates a lower level of overlapping with other alternatives, and lower value of path size factor indicates the links in the route appear more times in alternative route choice set. For a unique route, the path size factor equals to 1, and it would be less than 1 for a route with partial overlapping. Due to the IIA problem, the utility would be over-estimated for overlapped routes. Adding path size factor in the utility function considers the effect of overlapping to the route's desirability (utility), and corrects the over-estimation to produce more accurate results. Because the logarithm form of path size factor has a value below 0, in order to correct the over-estimation of overlapped routes in the utility, the coefficient of the path size factor (b) in the utility should be positive (Bliemer & Bovy, 2008).

Within the literature, many route choice studies have applied Path-size logit models. For example, Menghini et al. (2009) tried various alternative route sets and combinations of factors with the Path-size logit model. The Path-size models performed well with Rho-squares larger than 0.5, and the path size factors were all statistically significant. Hood and Charlton (2013) additionally performed a GPS-based route choice analysis in San Francisco using the Path-size logit model. The estimation of path size factor was significantly larger than zero, and all the coefficients and significances of other variables were reasonable explained.

Frejinger and Bierlaire (2007) compared results from MNL and PSL, and developed a new route choice model called subnetwork model. They found that although the results from MNL and PSL did not have a large difference, according to the log-likelihood and Rho-square, PSL still performed better than MNL in all cases. However, one of the limitations of modified logit models was that they cannot correctly capture the correlations among attributes, and the best model option needed to be found by iteration approach.

Both the C-logit model and the Path-size logit model have been shown to have efficient model estimation processes (Ramming, 2001), and produce better results in route choice behavior analysis when compared to MNL (Bliemer & Bovy, 2009). Ramming (2001) compared the estimation time of various route choice models, among which MNL, C-logit and PSL only took 5 minutes, while cross nested logit model and other complicated models (e.g. Logic Kernel) took more than 1 hour. In addition, Path-size logit model shows a better robustness with respect to the size of alternative route set (Prato & Bekhor, 2007; Ramming, 2001). Furthermore, based on the log-likelihood measures, the Path-size logit model outperforms the C-logit model as well some other complicated model forms (Prato & Bekhor, 2007; Ramming, 2001), and so many previous studies have adopted PSL as modification of MNL.

2.4 Factors Affecting Route Choice

Based on data collected by surveys and GPS-enabled devices, researchers have found that cyclists do not choose their routes randomly, but rather choose their routes by following specific rules affected by a combination of factors (Howard & Burns, 2001). Therefore, the last challenge in route choice analysis is identifying the factors affecting cyclists' decisions. Based on previous studies, the factors can be roughly divided into two groups: alternative route dependent factors (trip distance, travel time, network characteristics) and generic trip attributes (e.g. socio-demographic characteristics and trip purpose). These factors are highlighted in this section.

2.4.1 Alternative Route Dependent Factors

Alternative route dependent factors including trip distance (travel time) and network characteristics are those related to alternative routes, and varies when the route changes. The values of these factors vary among alternative route set, which were considered significant in route choice behavior, and included in most of the previous analysis.

2.4.1.1 Trip Distance or Travel Time

Within the literature, among all the route dependent factors, trip distance and travel time were considered the most important ones (Aultman-Hall et al., 1997; Howard & Burns, 2001). The National Travel Survey represented that 58% of cycle trips were around 3.2 km, and only 5% exceeded 8 km (Department of Transport, 2013). According to 2001 National Household Travel Survey (NHTS), 41% of all trips in 2001 were shorter than 2 miles, and 28% were shorter than 1 mile (Pucher & Renne, 2003). Vandenbulcke et al. (2009) also concluded that for most cyclists, 10 km was the limitation of commuting trips according to the sample trips in Belgium.

Bovy and Bradley (1985) examined cyclists' route choice behavior in Delft, Netherlands based on stated preference data, and found that travel time always played the most important role in cyclists' decision followed by road types. In addition, Aultman-Hall et al. (1997) compared the actual routes with the shortest possible paths in Guelph, Ontario, and found that only 14.6% of the routes were exactly the same as the shortest paths; however, many of the routes chosen still showed an obvious preference to the shortest path, as 37.5% of routes were within 0.1 km of the minimum distance, with average deviation from the shortest path only being 0.4 km. Furthermore, Howard and Burns (2001) compared the shortest route, most direct route, and safest route with the observed route, and also found that for bicycle commuters, the shortest path was the one had a majority of arcs in common with the actual route (50.6%) when compared to that of the other two routes (43.8% for the most direct route, and 20.4% for the safest route).

2.4.1.2 Network Factors

Despite the importance of distance and travel time, cyclists do not always choose the shortest possible path. In the study done by Aultman-Hall et al. (1997), more than 80% of the chosen routes were found to not follow the shortest paths. They were therefore affected by other factors. Within the literature, one important class of factors commonly found to influence decision

process are network factors. Network factors include motor traffic volume, surface gradient, road type, bicycle facility type, existence of parallel on-street parking as well as road surface conditions. In previous studies, network factors are further divided based on two kinds of related effects, namely, factors affect the perceived safety risk and factors affect the comfort level.

The term perceived risk or stress level was first mentioned by Davis (1987), and supported by Sorton and Walsh (1994) and Hopkinson and Wardman (1996) demonstrating that 91% of surveyed cyclists preferred to cycle more if there was a safe cycling network, and risk reduction was more important than reducing travel time. Parkin et al. (2007) also supported perceived risk as one of main factors in route acceptance analysis. Aultman-Hall and Kaltenecker (1999) did an analysis on the safety of bicycle commuters in Toronto, and found that the injury rate for cycling were 26 to 28 times higher than that for automobile travel, which means the safety is more sensitive to cyclists than to automobile travelers.

Some researches presented the safety or perceived risk by perceived risk index or stress level. Davis (1987) first developed a mathematical model to index bicycle safety by roadway segment index and intersection evaluation index. The road segment index was a function of daily traffic, number of lanes, speed limit, lane width, pavement surface conditions, and location factors, while the intersection evaluation index was related to the cross street volume, traffic volume, geometric factors and signalization factors. However, in his method, the involvement of many construction related factors (e.g. pavement surface conditions, geometric factors and location factors) decreased the influence of other objective factors (e.g. traffic volume and speed limit). Later improvements by Epperson (1994) also did not perform well in safety analysis and route choice analysis.

Ladis et al. (1997) modeled the human perceived bicycle level of service based on 150 cyclists and 4,300 trip observations in Tampa, Florida. They modeled the perceived bicycle level of service by linear regression with traffic volume, speed of motor vehicles, pavement surface conditions and distance to motor vehicle lanes. They found that extra pavement width and a facility separating two traffic modes (motor and non-motor) increased the level of service to a safer condition. Comparing the separated stripping bike lanes and wide curbs, bike lanes were more welcomed than wide curbs for cyclists. A separated stripping bike lane may increase nearly 50% of quality-of-service score.

In addition, Sorton and Walsh (1994) estimated stress level by three proposed variables, lane volume, curb width, and motor speed limit, and rated the variables into 5 scores based on the perceived risk considered by three types of bicyclists: experienced, casual and youth. This study was vague in classifications, and none of the scores for three variables were exactly the same as the proposed stress level according to the surveys. However, the results still confirmed the negative effects of high traffic volume and speed limits, and the positive effect of curb width. It also proved the fact that different types of bicyclists have different attitudes towards same variables. For example, experienced bicyclists could tolerate higher speed and higher volume of motor vehicles than casual and youth bicyclists.

Parkin et al. (2007) further estimated the route preferences of road type, junction and turn type, bicycle facility type, average traffic volume and average number of parked vehicles for 144 commuters in Bolton between January and July 2002. They found that residential roads with

on-street parking and busy roads have higher perceived risk. Although bicycle facilities reduced the perceived risk, the road type and parking played more important roles in risk estimation. For generic features, the proportions of off-road and adjacent bicycle facilities largely decreased the perceived risk by 1.15 to 1.67, while other features such as average flow, average number of parking vehicles, and different kinds of junctions had positive effects on perceived risk. Higher perceived risk causes lower possibility of choosing the route. Therefore, bicycle facilities (off-road, adjacent) had positive coefficients (1.886 and 1.938) in the route acceptability models.

Bicycle facility type and road type are two main factors that may affect cyclists' perceived risk and route acceptance according to the literature. In the 1999 Cycling Survey, 93% of cyclists were comfortable cycling on bike trails (City of Toronto, 2011). They supposed streets with bike trails, which separates motor and non-motor traffic were safer than roads without bicycle facilities. In addition, 87% of them enjoyed cycling on local roads. 53% of cyclists felt easy to ride on major roads with bike lanes, but only 18% of cyclists were comfortable with major roads without bike lanes. Local streets were considered safer than major streets without bicycle facilities. Furthermore, 38% of participants believed that more on-street bike lanes would encourage cycling in Toronto, and 14% of them thought off-street facilities would have better consequences than on-street facilities.

Abraham et al. (2002) developed a logit model for estimating route choice behavior in the city of Calgary with stated preference data. After comparing the coefficients, they found that, road type was more important than bicycle facility type. A 10-minute change from arterial to residential street in a route increased the utility by 0.5 unite, while a 10-minute change to bike lane increased utility by 0.4 unite. Among all types of bicycle facilities, pathway was the most preferred, followed by bike lane. A 10-minute change from arterial without bicycle facility to arterial with pathway increased the utility by 0.6 unite, while wide curb lane only resulted in a 0.35 unite increase in the utility. The results confirmed the inference that off-street facilities were substantially more onerous, and cyclist strongly preferred low-traffic residential roads.

Contradicting results have also been reported in the literature, such as a study by Aultman-hall and Kaltenecker (1999), which employed statistical analysis based on the reported cycling collision and injury rates. They found streets with on-street facilities have the lowest collision and injury rates followed by sidewalks and off-street paths. These results contradict other studies and surveys, such as work done by the City of Toronto (2011) and Abraham et al. (2002), which suggest that off-street bicycle facilities were safer than on-street facilities. One possible explanation is that the skills of cyclists on sidewalks and off-street pathways varied substantially, so collisions and injuries may happen more often on sidewalks and pathways. In addition, as cyclists need to share roads with pedestrians and rollerblades on sidewalks and off-street paths, collisions may happen between pedestrians and cyclists.

Stinson and Bhat (2003) did a study further analyzing the effects of 11 link-level and route-level factors (e.g. roadway class, bicycle facility type, travel time, facility continuity, stops, etc.) to human perceived risk and route choice behavior by stated preference surveys across the U.S and, with more detail, in Austin, Texas through a discrete choice modeling framework. They found an obvious preference for bicycle lanes and separate paths. An additional separate path built may increase the probability of choosing the link by a factor of 1.78, and an additional bicycle lane by a factor of 2.07, which was higher than separate path. The study also confirmed the results in

other studies that residential streets were the most preferred by cyclists, followed by minor arterial. If one link was changed from a residential street to a major arterial, they found that this would decrease the probability of that link being chosen by a factor of 1.76. In addition, the on-streets parking were avoided by cyclists with a factor of 0.43 decrease in probability.

Hood et al. (2001) grouped bicycle facilities into three classes, off-street bike paths, separated stripping bike lanes, and on-street shared lanes, and found similar results with Aultman-hall and Kaltenecker (1999). They concluded that bike lane was preferred on average followed by bike path. Cyclists were willing to make 0.49 to 0.92 km detours to choose routes with bicycle facilities. However, in their study, traffic volume had a positive coefficient, which was counter-intuitive and rejected. Tilahun et al. (2007) improved the analysis by further dividing facilities into five types based on their characteristics: off-road, in-traffic, bike lanes, and on-street parking. They also found that bike lane was more preferred than off-street facilities. A bike-lane improvement was valued 16.41 minutes in travel time, the off-road improvement was less important with a value of 5.13 minutes in travel time. Bike lane was identified as the most important improvements for cyclists because the perceived safety would increase without reducing the speed.

They also confirmed the effect of parallel on-street parking. No parallel on-street parking was valued 9.27 minutes in travel time, which was even more important than off-road bicycle facilities. Krizek (2006) also showed the importance of parallel on-street parking. A trade-off analysis showed that the absence of on-street parking was equivalent to 8.9 minutes in travel time, while an off-road facility worth only 5.2 minutes.

Sener et al. (2009) continued estimating 10 factors for non-commute trips and 11 factors for commuting trips produced by 1621 individuals in Texas. The attributes were divided into 6 types: personal characteristics, on-street parking, bicycle facility type, and roadway physical, functional, operational characteristics. They confirmed the positive effects of bicycle facilities especially bike lanes. They also found that cyclists preferred no on-street parking followed by angled parking, which provides more maneuvering room and more time for cyclists to react. Parallel on-street parking was the least preferred, and a route with parallel parking was around 0.5 utility unite less attractive than a route with no parking for all kinds of cyclists.

In general, most of the previous studies confirmed the positive effects of bicycle facilities, and low-traffic residential streets. Road type was confirmed more important than bicycle facility type. Among all kinds of bicycle facilities, some researches preferred off-street pathway, while others still though separate stripping bike lane was enough for them considering both safety and speed. Previous studies also concluded the negative effect of on-street parking, among which parallel parking was the least preferred.

Comfort level parameters describe whether cyclists are comfortable during the entire route. Although safety issue also affects cyclists' comfort level, comfort level is more related with cyclists' preferences to gradients, number of turns, scenery and on-going facilities. It was confirmed that all cyclists avoid high number of traffic controls to prevent delays (Sener et al., 2009). Aultman-Hall et al. (1997) compared cyclists' routes with the shortest path. The number of turns in the actual route was much smaller than that in the shortest path, which means that cyclists considered the signal delay a significant factor when choosing routes. Parkin et al.

(2007), in addition, found negative preferences in junctions, among which roundabout was considered the most dangerous type of junction, followed by right turn. One additional right turn decrease the probability of choosing the route by 0.343, and 0.115 for one additional signal control. Furthermore, Hood et al. (2001) got similar result that cyclists were willing to detour an extra 0.17 km to avoid a turn.

A number of studies have assessed the effects of road gradient on the route choice behavior of cyclists. For example, Menghini et al. (2009) and Hood et al. (2001) both considered trip length, average absolute gradient, maximum gradient, percentage of various bike paths, and number of traffic lights in the Path-size logit models for route choice behavior analysis. They found that chosen routes were always shorter, less steep, with fewer signal lights and more bicycle paths when compared statistically with non-chosen paths.

However, contradicting findings were found in gradient, for people's attitudes towards gradient were quite different. The results may also be easily affected by sample selection. For example, most of studies including Stinson and Bhat (2003), Menghini et al. (2009) and Hood et al. (2001) found flat gradient had positive effect on route choice, while Sener et al. (2009) found that cyclists prefer moderately hilly terrains as they want to prevent boredom.

Scenery and on-going facilities were more sensitive to recreational trips than commuting trips, as discussed in later section. However, in general, on-going facilities such as parking spots, restaurants along the routes would increase the attractiveness of the route if not concluded insignificant (Chen & Chen, 2010).

2.4.2 Generic Trip Attributes

Besides route dependent attributes, generic trip attributes such as socio-demographic factors and trip purposes are common to all alternative routes. In other words, the values of generic trip attributes do not change among the different alternative routes, but vary by different trip instances. The influences of these generic trip attributes on route choice behavior have been mentioned in some studies, and considered significant in route choice decisions. However, the results of the generic trip attributes are often not transferrable, and are easily affected by the sample size and diversity of study areas, so studies may determine different results by sampling different sample and in different experiment locations.

2.4.2.1 Socio-demographic Factors

As key generic trip attributes, socio-demographic factors including household income, gender, age, etc. were proved in some researches to have slightly influences on route choice decisions when compared to other variables (Hulse & Shafizadeh, 1996; Sener et al., 2009). However, few studies have been done involving these factors due to the complexity of data collection and cyclists' concern about the privacy issue. Some studies that included them found that demographic factors to be not significant when compared to trip operational and network factors. For example, both Hood et al. (2001) and Abraham et al. (2002) found none of the demographic factor had significant effects on route choice decisions when compared to other factors (e.g. travel time, bicycle facilities, road types, etc.). Krizek (2006) confirmed that income and gender

were not significant in mixed logit regression model, while household size had slight negative effect on travel time. A household with more than 2 people always chose shorter routes when compared to single or two-person households.

Antonakos (1994) estimated environmental and personal factors by survey data, and found that gender slightly influenced commute route choice. Women preferred wide curb and bike paths more than men, and cycling experience were positively correlated with the use of wide curb lanes in recreational trips. Gender was also found affecting route choice by its preference on on-street parking (Sener et al., 2009). Males were found to be more sensitive to street parking than females due to their higher speeds (Sener et al., 2009).

Hulse and Shafizadeh (1996) included household income, gender and age in their analysis, and found that household income had an influence on commute distance. Households with an annual income between \$7,500 and \$15,000 tended to have shorter commute travel times when compared to other ranges of income. A year later, Shafizadeh and Niemeier (1997) additionally analyzed the interactions between demographic factors and commute travel time separated by regions, and found that regions affected the influences of demographic factors. For example, the results showed that income had a positive effect to travel time when destinations were within central business district (CBD), while income had a negative effect if the destinations were outside CBD. That may be because the residence of people with higher income was further than those with lower income (Shafizadeh & Niemeier, 1997).

2.4.2.2 Trip Purposes (Commuting Trips vs. Recreational Trips)

In general, trip purposes do not have significant effects on cyclists' route choice behavior, commuting and recreational cyclists had quite similar preferences (Sener et al., 2009); however, some researchers found different results based on commuting and recreational trips (Broach et al., 2012; Chen & Chen, 2010). Antonakos (1994) performed a stated preference survey by asking cyclists to assign scores for four recreational bicycle tours based on various environmental and personal factors in Michigan, and tried to estimate the influences of factors to recreational route choice decisions. For both commuting and recreational trips, bike lanes, wide curb lanes and bike paths were preferred. Commuting cyclists were found to consider directness (trip distance or travel time) and flatness (gradient) more important, while recreational cyclists valued scenery, and did not care much about distance. The study also found that recreational cyclists enjoyed riding on wide curb lanes, while in commuting trips, bike lanes were the most preferred bicycle facility as bike lanes are more convenient to errand considering both safety and traveling speed.

Chen and Chen (2010) developed a multinomial logit model to quantitatively simulate recreational route choice decisions. They considered 6 attributions including trip length, road types, slope, bicycle facility types, and found that the coefficients of bike route (-0.84 and -1.08) were negative suggesting bike path was more preferred than bike route. The other facilities along the route such as restrooms, bicycle parking spots, maintenance equipment, information centers had positive coefficients (from 0.08 to 0.44), and were verified more significant for recreational trips than commuting trips especially for long time recreational trips.

In addition, Broach et al. (2012) analyzed 1,449 commute and non-commute route choices, and found that cyclists for both commute and non-commute purposes preferred separated bike paths, followed by bike boulevards. They found that bicycle commuters were more sensitive to distance and seemed unaffected by other factors as non-commuting cyclists. Furthermore, commuting cyclists were found to be more sensitive to high traffic volume than non-commute cyclists due to the fact that commuting trips often happen during rush hours. Finally, they found that turn frequency had significant negative effect on the route choice for all purposes, while non-commute cyclists were more sensitive to this factor than commuting cyclists. Their model presented a 7.4% increase in non-commute distance or a 4.2% increase in commute distance, which was equivalent to one additional turn per mile (0.6 turn/km).

However, few studies involving recreational or other non-commute trips have been done due to the difficulties in data collection and the limited sample size. Comparison between these two purposes, however, highlights the differences and provides valuable guidance for governments and planners to improve the quality of cycling network wisely.

2.5 Summary of Previous Studies

This chapter describes the four challenges in previous studies about route choice analysis: data collection, alternatives generation, model estimation and factors identification. Earlier studies were mainly based on revealed or stated preference surveys to ask respondents choosing the routes they preferred or drawing their recent routes on a map or in a GIS software (Broach et al., 2012; Howard & Burns, 2001; Hyodo et al., 2000; Stinson & Bhat, 2003; Tilahun et al., 2007; Yang & Mesbah, 2013). However, surveys are difficult to implement and are time-consuming. As GPS-enabled devices have become popular in recent decades, route choice analysis based on GPS location data increasingly presents significant advantages over traditional survey methods in both time, cost and ease of the implementation. However, one of the greatest challenges using GPS data is map-matching. Previous studies provided various methods to snap GPS points to network links (Hood et al., 2011; Menghini et al., 2009; Quddus et al., 2003; White et al., 2000), but none of them has significant advantages in accuracy, complexity and efficient scaling on large networks.

With revealed data records from revealed preference surveys and GPS tracking, only chosen routes are known, the second challenge is the way to identify alternatives following cyclists' perceptions. Shortest-path algorithms such as the Bellman-Ford algorithm, Dijkstra's algorithm and the A* algorithm are commonly used for this task and have been introduced above along with their benefits and limitations (e.g. Aultman-Hall et al., 1997; Fu et al., 2006; Rieser-Schüssler et al., 2013; Zeng & Church, 2009). Of these, due to its simplicity and ease of implementation, Dijkstra's algorithm is the most commonly used algorithm in transportation analysis as well as some network simulation software (ArcGIS and MATSim) (Atkinson et al., 1986; Brodal, 1996; Fredman & Tarjan, 1987). Improvements such as a priority queue in Dijkstra's algorithm were confirmed better performance in previous studies (Atkinson et al., 1986).

The alternative route set contains reasonable, unique route options, which was generated by the successive iterations of the shortest path algorithm. Two approaches were mentioned in previous studies: stochastic choice set generation and deterministic choice set generation. Stochastic

choice set generation, in which link costs follow statistic distributions, has high efficiency but has a limitation in alternative similarity (Bliemer & Bovy, 2008; Nielsen et al. 2000; Prato & Bekhor, 2007; Ramming, 2001). Deterministic choice set generation methods (k-shortest path, path labeling, link penalty and link elimination) use label as link costs to generate meaningful paths such as trip distance and perceived risk (Broach et al., 2012; Howard & Burns, 2001; Menghini et al., 2009; Su et al., 2010; Menghini et al., 2009). Link penalty and link elimination were proposed for avoiding overlapping among alternatives by using a factor (link penalty factor) to increase the link costs or purposefully eliminates selected links in each iteration (Lim & Kim, 2005; Rieser-Schüssler et al., 2013).

In order to estimate the influences of cyclists' route choosing behavior, previous route choice models were based on the discrete choice modeling framework. Two random utility models are commonly applied, the Probit model and the Logit model, which differ in the distribution of the random error they use. The Probit model adopts normal distribution for random error, which requires simulation (Daganzo & Sheffi, 1977; Yai et al., 1997). The numerical integration and simulation make this model difficult to be implemented, and far more computational intensive. Logit models are commonly applied in route choice analysis, but are vulnerable to the independent irrelevant of alternatives (IIA) property that requires alternatives to be perfectly distinct (Hopkinson & Wardman, 1996; Stinson & Bhat, 2003; Usyukov, 2013). In order to avoid IIA problem, modified logit models such as C-logit model and Path-size logit model have been proposed. This models avoid the problem by using factors that can estimate how much the overlapping exist in alternative routes (Bliemer & Bovy, 2008; Cascetta et al., 1996; Hood & Charlton, 2013; Menghini et al., 2009). Path-size logit models have been shown to be robust with respect to the size of alternative route set, and, based on comparisons of the log-likelihood measures, generally outperform C-logit models, multinomial logit models and other complicated models in most of the studies done (Bliemer & Bovy, 2009; Frejinger & Bierlaire, 2007; Ramming, 2001).

Finally, many previous studies have confirmed the effects of both alternative route dependent and generic trip attributes. Route dependent factors are commonly identified in the literature, which includes trip distance, travel time and network characteristics. A number of studies highlighted above conducted models to estimate the influences of those factors on route choice decisions. In general, cyclists prefer shorter, safer and more direct routes. They also prefer low traffic volume routes and local streets with more bicycle facilities and less on-street parking (Abraham et al., 2002; Aultman-hall & Kaltenecker, 1999; Hood et al., 2001; Parkin et al., 2007; Sener et al., 2009; Stinson & Bhat, 2003; Tilahun et al., 2007). However, contradicting findings have been reported in different studies. For example, Abraham et al. (2002) and Parkin et al. (2007) found that separated off-street bicycle facilities like pathways were more preferred than on-street facilities like bike lanes, while Aultman-hall and Kaltenecker (1999), Hood et al. (2001), Sener et al. (2009), Stinson and Bhat (2003) and Tilahun et al. (2007) all concluded that considering both safety and speed, bike lane was the most preferred bicycle facility type followed by the separate pathway. In addition, in term of the gradient, Menghini et al. (2009) and Hood et al. (2001) found that cyclists preferred flat roads, while Sener et al. (2009) found that moderate slope was more preferred to avoid boredom.

Generic trip attributes includes trip purposes and socio-demographic characteristics, which remain the same for all alternative routes. In previous studies, recreational cyclists were found

less sensitive to trip distance and travel time, but care safety, scenery, and on-route facilities (e.g. fountains, parking spots, restaurants, etc.) more than commuting cyclists (Broach et al., 2012; Chen & Chen, 2010). Local streets were found to be more preferred than major roads, and considered more important than bicycle facilities (Abraham et al., 2002; Stinson & Bhat, 2003). Demographic information was confirmed having slightly influences on cyclists' choices if they were not concluded insignificant (Antonakos, 1994; Hulse & Shafizadeh, 1996; Krizek, 2006; Sener et al., 2009). As generic trip attributes were easily affected by sample cyclists and study areas, few studies include this type of attribute in their analysis and some inconsistent results were found in different studies.

In conclusion, many previous studies lack a systematic investigation of the various aspects of route choice modeling, particularly with respect to alternative route set generation and model selection on the basis of complexity, accuracy and computational power. In addition, previous studies have limitations in terms of generalizability and transferability of their findings and conclusions due to the use of choice data with low resolution, small sample sizes, and incomplete attributes. This research aims to fill the gap in the literature and develops systematic methods to perform the cyclists' route choice analysis, as detailed in the following chapters.

Chapter 3 Data Description

This chapter describes the data used in this research including GPS observations obtained from the participating cyclists, network related datasets (e.g. bikeway network, digital elevation model, traffic volume, transit stops and parallel on-street parking permitted areas), individual information (user profile data), and weather data. The study area of this research is limited to the City of Toronto, including the downtown core and surrounding boroughs.

3.1 Study Area

All of the recorded trips originate and terminate in the City of Toronto, including Etobicoke, York, North York, Scarborough, East York and Old Toronto with an area of 630 km², which, according to the 2011 census in Canada, has a population of 2,615,060 and is Canada's most populous metropolitan area (City of Toronto, 2016).

According to the 2001 census, 1.7% of people in the City of Toronto use their bicycle regularly to commute work. The City of Toronto has the highest bicycle mode share of any municipality in the Toronto and Hamilton Census Metropolitan Areas (City of Toronto, 2016). The average bicycle mode share across the Toronto and Hamilton area is 0.5%, and the average bicycle mode share across Canada is 1.3% (City of Toronto, 2016). Figure 3-1 shows the basic network in the study area, and also a trip example in the dataset.

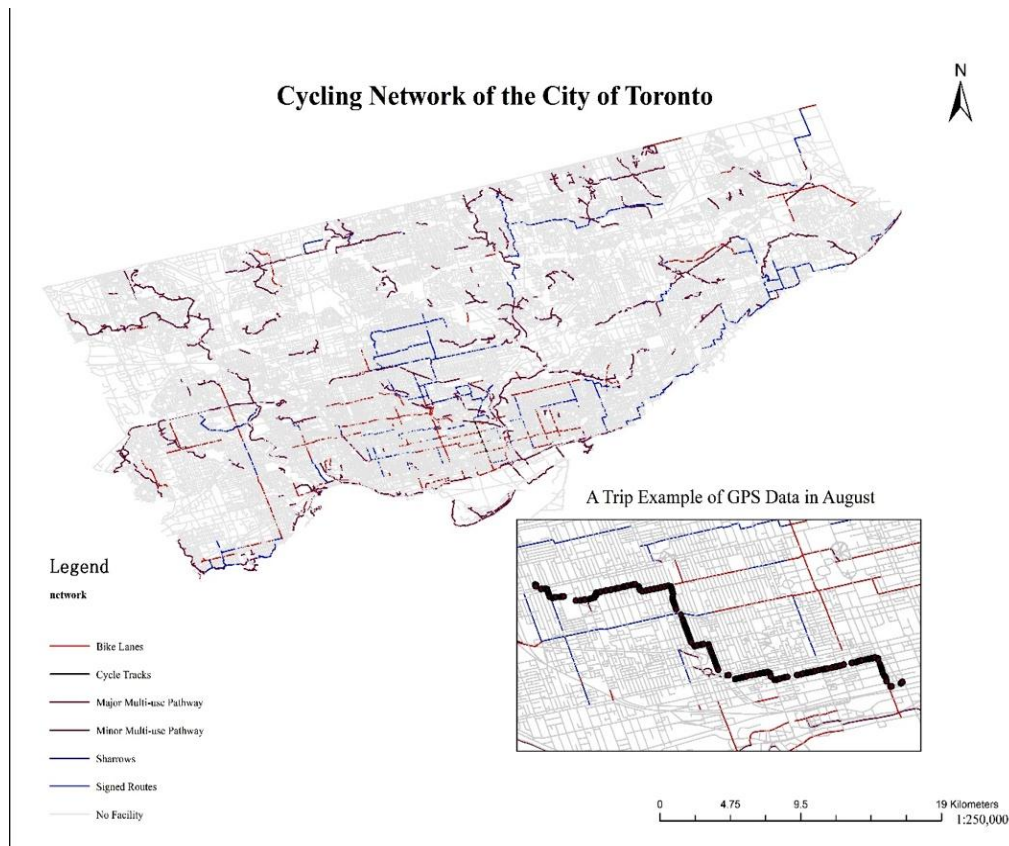


Figure 3-1 Cycling Network in the Study Area

3.2 Data Sources

Revealed GPS observations

For this study, per-second GPS observations were collected by a smartphone application, CycleTrack, developed by Brisksynergies. Participants open the application when they start cycling and the smartphone then collects the location of the participant every second as a latitude and longitude coordinate. The collected data points are then uploaded onto the server. At the end of the trip, the application asks participants to specify the purpose of their trip and add their personal information. Participants can also skip this step and turn off the application after the trip ends.

GPS observations track the routes of cyclists and provides a more accurate way to collect revealed data. However, some of the participants may forget to turn off the application when they arrive at their destination or change traveling modes. Smartphone GPS locating services also have system errors, and while these errors are relatively low in urban areas with good visibility, they can still be substantial. For most smartphones, the average error is usually around 20 metres, but can range anywhere from 5 to 35 meters (Paek et al., 2010). It is therefore important to pre-process collected GPS data by filtering outliers not in the study area.

In this study, around 36 million observations were collected from May 20th, 2014 to January 25th, 2015. In addition to latitude and longitude, each observation is accompanied by a timestamp identifying the recorded time and the sequence in the trip. Figure 3-2 shows an example of GPS observations.

trip_id	coord_id	recorded_at	longitude	latitude	altitude
14325	11335515	2014-06-12 13:59:18.717	-79.464114	43.677043	89.9000015258789

Figure 3-2 An Example of GPS Observations

User profile data

An optional survey is included in the smartphone application to collect profile information, including age, income, gender, rider history, cycling level or comfort level, whether or not they cycle in winter, and email. One device or account is linked with one user; personal information is entered when signing into the application. However, as participants are not required to complete the user survey, many opted not to complete it due to privacy concerns. Of the collectable fields, gender had the poorest response rate, with 72.8% of participants preferring not to disclose their gender. Response rates to most of the other fields were generally high, with around 80% of respondents answering all other questions.

The unique value used to identify a user profile is predefined by the City of Toronto. Participants do not need to use words to answer the questions; they just need to pick which group they are in. As shown in Appendix II, each attribute is classified in 3 to 7 groups, and the answers are integers from 1 to 7. Figure 3-3 shows an example of user profile data. Appendix II provides detailed user information for each attribute.

user_id	app_user_id	age	income	gender	rider_history	cycling_level	winter	email
3137	3213	4	6	0	0	3	2	

Figure 3-3 An Example of User Profile Data

Network related data

Network related datasets include the bikeway network, digital elevation model (DEM), traffic volume, transit stop locations and parallel on-street parking permitted areas. Most of these datasets were obtained from Toronto Open Data except for DEM, which was obtained from Geospatial Centre at the University of Waterloo and the traffic volume (AADT) data, which was obtained from Toronto Traffic Management Center.

The bikeway network contains one direction roads with road type, bicycle facility information, as shown in Figure 3-1. The road type and bicycle facility information is included as attributes in bikeway network, and has been reclassified into several groups. The road types are classified into major arterial, minor arterial, collector, local streets and others (see Table 4-1), and the bicycle facilities are classified into bike lane, shared lane, multi-use bikeway, cycle tracks, suggested on-street route, contra-flow bike lanes, suggested on-street connection, park roads cycling connection, and informal dirt footpath (see Table 4-2). Detailed descriptions based on road classification management system are shown in Table 4-1 and Table 4-2 (City of Toronto, 2016).

However, as the bikeway network only contains one-direction network links, in a later pre-processing step, a bidirectional bikeway network was created and intersections (nodes) were identified based on this dataset.

Transit stops were obtained in text files containing location descriptions (e.g. Davenport Rd At Bedford Rd), and their locations were first geocoded. Parallel on-street parking permitted areas were obtained in shapefile format (GIS format) describing geocoded polygons where parallel on-street parking is permitted. All the streets in the polygons have parallel on-street parking. Figure 3-4 shows the map of bus stops and parallel on-street parking areas.

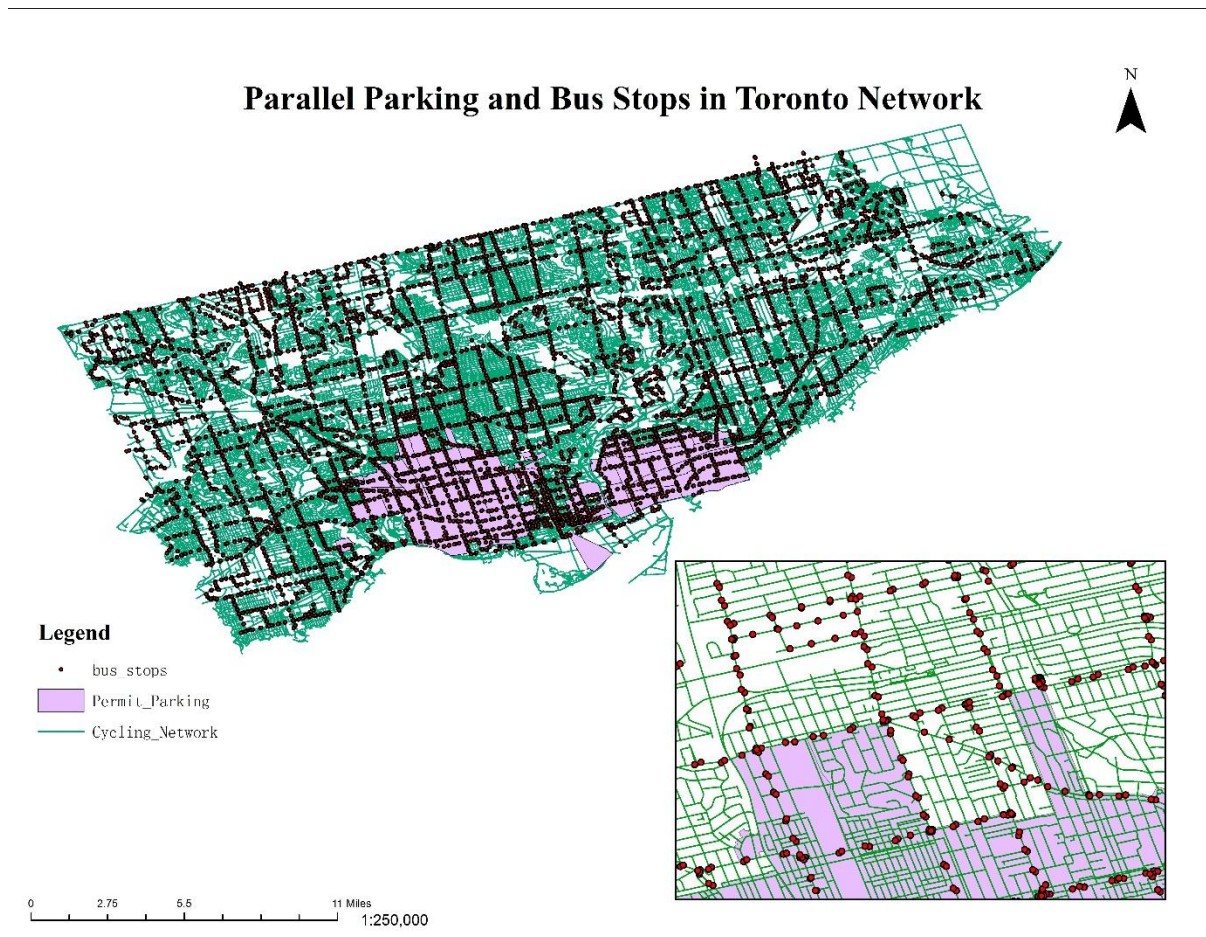


Figure 3-4 Map of Parallel Parking Area and Bus Stops

The digital elevation model (DEM) was obtained as a raster image, with each pixel of the image representing elevation information at that point. The model was coded in a GIS format covering the entire city of Toronto and had a cell resolution of 10 meters. Geospatial software (e.g. ArcGIS) tools were then used to analyze the slopes provided in the DEM and code them to the links imported from Toronto's Open Data.

The annual average daily traffic volume (AADT) was obtained from Toronto Traffic Management Center, showing the AADT for every main streets in a time period from May 2014 to February 2015. However, some streets in the bicycle network were not included and did not have volume records. Links without traffic volumes were assigned an estimated average AADT value calculated by values of the links with same road type in the traffic volume dataset.

Weather data

Weather data was obtained from Environment Canada containing all the temperature, precipitation (mm) and snow amount (cm) from May 2014 to February 2015. The precipitation amount is only for the rain precipitation, and snow amount is separately recorded.

Table 3-1 shows a summary of all the data sources used in this research: GPS observations as well as trip purpose, user profile information, and network related datasets.

Table 3-1 Summary of Data Sources

Dataset	Data Format	Source	Content	Columns
GPS Data	csv	CycleTrack	Tracing the locations of cyclists with latitude and longitude	trip_id: identify the trip of the location coord_id: identify each GPS point recorded_at: the timestamp of the point longitude: longitude of the location latitude: latitude of the location altitude: altitude of the location speed: speed detected by the device hort_accuracy: the accuracy of the latitude vert_accuracy: the accuracy of the longitude
Trip Purpose	csv	CycleTrack Survey	Describing the purpose of the trips detected by application	trip_id: identify the trips purpose: text describing the purposes including commute, shopping, exercise, etc. In this study, only commuting trips are studied
User Profile	csv	CycleTrack Survey	Describing the demographic or social-economic characteristics of cyclists; Appendix II shows the values of people's choices	user_id: identify the users app_user_id: identify the application age: age of the cyclists income: household income gender: gender of the cyclists rider_history: how long the cyclists riding cycling_level: comfort level of the current road conditions winter: whether the cyclist cycling in winter email: emails for backward surveys
Bikeway Network	shapefile (WGS 84)	Toronto Open Data	The GIS shapefile of edges of Toronto bikeway network	link_id (geo_id): identify links link_name: the street names one_way_di/one_way_desc: describe whether it is a one way street fcode/type_desc: describe the type of the roads like arterial, local, collector etc. bicycle_type: describe the type of bicycle facility of the road like bike

				lanes, sharrows, contra-flow bike lane, etc.
Digital Elevation Model (DEM)	tif	Geospatial Center of University of Waterloo	The 10- meter DEM of the entire Ontario from OMNI	Raster dataset containing the elevations of every pixel
Traffic Volume Data	csv	City of Toronto	The traffic volume detected by sensors in the main streets in the City of Toronto	Contains the auto-traffic volume in most of the main streets.
Transit Stops	csv	Toronto Open Data	The locations of transit stops	latitude, longitude: specify the location of transit stops route_num: the number of transit route
Weather Data	Excel	Environment Canada	The weather of the date from May 2014 to January 2015	mean temperature, total precipitation in mm, snow amount in cm
Parallel Parking Permitted Area	shapefile (WGS84)	Toronto Open Data	The areas, where parallel parking is permitted	name: the name of the area

Chapter 4 Data Processing and Exploratory Data Analysis

Chapter 3 describes the data sources used in this study including GPS observations as well as user profiles and trip purposes, network related GIS datasets (bikeway network, digital elevation model, bus stops, parallel parking area, and traffic volume), and weather data. However, raw data must be preprocessed before they can be used for subsequent data analysis and model development. In order to analyze the routes cyclists travelled on, GPS point data need to be matched to the network. This chapter describes the procedures in data pre-processing and map matching. Exploratory data analysis is also presented in the last section.

4.1 Data Pre-processing

Data pre-processing is an integral component of the model development process, aiming at preparing and integrating all data sources for subsequent analysis. The flow of this pre-processing step is shown in Figure 4-1. After all the datasets being processed, a geospatial data management system (PostgreSQL) was used to manage all the data stored in a common database.

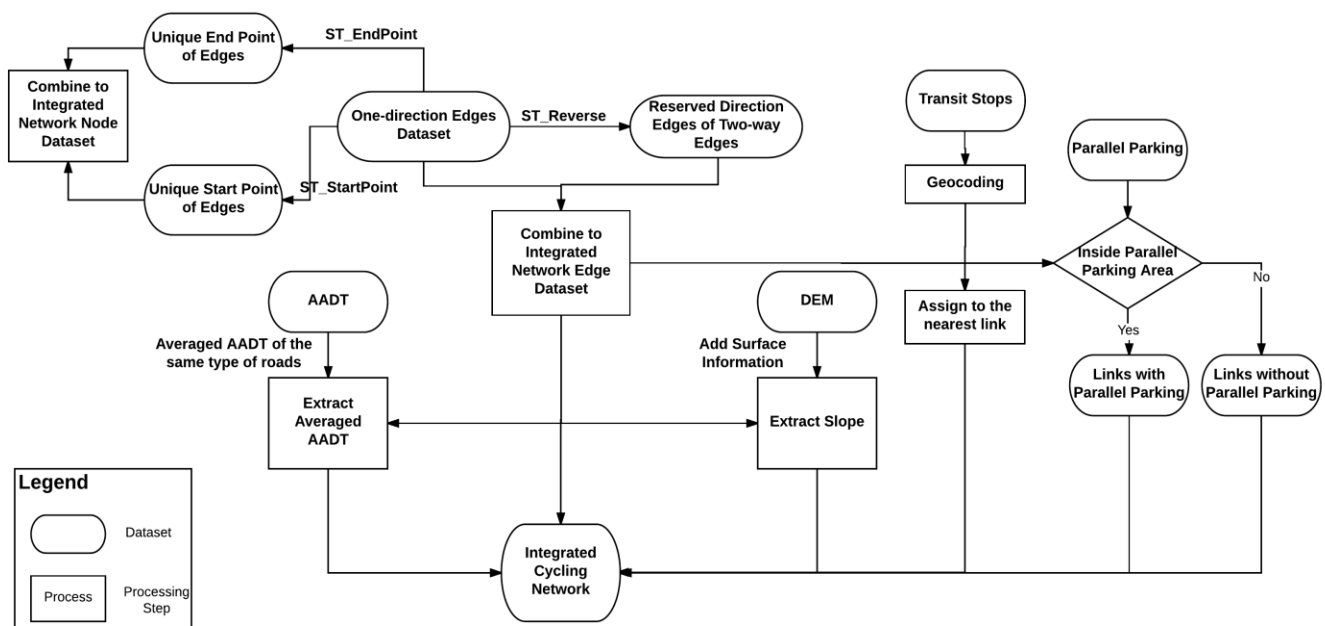


Figure 4-1 Pre-Processing of Network

Geospatial software (ArcGIS) and GIS-enabled database (PostgreSQL) were used to process the data. ArcGIS is a geographic information system that can be used for creating, analyzing and discovering maps and geographic information in a geodatabase. There are many tools that can be used to perform various functions such as buffering and spatial joining.

PostgreSQL is a powerful, open-source, relational database system that can store, manage and process geospatial information. It supports PostGIS functions – an extension supporting spatial information data types (e.g. point, line string and space), and functions processing the spatial

information (e.g. buffer and calculating distance). All the spatial data types are stored as a batch of characters describing the features in the database, and can be easily accessed by PL/pgSQL using scripts (Momjian, 2001).

Bikeway network

As discussed in Section 3.2, the bikeway network we obtained from Toronto Open Data contains only centerline of roads (one-direction roads); however, alternative route generation (see Section 2.2) needs bidirectional links as well as nodes of the network. The nodes were extracted by the unique start and end points of links (ST_StratPoint and ST_EndPoint), and the bidirectional network was created by including both existing links and reversed links (ST_Reverse) in a new dataset. One-way streets were excluded from this process, as no link in the opposite direction was needed. In addition to generating two-way links, the distance of all the links was calculated in the dataset (ST_Length).

Digital elevation model (DEM)

The DEM obtained from the Geospatial Centre at the University of Waterloo was processed in ArcGIS by extracting the value of the slope and then adding it to the bikeway network layer. “Add surface information” is a tool in ArcGIS that can extract attributions from a surface data. Minimum, mean and maximum slope were measured in percentage unit (grade) along the line feature (ArcGIS Resource Center, 2016).

Transit stops

The transit stop dataset was obtained from Toronto Open Data as a text file containing location descriptions for all stops. This data was geocoded by an Excel geocoding tool (version 3.6) developed by Max Rice and Juice Analytics (2015) to a point layer. For the subsequent analysis, all the transit stops were then assigned to the nearest link by calculating the Euclidean distance between each bus stop and all the network edges using the ST_Distance_Sphere function in PostgreSQL. The number of transit stops on each link was appended to the link layer as an additional attribute. For example, the largest number of bus stops on a link (Steeles Ave East) is 55.

Parallel on-street parking area

Parallel parking areas are represented as polygons and the associated data were obtained from Toronto Open Data. A spatial join tool was used to identify the links in the bikeway network that have parallel parking. “Spatial Join” is a tool in ArcGIS that can connect two layers based on the spatial relationship (completely within) between them (ArcGIS Resource Center, 2016). The features in the bikeway network were joined to the parallel parking permitted areas, and the successfully joined ones have parallel parking as an appended attribute.

Annual average daily traffic volume

In addition, as discussed in Section 3.2, the annual daily traffic volume (AADT) dataset was acquired from Toronto Traffic Management Center, which contains AADT information for roads equipped with sensors. A new dataset containing the average AADT for each road type was

generalized from the original AADT dataset. The road classification management system in Toronto shows that for each road type, the AADT on all links are much similar. Therefore, for those links without equipped sensors, the value of AADT was assigned based on the average AADT of the same road type.

Final Integrated Network

The final integrated network represents all roads and bikeways in Toronto as bidirectional links with all relevant information appended to them as attributes. The integrated network dataset contains the name, road type, bicycle facility, grade (slope), the number of transit stops, parallel on-street parking status, and the road type averaged AADT information.

4.2 Map Matching

GPS observations were collected and organized as a point level dataset, containing second-by-second locations of the routes of all participating cyclists. In order to analyze the route choice behavior of participating cyclists, it was first necessary to convert point level GPS observations to a trip level sequence of links. This section describes the GPS data map-matching method, in which individual points were matched to links in the integrated network developed in Section 4.1.

Map-matching is an important topic in both transportation engineering and geomatics. The basic solution of map-matching is snapping the nearest edge to the observation. As discussed in Section 2.1.2, this method contains errors due to the inaccuracy of GPS observations and links (White, et al., 2000). An application of this method was first tried in this research; however, problems in the process for two types of points became apparent: points at intersections and parallel streets. Figure 4-2 highlights an example of this problem. In this figure, GPS observations, P1 to P4, could be matched to the vertical link. However, when considering the tendency of the whole trip, these points should instead be matched to the horizontal link. Similarly, as Toronto's network contains some parallel streets that are separated only by 20 to 50 meters between, when map-matching with the closest link, P1 to P4, the second half of the figure would be matched to the bottom link. However, when the entire route is considered, these points should be matched to the upper link.

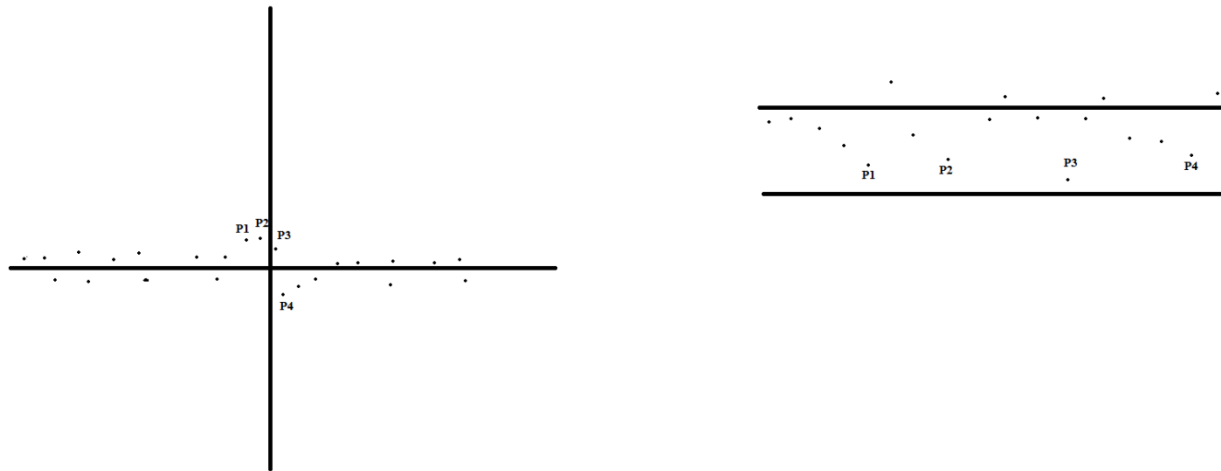


Figure 4-2 Map-matching Issues in Intersections and Parallel Streets

Therefore, a more sophisticated method that considers the overall route structure of individual trips should be applied. Considering the computational power and complexity of implementation, in this research, Strauss, et al. (2015)’s method was adopted. This method is similar to the method discussed previously, but uses buffers to do the map-matching in ArcGIS. To completely avoid the problem above, some additional modifications were also applied.

In this research, after considering the distances between parallel streets (around 20 to 50 meters) and the accuracy of smartphone GPS services (also around 20 meters), a 20-meter buffer was created for each link in the network, and only points inside the buffer were matched to the link as shown in Figure 4-4 (left).

The problems at intersections and closely paralleled streets were solved by identifying and grouping affected observations. A 20-meter buffer was created for each intersection, and observations inside the intersection buffers were considered as part of the “intersection” problem (Figure 4-4 middle). For closely paralleled streets with overlapping buffers, observations in more than one link buffer were considered as part of the “parallel streets” problem (Figure 4-4 right). Those two kinds of observations were identified as arbitrary observations requiring further analysis, while all other observations were considered as correctly map-matched GPS points. GPS records that could not be matched with any buffer were removed directly. These points were considered as being outside the study area or were assumed to contain large positioning errors.

In ArcGIS, the whole process was developed as a tool in toolbox as shown in Figure 4-3. Buffer and Spatial Join are two main tools used in ArcGIS to perform map-matching. “Buffer” creates circle-based buffer polygons for the input features with a specific Euclidean distance. “Spatial Join” can assign a link identifier from the network to GPS observations if the observations are completed within the buffer of the link. The end result is two point datasets, one containing the observations correctly map-matched, the other one containing the arbitrary unmatched observations.

The arbitrary observations were then imported into PostgreSQL, and assigned to a link based on a vote between the previous 10 and next 10 points by PLpgSQL scripts, as shown in Appendix II. After all GPS records have been matched to the network, trips can be generated following the structure of origin and destination point, departure and arrival time, route length, average speed and the sequence of links cyclists travel on.

Due to the GPS accuracy, some of the trips contain small gaps, as shown in Figure 4-5. These small gaps were connected manually or by shortest path algorithms. As the trip data comes from volunteers, some participants may record non-cycling trips while using the app. To mitigate this issue, trips with average speeds exceeding 36 km/h (or 10 m/s) were deleted. This value was chosen after reviewing relevant consulting resources from local cycling clubs. For example, Toronto's Cabbagetown Cycling Club groups cyclists into 5 groups based on riding skills. Cyclists in the group with the most experience (A* level) generally have average speeds of between 30 to 32 km/h (Cabbagetown Cycling Club, 2015).



Figure 4-3 Tool for Map-matching Processing in ArcGIS

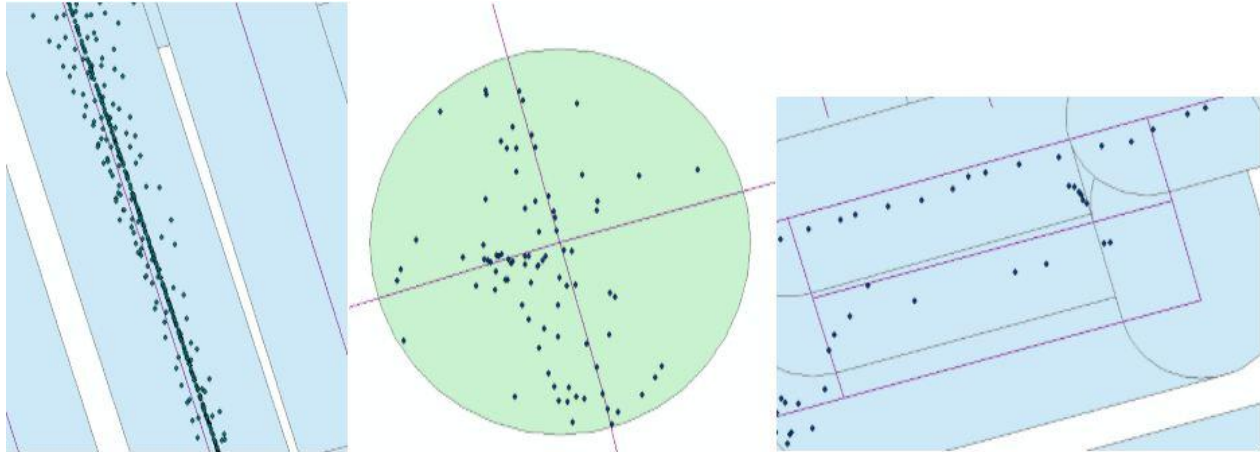


Figure 4-4 Buffers for Links (left), Intersections (middle) and Parallel Streets (right)

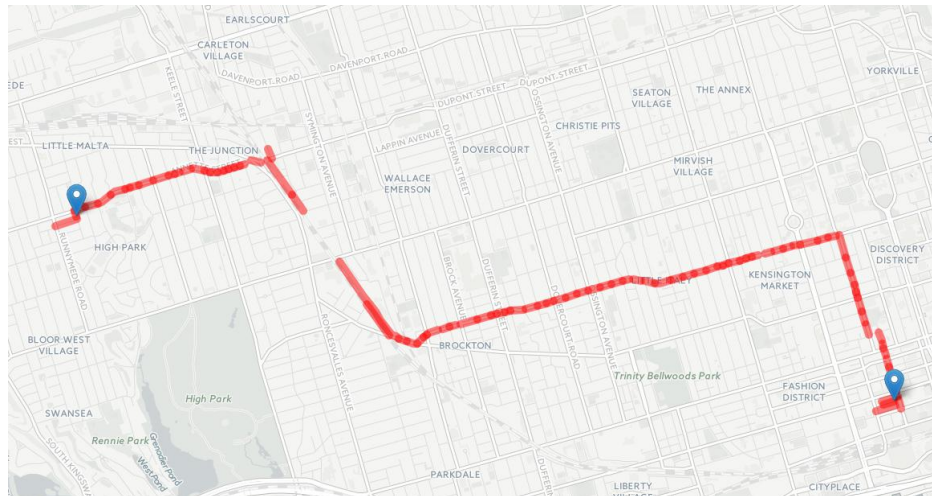


Figure 4-5 An Example of Routes with Gaps

All of the data pre-processing and map-matching steps were implemented using a combination of ArcGIS and PostgreSQL. Generally, ArcGIS was used for graphic processing operations, such as erasing points, creating buffers, spatial joining of GPS records with link buffers, selecting specific points by locations, merging two datasets, or adding surface slope information onto a feature layer while PostgreSQL, as a PostGIS enabled database system, was used for filtering and analyzing with SQL queries or by scripts performing self-programmed functions. In this study, building the bikeway network, map-matching and trip generation were all performed in PL/pgSQL in a PostgreSQL database. The detailed information about the tools in ArcGIS and PostGIS functions used in PostgreSQL are provided in Appendix I.

4.3 Exploratory Data Analysis

Recorded Trips

After the map-matching and trip generation processes, a total of 51,449 trips were identified, taking place between May 20th, 2014 and January 25th, 2015. Among these trips, 85.6% were

made between May to October. Cycling in winter is not as popular as in spring. Figure 4-6 shows the distribution of trips over different months. Very few (6.9%) trips happened in December and January.

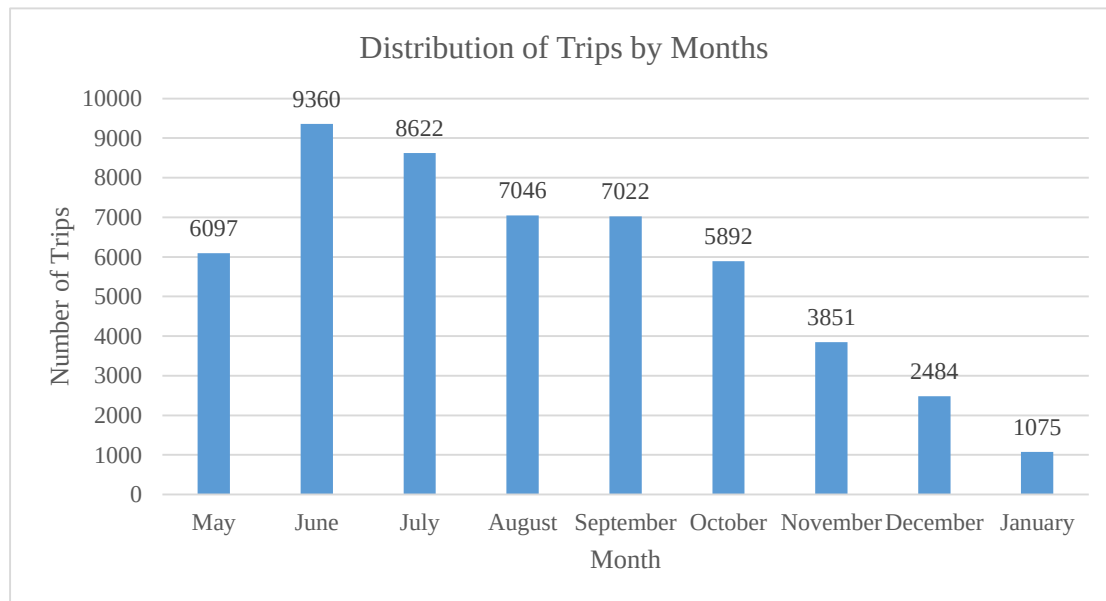


Figure 4-6 Distribution of Trips by Months

Ten types of purposes were recorded including commuting, shopping, social or leisure, exercise, etc. Most of the trips (65.11%) were for commuting. Around 2.75% of the trips were for exercise or sports, and 16.14% of the trips were for social or shopping. Other purposes (e.g. ON, school, business, home and others) comprised up to 16% of the total trips. The distribution of trips by purposes is shown in Figure 4-7 and Figure 4-8. In this study, only commuting trips and recreational trips were studied, with the later including exercise or sports, school, social or leisure, and shopping trips. Recreational trips consist of 18.9% of the sample versus commuting trips making up 65.11%.

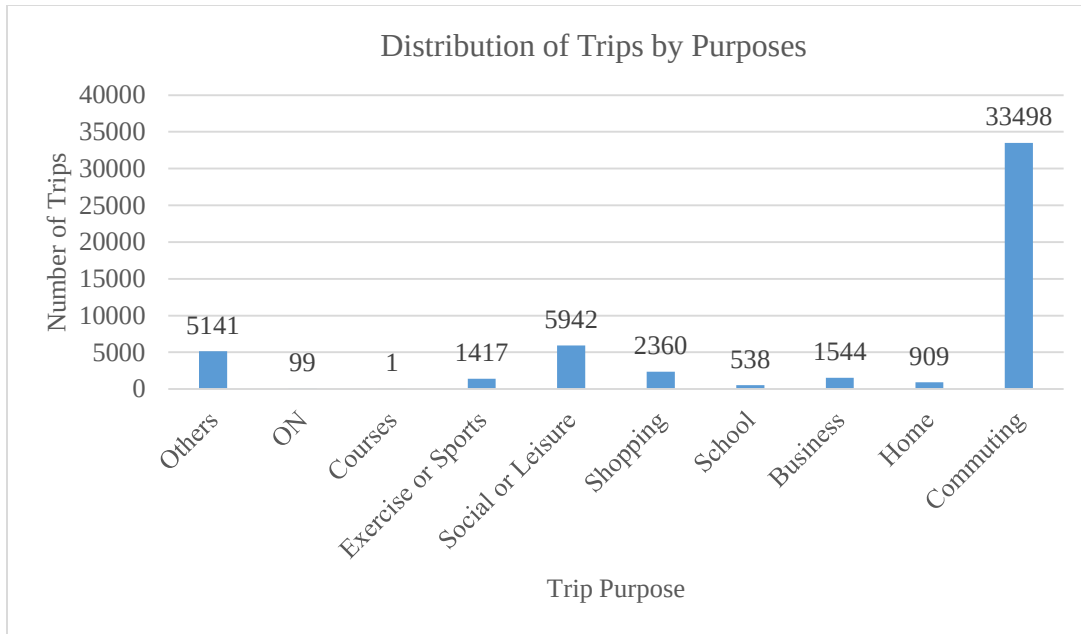


Figure 4-7 Distribution of Trips by Purposes

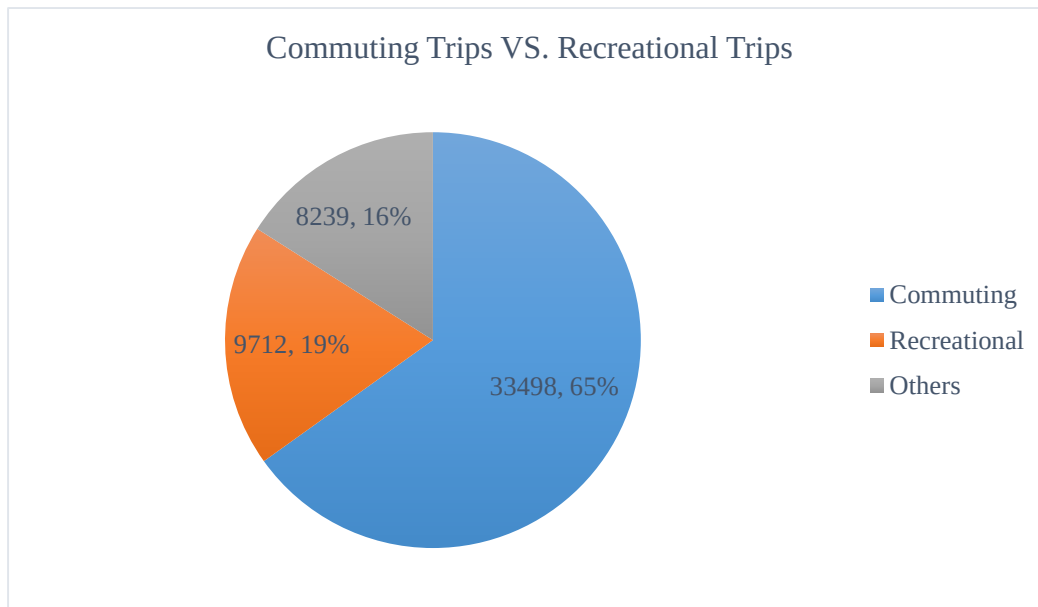


Figure 4-8 The Number of Commuting Trips VS. Recreational Trips

The distances of the trips were between 10 m to 102.5 km, and nearly half of the trips were between 2 km to 7 km. Figure 4-9 shows the number of trips in various ranges of distance. For comparison, 12.27% of the trips were between 6 km to 7 km, among which 79.18%, 4998 trips were commuting trips. For longer trips, nearly half of the trips were recreational trips when trip distance is more than 15 km.

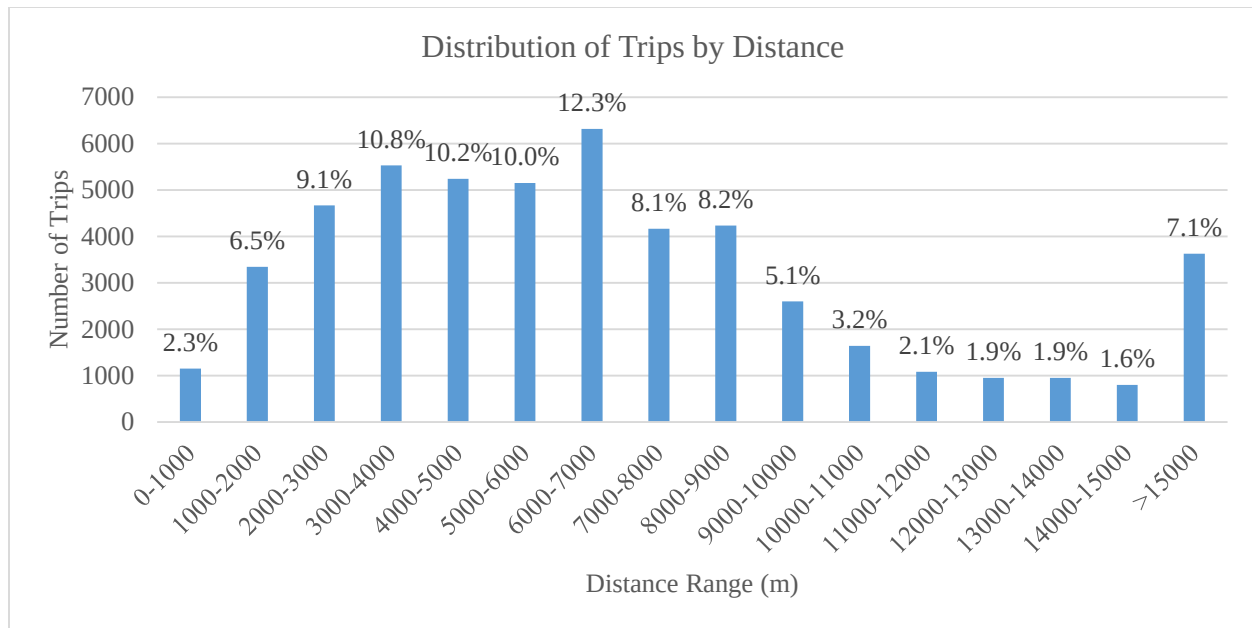


Figure 4-9 Distribution of Trips by Distance

The trip dataset included some extreme short trips or trips with unreasonably long detours (e.g. for a coffee or meal). These trips are not meaningful for route choice behavior modeling. As a result, the pre-preprocessing step introduced a filter to remove extremely short trips (< 500 m) and trips with unreasonable detour (distance is more than four times of the shortest distance). After filtering, a total of 33,222 commuting trips and 9611 recreational trips remained.

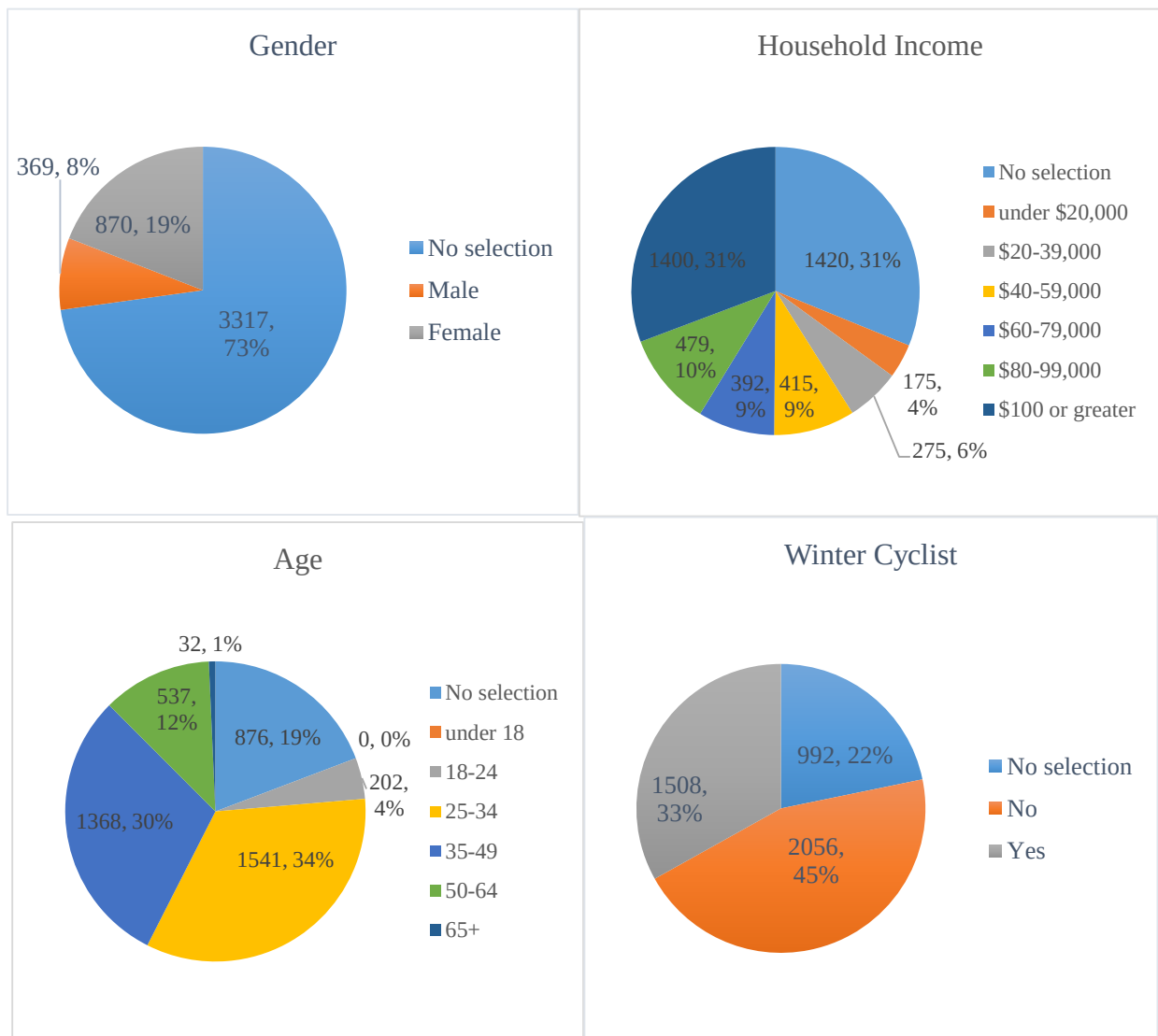
Participating Cyclists

A total of 4556 cyclists participated in this study. Some of them opted out from the personal information survey. Multiple trips produced by the same cyclists may bias the results, as the preferences of the cyclists making multiple trips may be correlated or over-represented in the analysis. However, as more than 20% of cyclists did not record their personal information, and over 70% of cyclists did not provide personal information, it was not possible to conclusively determine if a trip was made by the same person at all times. As a result, all trips were considered and treated as independent. Of the cyclists that provided gender information, 29.8% were male and 70.2% were female, which is different from literature. However, since 73% of cyclists did not have a selection in gender, it was difficult to draw any substantial conclusions based on the gender of cyclists in the City of Toronto. In most previous studies, males produced more and longer trips than females. However, in this study, females produced more than 20% of trips than male. This may be due to the missing gender data in personal information, which may contain a lot of male trip-markers.

In addition, around 44.6% of all cyclists who responded household income, come from households with more than \$100,000 or greater, although nearly 30% do not have a selection in household. This is different with the literature suggesting richer people have a higher possibility of choosing personal cars for travel. However, as cycling is viewed as a healthy activity, even people who can afford a car may choose to use a bicycle for commuting or recreational purposes.

Furthermore, of the cyclists that provided age information, 41.9% are between 25 and 34 years old, with the next most common age between 35 and 49 making up 37.2%. Over half (58.2%) of the participants reported that they have cycled since they were children, and approximately half (57.7%) of participants also reported that they do not cycle in the winter.

Bicycle facilities were confirmed to have effects on cyclists' mode choice as they have lower perceived risk as indicated in the literature. Besides the 22% of cyclists who do not have a selection on comfort level, 17.5% of cyclists are willing to cycle on roads with specific bicycle facilities but prefer not to share the road with motor vehicles, and 23% of cyclists feel uncomfortable to cycle on roads with motor vehicles even when bicycle facilities exist. Figure 4-10 shows the distributions of gender, household income, age, winter cycling, rider history, and comfort level among 4,556 cyclists.



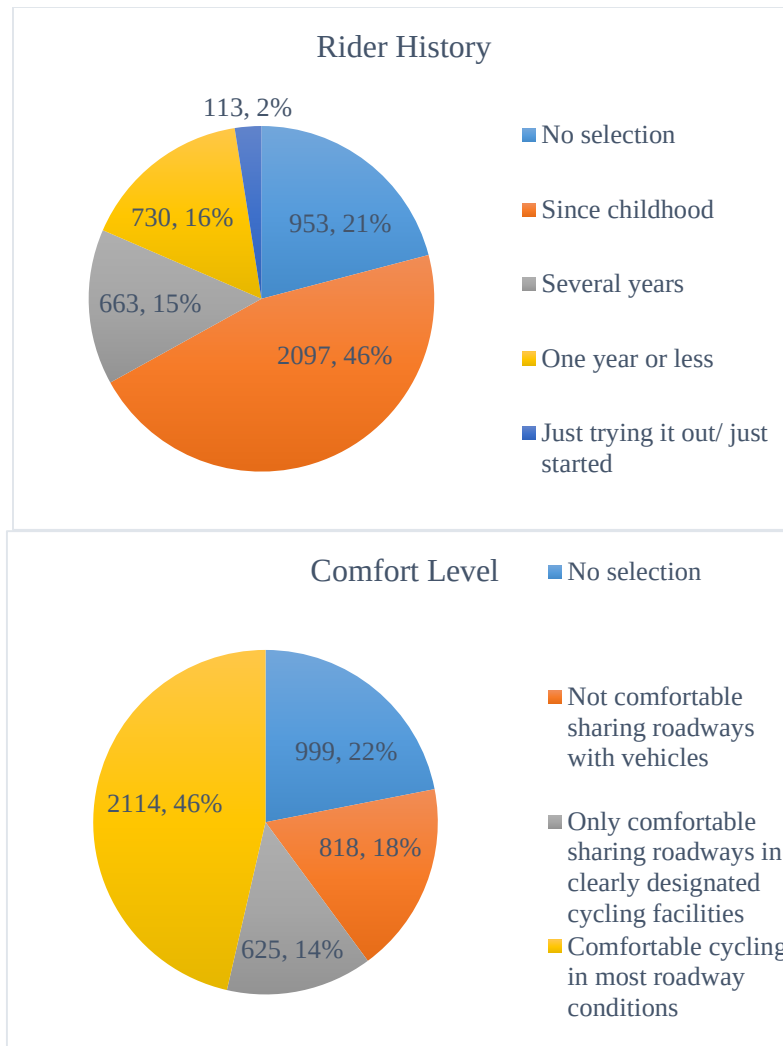


Figure 4-10 Distribution of Gender, Income, Age, Winter Cycling, Rider History and Comfort Level in Personal Characteristics

More detailed distributions are shown in Appendix II. According to the statistics summary, nearly half (44% and 45%) of the trips were produced by cyclists between 25 and 34 years old in both commuting and recreational trips. Around 41% of commuting trips and 13% of recreational trips were produced by cyclists that come from a household with more than \$100,000 income per year. More than half of the cyclists started riding since they were children. Those results are consistent with the literature.

Among 33,222 commuting trips, 28% of them produced by cyclists considered it difficult to cycle on roads without cycling facilities, and 24% of 9,611 recreational trips considered bicycle facilities necessary. As only 12% of existing network links have bicycle facilities, the government of Toronto needs to improve the quality of cycling networks to encourage more cyclists.

Network Characteristics

The final cycling network in the study area contains 14,285 km roads, 122,396 links, and 40,808 nodes. Roads in the network were originally divided into 12 road types based on the City of Toronto's classification system (e.g. arterial roads, collectors, local streets, busways, laneways, ramps, etc.). Cyclists are permitted to cycle on all types of roads except interchange-controlled expressways. Arterials, collectors, and local streets represent most of the network (64.88%); therefore, the 12 original road types were reclassified into 5 types: major arterials, minor arterials, collectors, local streets and others. Arterial roads are designed for relatively fast movement, and the AADT for arterial roads is usually between 8,000 to more than 20,000 depending on the level of road service (City of Toronto, 2016). Bicycle facilities are always desired by cyclists for the high traffic volume roadways. Collectors are similar to arterial roads, and are used for both car movements and accessing private property. The AADT of collectors is between 2,500 and 8,000. Rush hour peak periods may have more traffic volume than other times of day. Residential or local streets make up nearly 39% of the entire network, which is the largest value among all the road types. Arterial roads including major arterial and minor arterials only represent around 15% of the network. Table 4-1 shows descriptions, examples and basic statistical information for four types of roads, and Figure 4-11 shows the distribution of different types of roads.

Table 4-1 Description of Road Types

Road Type	Description	Example	Number of Links (Distance in km)
Major Arterial	The major function is car movement; AADT is greater than 20000; Specific bicycle facilities are desired.	 Yonge Street	11,769, 9.62% (1437.65, 10.06%)
Minor Arterial	The major function is car movement; AADT is between 8000 and 20000; Bicycle facilities are better designed.	 Gerrard Street East	7,049, 5.76% (820.29, 5.74%)
Collector	Provide access to property and traffic movement are two main functions; AADT is between 2500 to 8000; Specific bicycle facilities are required.		12,418, 10.15% (1534.53, 10.74%)

		Elmhurst Drive	
Local Streets	The main function is provide access to property; AADT is depends on the residences, and usually less than 2500; Bicycle facilities are desired.		48,169, 39.36% (6552.00, 45.86%)
		Hoshlega Drive	

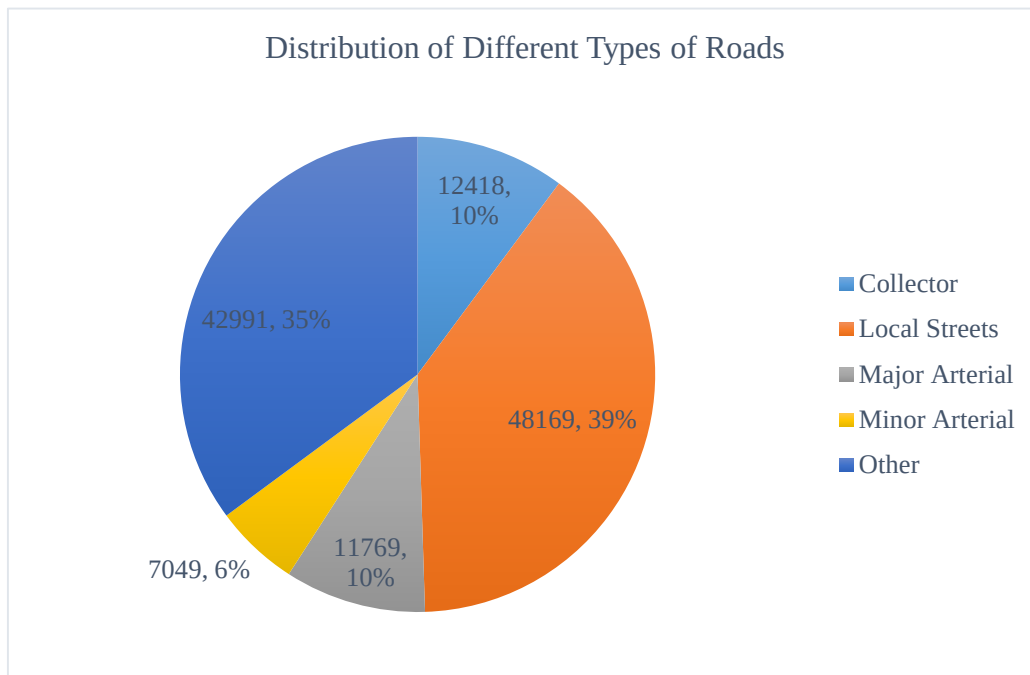
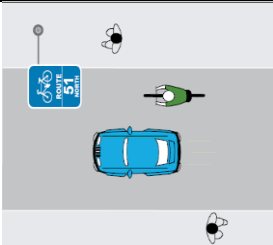
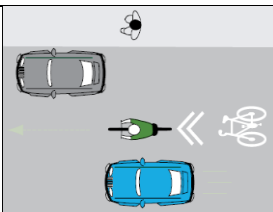
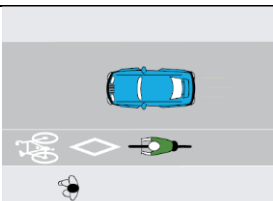
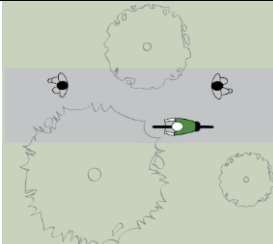


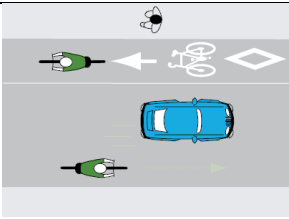
Figure 4-11 Distribution of Types of Edges

Although previous studies have indicated that cyclists strongly prefer high quality bicycle facilities, Toronto's bicycle network is not well developed, and only 12% existing network links have bicycle facilities. Toronto's network contains 10 types of facilities, including signed routes, sharrows, bike lanes, cycle tracks, etc. Detailed information on these bicycle facilities is presented in Table 4-2. Despite their different classifications, many of these bicycle facilities share similar characteristics. For example, a signed route and sharrow all indicate sharing roads with motor vehicles but the latter is different in that it has pavement markings positing where cyclists should cycle and advising motorists to share the road with cyclists. Cyclists have higher priority on shared lanes. Therefore, in order to generalize bicycle facilities and keep the most important features, in this study, the types of bicycle facilities were reclassified into 5 types: Bike Lanes (separate from motor traffic), Cycle Tracks (bicycle only), Multi-Use Trails (all non-motor users - bicycles, pedestrians, rollerblades), Shared Roadways (shared with motor vehicles but with high priority) and Others, as illustrated in Table 4-2.

Among all links with bicycle facilities, both shared lanes and bike lanes make up of around 17% of links with bicycle facilities corresponding to 16.58% and 13.67% of the total link distance. Multi-use pathway is one of the most common bicycle facilities, and makes up 34.26% of all links with bicycle facilities corresponding to 39.5% of the link distance. Figure 4-12 and 4-13 present the distributions of reclassified bicycle facilities in the entire network and the network with bicycle facilities.

Table 4-2 Description of Bicycle Facility Types

Reclassify Facility	Bicycle Facility	Description	Example	Number of Links (Distance in km)
Shared Lane 2421, 2.08% (274.92, 1.92%)	Signed Route	Designated by Bikeway Network Signage as preferred route of cycling; Cyclists share roads with motor vehicle; Usually found in residential streets.		2121, 1.82% (251.61, 1.76%)
	Sharrow	Cyclists share roads with motor vehicle; Marks show the position cyclists cycling, usually away from curb and parking cars.		300, 0.26% (23.31, 0.16%)
Bike Lane 2497, 2.15% (226.65, 1.59%)	Bike Lane	Dedicated space for cyclists. Motorists are not allowed to park, stand or drive in that area.		2497, 2.15% (226.65, 1.59%)
Multi-use Pathway 4776, 4.10% (654.77, 4.58%)	Multi-use Pathway	Cyclists share the space with pedestrians and in-line skaters; The area is physically separated from motorists; Sometimes called “off-road facility”.		4776, 4.10% (654.77, 4.58%)
Cycle Track 44, 0.04% (5.3, 0.04%)	Cycle Track	A physical separate bike lane for cycling only.		44, 0.04% (5.3, 0.04%)

Others 4203, 3.61% (496.26, 3.47%)	Suggested On-Street Route	Route suggested by cyclists but not in the cycling network yet; No indication of facility type.		2547, 2.19% (282.87, 1.98%)
	Contra-Flow Bike Lanes	Allow cyclists to travel in the opposite direction on one-way streets; Display as bike lane in the opposite direction; No automobile facilities in the opposite direction.		74, 0.06% (5.18, 0.04%)
	Suggested On-Street Connection	Connections between different types of road suggested by cyclists; Not sure whether has facility or not.		1196, 1.03% (142.98, 1.00%)
	Park Roads Cycling Connection	Cycling routes in park area; Always share with pedestrian and skaters; If the park allows motor vehicles, cyclists also share roads with motorists.		356, 0.31% (56.36, 0.39%)
	Informal Dirt Footpath	Footpath allows cyclists, pedestrian, and other modes except motor vehicle; The quality of road is not as good as multi-use pathway and cycle tracks.		30, 0.03% (8.87, 0.06%)
No Facility				102461, 88.02% (12627.98, 88.39%)

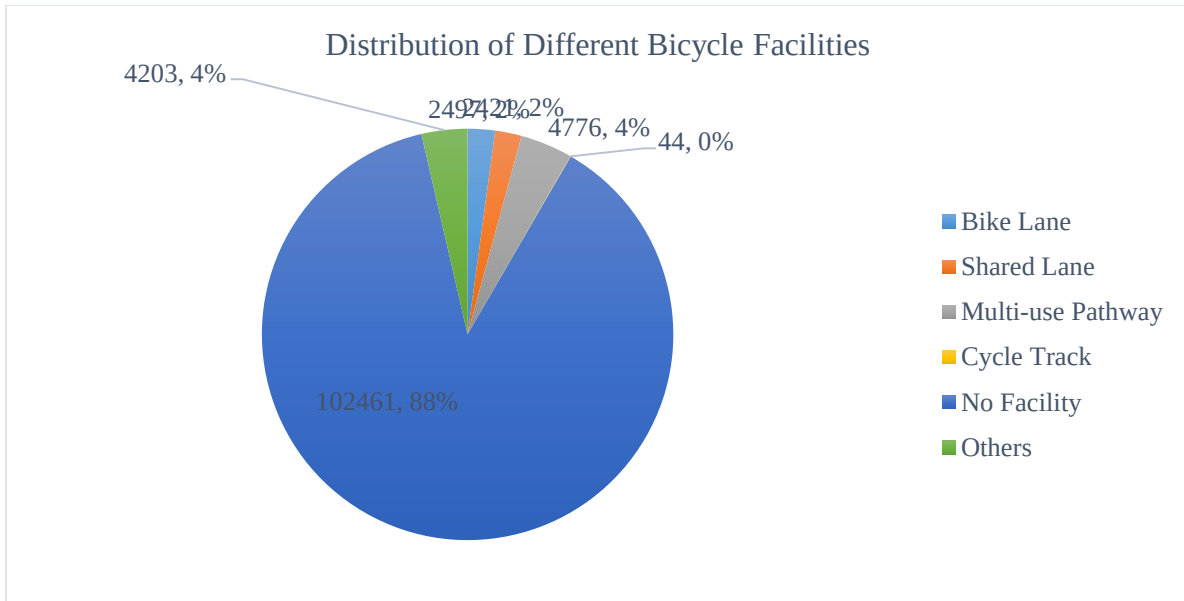


Figure 4-12 Distribution of Different Bicycle Facility Types

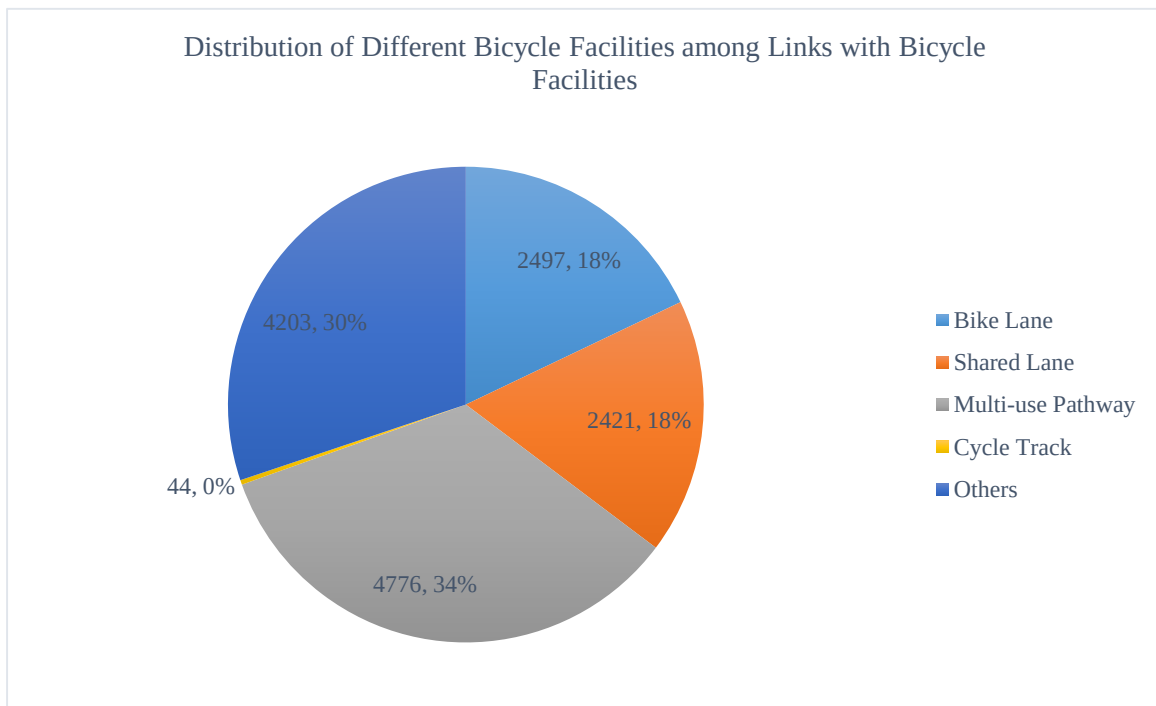


Figure 4-13 Distribution of Bicycle Facility Types Except No Facility

Finally, 10.65% of links in the network have bus stops, with Steeles Ave East having the largest number of bus stops (55 stops). 4.29 stops in average appear per kilometer for both commuting and recreational trips. Additionally, 22.65% of links with a total distance of 2,410 km allow on-street parking. Most of these links are local and arterial streets in the downtown area.

Chapter 5 Development of Route Choice Models

Developing route choice models has proven to be an efficient way to obtain a quantitative understanding about cyclists' route choice behavior (Sener et al., 2009; Stinson and Bhat, 2003). The approach described in this chapter is largely based on the fundamental concepts of discrete choice theory. In particular, the form of path-size logit model (PSL), which is a modified multinomial logit model (MNL), was selected as the modeling framework to address the issue of overlapping among alternative routes. A number of path-size logit route choice models were estimated for both commuting and recreational trips with various modeling options and combinations of candidate factors. The following sections provide details on the model estimations and evaluation of models as well as discussions about the results.

5.1 Discrete Choice Model - Framework

The route choices made by individual cyclists belong to the type of decision problems that can be best captured by a discrete choice modeling framework. Discrete choice models identify patterns in the choices made by individuals between various alternatives based on a few independent attributes. According to the process developed by Moshe Ben-Akiva and Steven Lerman (1985), discrete choice theory consists of four elements:

- **Decision-maker:** the discrete choice model is disaggregate in nature, considering the individual decisions made by each individual decision maker. Each decision maker is assumed to be characterized by his/her measurable socio-demographic characteristics such as age, gender, education and income.
- **Alternatives:** the discrete choice set contains a finite number of alternatives that can be explicitly listed. For example, the available traveling modes would form an alternative choice set. The choice set can also be the subset of universal alternative choices that are available to a particular individual based on an individual's situation or awareness of the alternatives. Therefore, identification of the choice set that each individual considers during his/her decision making process is an important part of discrete choice modeling.
- **Attributes:** attributes of all alternatives must be defined, which can be a directly measurable quantity or a function of other variables.
- **Decision rule:** it is assumed that decision-maker assesses the attractiveness of alternatives in terms of utility and always choose the option with the highest value of utility.

The route choice models in this research were developed to estimate the decision making process of participating cyclists, including their consideration of route related factors, personal attributes, and weather factors. The decision making rule was assumed following random utility theory.

Utility theory or random utility theory was first developed in the field of economics and has since been widely used in demand modeling and contains two principal assumptions. The first is that the attractiveness of a choice option can be quantified through a measure called "utility", and the second is that individuals always choose the options with the highest utility. The utility

function consists a deterministic part, which is a function of measureable factors, and a random part estimating the unmeasurable factors such as mood of the decision-makers, systematic and measurement errors, etc. As the random part in the utility function is hard to be measured, the probability that a given option is chosen is equal to the probability that the utility of the option is greater than utility values of all other options, which can be presented as follows (Ben-Akiva & Lerman, 1985).

$$P(A_i) = P(U_i > U_j) = P(V_i + \varepsilon_i > V_j + \varepsilon_j) \quad \text{for all other } A_j \in A \text{ except } A_i \quad (5-1)$$

where A_i = the i th option

A = the full set of options being considered

$P(A_i)$ = the probability that option i is chosen

U_i = the utility of the option i , containing two components, the deterministic part (V_i), usually defined as a linear function of all or subsets of the factors considered to have effects on route choice, and an error term (ε_i) associated with the option.

To determine the probability in Equation 5-1, the distribution of the random errors must be known or assumed. Utility models are classified based on their assumed error distribution; among these, the Probit and Logit models are the most common (Prato, 2009). This research employs the more popular Logit model, in which the error terms are assumed to be identically and independently distributed (IID), following the Gumbel Type I distribution:

$$P(\xi_i < x) = e^{-e^{-x}} \quad (5-2)$$

With this distribution, the following multinomial logit model (MNL) can be derived:

$$P(A_i) = \frac{e^{V_i}}{\sum_{A_j \in A} e^{V_j}} \quad (5-3)$$

The procedure of developing discrete choice models for route choice behavior analysis is divided into four steps: identifying the attributes affecting route choice behavior, designing the model specification, generating a reasonable alternative route choice set, and estimating the route choice behavior models (Path-size logit model). The following sections provide a detailed discussion on each of these four steps.

5.2 Attributes Affecting Route Choice Behavior

As the attribute set is one of the key elements in discrete choice theory, this section identifies all the attributes affecting route choices before model estimation. As discussed in Section 2.4, cyclists' route choice behavior is mainly affected by alternative route dependent attributes (e.g. trip distance, network characteristics) and generic trip attributes (e.g. cyclist's socio-demographic characteristics, trip purpose, weather features, time of the trip). Although many studies have examined these factors, most have considered only one type of factor. In this study, a combination of these types of attributes were chosen for analysis

Attributes of Alternative Routes

As discussed in Section 2.4, trip distance or travel time was the most important factor for route choice decisions. Many studies have shown that travel time plays a more important role in route choice decisions than trip distance (Bovy & Bradley, 1985); however, due to the lack of information about intersection signal timing and the speeds of cyclists, travel time was excluded in this research. Despite this, distance was assumed to have a high correlation with travel time, and therefore captures the effect of travel time on route choice behavior.

Besides trip distance, previous studies have shown that network factors related to perceived risk and comfort level also influence cyclists' decision making (Section 2.4). The network factors considered in this research consist of road type, bicycle facility type, the number of transit stops, annual average daily auto-vehicle traffic volume (AADT), the existence of parallel on-street parking, and the energy consumption. The dataset of Toronto's road network groups all links into the following four categories: major arterials, minor arterials, collectors, local streets and other roads, which is based on the links' functionality, average traffic volume and level of service. Many previous studies considered the road type was statistical significant for cyclists on their route decision making process.

The dataset also groups bicycle facilities into four categories, according to Toronto classification management system (City of Toronto, 2016), including shared lanes, bike lanes, multi-use pathways, cycle tracks, and other facilities. The principle of classifying bicycle facilities is based on their functionality and independence from motor road lanes. Multi-use pathways and cycle tracks are considered to be separate dedicated bicycle facilities, which were considered statistical significant in route decision making in the literature (Sener et al., 2009; Stinson & Bhat, 2003).

In addition to road and bicycle facility type, this study also examined the effect of transit stops on route choice behavior, as these were theorized to affect cyclist's perception of safety. Few of the previous studies included this factor in their analysis; however, as public transit system in Toronto is more advanced than that in other cities, it was included as an attribute to estimate the effect of public transit on cycling.

As discussed in the literature (Section 2.4), on-street parking is another important network factor, which may increase cyclists' perceived risk (Winters et al., 2012). Roads with on-street parking, especially parallel on-street parking, are narrower, and cyclists must travel together with motor vehicles close to the center line if there is no dedicated bicycle facilities. Therefore, parallel on-street parking was identified as one of the key variables in this research.

Finally, energy consumption caused by gradient and distance has been shown to affect cyclists comfort level and, consequently, their route choice behaviour (Menghini et al., 2009; Hood et al., 2001). In this research, energy consumption estimates a cyclist's ability to overcome rolling resistance, air resistance and uphill gradient. This study considers the effect of energy consumption using an equation developed by Olds et al. (1993) from biomechanical properties, which has been used in various power monitoring devices (e.g., SRM and Power Tap) (Froster et al., 1994). The energy cost of cycling is defined as:

$$E_{tot} = E_{Rr} + E_{Ra} + E_{grade}$$

$$= C_{Rr} \cos(\arctan(S)) (M + M_b)d + E_{Ra} + (M + M_b)g \sin(\arctan(S_{uphill}))d \quad (5-5)$$

where E_{Rr} = the energy cost in overcoming rolling resistant (J)

E_{Ra} = the energy cost in overcoming air resistant (J), which is assumed to be constant in the context of this research (the average is 0.19 J)

E_{grade} = the energy cost in overcoming uphill grade (J)

C_{Rr} = the factor related with tire pressure and tread, wheel radius, and road surface. The default value is 0.0457

S = the slope of the road section

d = distance (m)

M = the mass of the rider

M_b = the mass of the bicycle (kg) (the average is 10.3 kg)

The energy cost for overcoming rolling resistant and air resistant were considered constant using the average value in Olds et al.'s (1993) research. While the mass of the rider is affected by the gender. In this research, females were considered as 74.4 kg and males were assumed with a weight of 80.3 kg (Spitzer et al., 1999). The other cyclists did not indicate their gender were assumed with a weight of 77.4 kg.

In addition to the alternative route dependent factors, which vary among different alternatives, some generic trip attributes also affect cyclists' route choice decisions. The generic trip attributes are attributes with the same value among all the alternatives, which in this research can be divided into trip purpose (commuting trips and recreational trips), demographic characteristics (e.g. gender, age and rider history), weather characteristics (e.g. temperature) and time of the trip (month) in this research.

Trip purpose

Previous studies have confirmed the different main factors that influence cyclists' route choice decisions when they make commuting and recreational trips (Broach et al., 2012). In addition, different kinds of cyclists may have different sensitivity to the same factors (Chen & Chen, 2010). In order to estimate the differences between commuting and recreational route choices, in this research, the trip purpose was considered as one of the key generic trip attributes and both of commuting and recreational trips would be put into estimation. However, in this research, there are a total of 33,222 commuting trips, while only 9,611 recreational trips were recorded. The big difference between the sample sizes may cause some counter-intuitive results for recreational trip analysis.

Demographic characteristics

Another part of generic trip attributes is socio-demographic characteristics. In the literature, various results were concluded in different studies due to the variety of samples and study areas. Some studies concluded personal information not substantial in route choices (Abraham et al., 2002; Hood et al., 2001), while other studies determined slight influences of socio-demographic characteristics on route choice behavior such as gender, age, etc. (Krizek, 2006; Sener et al., 2009). In this study, although a lot of socio-demographic information was collected such as household income, comfort level, etc., only age, gender and rider history were selected as variables affecting route choice decisions. This is because a large number of cyclists opted out from the other factors, and other factors such as household income may influence the mode choices instead of cycling route choice.

Weather characteristics

Cycling, unlike driving, is an outside activity and can therefore be substantially affected by weather, including temperature, rain precipitation and snow amounts. However, as few trips were recording during snow and rain events, in this research only temperature was considered.

Month

The month the trip was produced in is similar in function to the weather characteristics. For the collected data, December and January are part of the winter season and have lower temperatures than other months. October and November are part of autumn and have moderate temperatures, while all other months are considered as spring and summer months, which have higher temperatures and the most number of cycling trips. As spring in Canada is short and the differences between spring and summer for cyclists are not obvious, so in this research, these two seasons were grouped together. Many previous studies did not consider the effect of weather characteristics and the month of the trip, because they only collected trip data over a small period of time. However, in this study, the larger scale of trip data collected makes estimation of the effect of weather characteristics and month of the trip on route choice behavior possible and necessary.

Table 5-1 describes all alternative routes dependent and generic trip attributes that were proposed to be involved in route choice models in this research.

Table 5-1 Attributes of Route Choice Models

Type	Attribution	Description
Alternative Routes Dependent Attributes	Travel Distance	The total length of a trip in km.
	Energy Consumption	The total energy cost of the trip calculated by Equation 19.
	Number of Stops	The total number of bus stops during the entire trip.
	Average Traffic Volume	The average volume of the trip. If there are live records of volume on the link, extract the live volume in the time period. Otherwise, used the average volume of the road type.
	Proportion of Major	The proportion of route distance occurring on major

	Arterial		arterial. It is between 0 and 1.
	Proportion of Minor Arterial		The proportion of route distance occurring on minor arterial. It is between 0 and 1.
	Proportion of Collector		The proportion of route distance occurring on collector. It is between 0 and 1.
	Proportion of Local		The proportion of route distance occurring on local streets. It is between 0 and 1.
	Proportion of Shared Lane		The proportion of route distance occurring on links with shared lane as bicycle facility. It is between 0 and 1.
	Proportion of Bike Lane		The proportion of route distance occurring on links with bike lane as bicycle facility. It is between 0 and 1.
	Proportion of Multiuse Pathway		The proportion of route distance occurring on links with multi-use pathway as bicycle facility. It is between 0 and 1.
	Proportion of Cycle Tracks		The proportion of route distance occurring on cycle track links. It is between 0 and 1.
	Proportion of No Bicycle Facility		The proportion of route distance occurring on links without bicycle facility. It is between 0 and 1.
	Proportion of Parallel Parking		The proportion of route distance occurring on links with parallel parking. It is between 0 and 1.
Generic Trip Attributes	Mean Temperature (degree)		The mean temperature of the day of trip produced.
	Month		The month when the trip was produced.
	Trip Purposes		The purposes of each trip (e.g. commuting trips and recreational trips)
	Demographic Factors	Age	Classification of age into 6 groups (see Section 3.2).
		Rider History	Classification of history of cycling into 5 groups (see Section 3.2).
		Gender	3 groups including no selection, male and female.

Correlation analysis

The inclusion of highly correlated variables could decrease the significance of the variables and performance of the entire model (Ben-Akiva, M. E., & Lerman, S. R., 1985). It could also lead to biased estimates of the model coefficients and thus the magnitude of the effect of the included variables. Correlation between the alternative route dependant factors such as energy consumption and trip distance, is expected to some degree as trip distance is one of the major inputs to the energy consumption function (Equation 19). Another example is between traffic volume and road type. The classification of roads in Toronto considers average traffic volume as one of key factors along with number of lanes, speed limit, and road surface conditions.

Therefore, before estimating route choice models, correlations among all the factors were checked. Appendix IV shows the results of correlation analysis for both commuting and recreational trips, and key correlations are discussed here.

The results confirmed the suspicion of high correlation between trip distance and energy consumption for both commuting (0.897) and recreational trips (0.979). However, trip distance and energy consumption still need to be estimated separately, as energy consumption is also related with road grade, which has been shown to have an influence on route choice decisions in the literature.

In addition, trip distance or energy consumption has a moderate level of correlation with the number of transit stops (around 0.67). This is caused by the fact that longer trips have a higher possibility to pass more transit stops.

Furthermore, the annual average daily traffic volume (AADT) has a low correlation with various road types except major arterial (between 0.1 and 0.65). The high correlation between AADT and major arterial streets (around 0.85) indicates traffic volume is a major factor in defining different types of arterial roads. However, the correlations between AADT and other road types are low, suggesting that traffic volume is not a good surrogate for road types, and they should be considered separately in the model.

Finally, the correlations between each type of bicycle facilities and other variables are between 0.06 and 0.6, and the correlations related to personal and weather factors are even lower with correlation coefficients ranging from 0.0004 to 0.25. Bicycle facilities, personal characteristics and weather factors are therefore confirmed as independent variables.

5.3 Model Estimation

In the previous section, attributes affecting route choice behavior have been identified, and their values were calculated in advance for each trip. Estimation of discrete choice models generally involves model specification, alternative set generation, and model estimation, which are discussed in this section. Since the number of feasible routes for each trip from an origin to a destination is extremely large, a finite subset of alternatives was identified based on some specific assumptions on the perceptions of cyclists. The discrete route choice model was built on random utility theory (logit model), and Path-size logit model was selected as modeling framework to estimate the effects of identified attributions on route choice decision as well as obtain the benefit of avoiding the issue caused by violating the independence from irrelevant alternatives (IIA) requirement.

5.3.1 Model Specification

As discussed previously, the decision rule applied to measure the probability of an alternative in the choice set being chosen is based on utility theory, specifically, random utility maximization. The utility term is partitioned into a deterministic part and a random error. The deterministic part of utility function is generally assumed to be a linear function of the contributing factors discussed in the previous section. For the problem in this research, all the factors listed in Section 5.1.1 are considered as generic variables as they are assumed to have the same coefficients for all

alternatives. For an alternative route i in the route choice set for a given, the deterministic part of the utility function is therefore defined as:

$$V_i = b_1 * D_i + b_2 * \frac{E_i}{1000} + b_3 * T_i + b_4 * \frac{AADT_i}{1000} + b_5 * MA_i + b_6 * MI_i + b_7 * C_i + b_8 * L_i + b_9 * SL_i + b_{10} * BL_i + b_{11} * MP_i + b_{12} * CT_i + b_{13} * NF_i + b_{14} * PP_i \quad (5-6)$$

where D_i = the distance of route i (km) or the logarithmic form of distance (km)

E_i = the energy consumption of route i (kj)

T_i = the number of transit stops of route i

$AADT_i$ = the annual average daily traffic volumes of route i (veh/day)

MA_i = the proportion of major arterial of route i

MI_i = the proportion of minor arterial of route i

C_i = the proportion of collector of route i

L_i = the proportion of local streets of route i

SL_i = the proportion of shared lane of route i

BL_i = the proportion of bike lane of route i

MP_i = the proportion of multiuse pathway of route i

CT_i = the proportion of cycle tracks of route i

NF_i = the proportion of streets without bicycle facility of route i

PP_i = the proportion of streets with parallel on-street parking of route i

b_{1-14} = the coefficients that determine the weights of individual attributes and are estimated using maximum likelihood method.

In addition to trip distance and energy consumption, network features such as road type and cycling facilities are associated with each link. The proportion of route distance occurring on those network features was chosen to represent all distance-based attributes for each route. This value is a decimal that ranges from 0 to 1, with 1 indicating that the route's entire distance covers that attribute's type and 0 indicating that the attribute is not a portion of that trip. The effects of the generic trip attributes would be estimated by comparing various models based on different sample trips.

As discussed in the previous section, trip distance has a high correlation with energy consumption, and traffic volume (AADT) is highly correlated with major arterial and multi-use pathway. In addition, as shown in Section 5.4.1, most trips happened in the central part of town

where parallel on-street parking is allowed. The extreme high proportion of parallel on-street parking of chosen routes may also cause bias in model results. Therefore, in this study, different forms of linear utility functions with different combinations of factors were tried to find the best-fit utility functions.

5.3.2 Generation of Alternative Routes

The alternative set defined in the discrete choice modeling framework should include a finite number of options that the decision makers consider to choose. For a route choice problem, identifying all the possible routes from an origin to a destination is challenging if not impossible due to the complexity and extent of the underlying network. Also, not all the possible routes are in fact being considered as alternatives by cyclists. For example, commuting cyclists would most likely not consider routes with excessively large detours from the shortest route as alternative. Therefore, the choice set should contain the routes that the cyclists are likely to consider before making a decision.

The basic idea of generating an alternative route set is based on the assumption that cyclists would consider routes that meet their specific route choice criteria such as minimizing travel distance, travel time, total risk or total energy consumption (Menghini et al., 2009; Rieser-Schüssler et al., 2012), as discussed in Section 2.2.1. As discussed in Section 2.2.2, two types of alternative route generation methods were commonly employed in the literature: stochastic choice set generation and deterministic choice set generation. This research adopted the deterministic choice set generation method. Furthermore, travel distance and energy consumption were considered as the main measures in generating alternative routes, assuming that all most likely candidate routes should be the ones with the lowest travel distance or energy consumption. To improve the chance of identifying the candidate routes that cyclist would actually consider, each criterion is used to identify multiple routes using an iterative path search process. Specifically, alternatives were generated by iteratively applying a shortest path search algorithm.

To avoid excessive overlapping among alternative routes, after each iteration, a link penalty multiplier of 1.5 (50% increase in travel distance or energy consumption) was applied to each link that was already chosen in the previous iteration, similar to Usyukoy (2013). If a link was selected in multiple iterations, the multiplier would be applied as a penalty every time it was selected. Note that this penalty was applied to the links on the routes that were chosen by the cyclists at the start of the process. Other multipliers were also tested and it was found that a value of 1.5 yielded routes with relatively low overlapping (from 0.7% to 30%), as discussed in Section 5.4.1. The final alternative route set contains 10 shortest distance paths, 10 minimal energy consumption routes plus the observed route.

The shortest path search in each iteration was done using Dijkstra's algorithm. Application of Dijkstra's algorithm involves exploring all paths to the various intersections of the network and keeping track of the minimum cost to get there from the origin. The route with the minimum cost is then selected out as the shortest path. Most of the shortest path generation functions in geospatial software is based on this algorithm such as ArcGIS and MATSim. In this research, the Dijkstra's algorithm was implemented in C# as shown in Appendix II.

5.3.3 Independence from Irrelevant Alternatives (IIA)

A type of Logit model, multinomial Path-size logit model, was applied in this research as the route choice model. This model was proposed to address one of the most important limitations of the traditional multinomial logit model, that is, the requirement of the Independence from Irrelevant Alternatives (IIA), which states that the ratio of the probabilities of choosing two alternatives for each individual is independent of the availability or attributes of any other alternatives (Ben-Akiva & Bierlaire, 1999; Hood et al., 2013; Menghini et al., 2009; Ramming, 2001).

Fundamentally, this means that the alternatives considered in the model should be purely discrete and alternatives should always have the same cross-elasticity. However, in route choice problems, taste variations, unobserved factors, measurement errors, relationships between observed factors to unobserved factors, and overlapping between alternatives can all lead to choice sets that may violate the IIA property.

Another effect of the IIA property statement says that a one-percent change in a variable of an alternative has the same impact on utility values of other alternatives, which means alternatives should always have the same cross-elasticity. If the alternatives are not distinct, the utility change of one alternative may influence other alternatives in different ways, depending on the cross-elasticity. The conflict is very common in real situations. For example, an improvement on public transit may cause some travellers to shift from carpooling, but may have less effect on personal cars. That is because the cross-elasticity of public transit to carpooling is greater than that to personal cars.

In route choice problems, although the alternative routes are different, there could be overlapping links between the alternatives. If the overlapping among alternatives is too large, the probabilities of choosing those alternatives would counter-intuitively increase due to the overlapping parts being considered more than once. The overlapping in alternatives violates the assumption of independence from irrelevant alternatives (IIA) property of MNL, and the probability of routes with more overlapping would be over-estimated (Broach et al., 2012).

From the perspective of route choice, previous studies have mainly approached the IIA problem in two ways. One approach minimizes IIA during the alternative routes generation, while the other extends the MNL framework to consider the effect path overlapping. In this research, we first applied a method to minimize overlapping in the alternative route generation process, as discussed in the previous section. A link penalty multiplier was also applied to minimize overlapping by increasing the generalized cost of the chosen links so that the selected links have a lower possibility of being selected again in the next iteration of the shortest path algorithm. Secondly, we applied Path-size logit model, a modified version of the multinomial logit model, to further reduce the effect of path overlapping, as detailed in the following section.

5.3.4 Path-Size Logit Models

As discussed previously, route overlapping violates the IIA property of the logit model, which may cause a counter-initiative increase of the probability of choosing routes with large overlapping. To address this overlapping issue, alternative model forms have been proposed in

the literature with the key idea of incorporating a correction factor, such as commonality factor and path size factor, in the utility function of multinomial logit model (MNL). Those factors approximately describe the amount of overlapping between each route and the other alternatives.

In this research, the Path-size logit (PSL) model was selected as the modeling framework, as it has been shown to have better performance than the C-logit model in the literature (Prato & Bekhor, 2007; Ramming, 2001).

As discussed in Section 2.3, the Path-size logit (PSL) model introduces a path size factor into the utility function, which was specified as:

$$PS_i^{rs} = \sum_{a \in \Gamma_i} \frac{l_a}{L_i^{rs}} * \frac{1}{\sum_{j \in C^{rs}} (\frac{L_i^{rs}}{L_j^{rs}})^\gamma \delta_{aj}} \quad (5-7)$$

where Γ_i is the set of links in alternative i , l_a is the length of link a , L_i^{rs} is the length of route i from origin r to destination s and δ_{aj} is 1 if route j includes link a and 0 if not (Frejinger & Blerlaire, 2008). γ was set to 0 in this research as Broach et al. (2012) suggested, as the alternative routes do not contain unreasonable detours. Then, the probability of choosing option i in a Path-size logit model (PSL) was defined as:

$$P(A_i) = \frac{e^{V_i + b \cdot \ln(PS_i^{rs})}}{\sum_{A_j \in A} e^{V_j + b \cdot \ln(PS_j^{rs})}} \quad (5-8)$$

where A is the alternative route choice set from origin r to destination s , and A_i is one of the routes in alternative route choice set. The value of the path-size factor (PS) is between 0 and 1, with a higher value indicating a lower level of overlapping with other alternatives. The effect of the path size factor on the utility should be positive, as it is affected by the natural logarithm of the path size factor (Bliemer & Bovy, 2009).

As other MNL, the Path-size logit model can be estimated using Maximum Likelihood Method, which estimates coefficients in the utility functions that are most likely to produce the observed distribution of choices for the sample (Ben-Akiva et al., 1985).

Many statistical software packages can be used to estimate multinomial logit (MNL) models. This study used Biogeme, which is an open source software tool for the maximum likelihood estimation of parametric models in general, with a special emphasis on discrete choice models (Bierlaire, 2003).

In this research, a number of alternative Path-size logit models were estimated with various modeling options and combinations of factors to find the effects of attributes discussed previously and the best-fit model form. Multiples models were also estimated based on different sample trips to check the sensitivity of the size of alternative route set (10 vs. 20) and the effects of generic trip attributes such as trip distance range (< 5km; 5-10 km; > 10 km), socio-demographic characteristics, trip purposes (commuting trips and recreational trips), weather features (temperature lower and higher than 5 degrees) and months (summer, autumn and winter periods).

The following alternative settings are of substantial importance:

- 1) Inclusion versus exclusion of trip distance and traffic volume, which have higher correlations with other variables
- 2) Inclusion versus exclusion of parallel parking fractions for commute trips: parallel on-street parking in this dataset showed a higher trip proportion in chosen routes, as shown in Section 5.3.2. Therefore, in case parallel on-street parking influence other factors, model excluding parallel on-street parking should be estimated.
- 3) Distance versus its logarithmic form ($\ln(\text{distance})$): the effect of trip distance has been found to be dominant as compared to other factors. As a result, some researchers (e.g., Broach, et al., 2012) have suggested using logarithmic form of distance, which may lead to improved model estimations.
- 4) Size of alternative route set for each trip;
- 5) Grouping of trips by trip distance
- 6) Grouping of trips by socio-demographic attributes (e.g. age, gender, rider history)
- 7) Grouping of trips by weather attributes (e.g. mean temperature)
- 8) Grouping of trips by trip purposes: commuting trips versus recreational trips.

5.4 Results and Discussions of Route Choice Models

The development of route choice model follows the steps of model specification, alternative path generation and model estimation. This section describes the results and discussions about alternative Path-size models based on the modeling options listed in previous section. The alternative route choice set is presented on the map and the effects of different types of factors are concluded from the model results.

5.4.1 Alternative Route Choice Set

The alternative route choice set was generated by a combination of path labeling and link penalty (Section 5.2.1). The size of alternative route choice set was 20 with 10 shortest distance paths generated using link length as the label, and 10 minimal energy consumption paths using energy consumption as the label. Figure 5-1 shows an example of a typical trip as well as its alternative routes. The red line shows the route used by the cyclists, which was determined by a map-matching algorithm; the bold blue and dark pink are the shortest path and most energy saving path, respectively.

This trip was made on June 4th, 2016 by a cyclist of between 25 to 34 years old, who had indicated he had cycled since childhood. The route is around 9 km long and passes a large number of transit stops – a total of 67 stops. The high number of transit stops is likely due to the fact that the trip occurred in the downtown area with dense transit coverage. Approximately 60% of the route travelled on major arterials, while only 7% travelled on local streets. Shared lanes

and bike lanes were the two main bicycle facilities along the route, among which 24% of the route occurred on links with bike lanes. However, due to the lack of bicycle facilities in Toronto's road network, most of the route (around 70%) occurred on links without any bicycle facility. This trip happened in the downtown area, where parallel street parking is allowed on most of the streets; 98% of the trip passed by links with parallel street parking.

As shown in the figure, the shortest path is a more direct route than the minimal energy consumption route and the observed route, while the energy saving path contains more turns to avoid uphill streets. As cyclists seldom make a long detour from an origin to a destination for saving energy, the overlapping between minimal energy consumption path and the chosen route is much smaller than that between shortest path and the chosen route. The distance of the shortest path is around 7 kilometers, of which 682.68 meters (approximately 10%) overlaps with the chosen one. In contrast, only 4.2% of the chosen route is overlapped with the minimal energy cost path.

The shortest path in this example has a 30% higher proportion of major roads than the chosen route. In addition, the proportion of shortest path without bicycle facilities is slightly higher than that of the chosen route. This is because most of the major roads in Toronto do not have any bicycle facilities due to the limited space. Shortest paths with higher proportion of major streets have lower proportion of bicycle facilities; thus, the proportion of the shortest path on streets with bike lanes decreases 3% when compared to the chosen route.

The minimal energy consumption path has longer distance (8.6 km); however, its energy consumption (1,520,300 J) is much lower than those of the shortest path (2,050,599 J) and the chosen route (2,225,804 J). The proportion of minimal energy consumption routes occurring on local streets (47.8%) is higher than both the chosen route and the shortest path (nearly 40% higher) due to the flatter road surface of the local streets. As most of the local streets in the network do not have any bicycle facilities due to their narrow width and low traffic volume, the proportion of links with bike lane tends to decrease on paths with higher proportions of local streets.

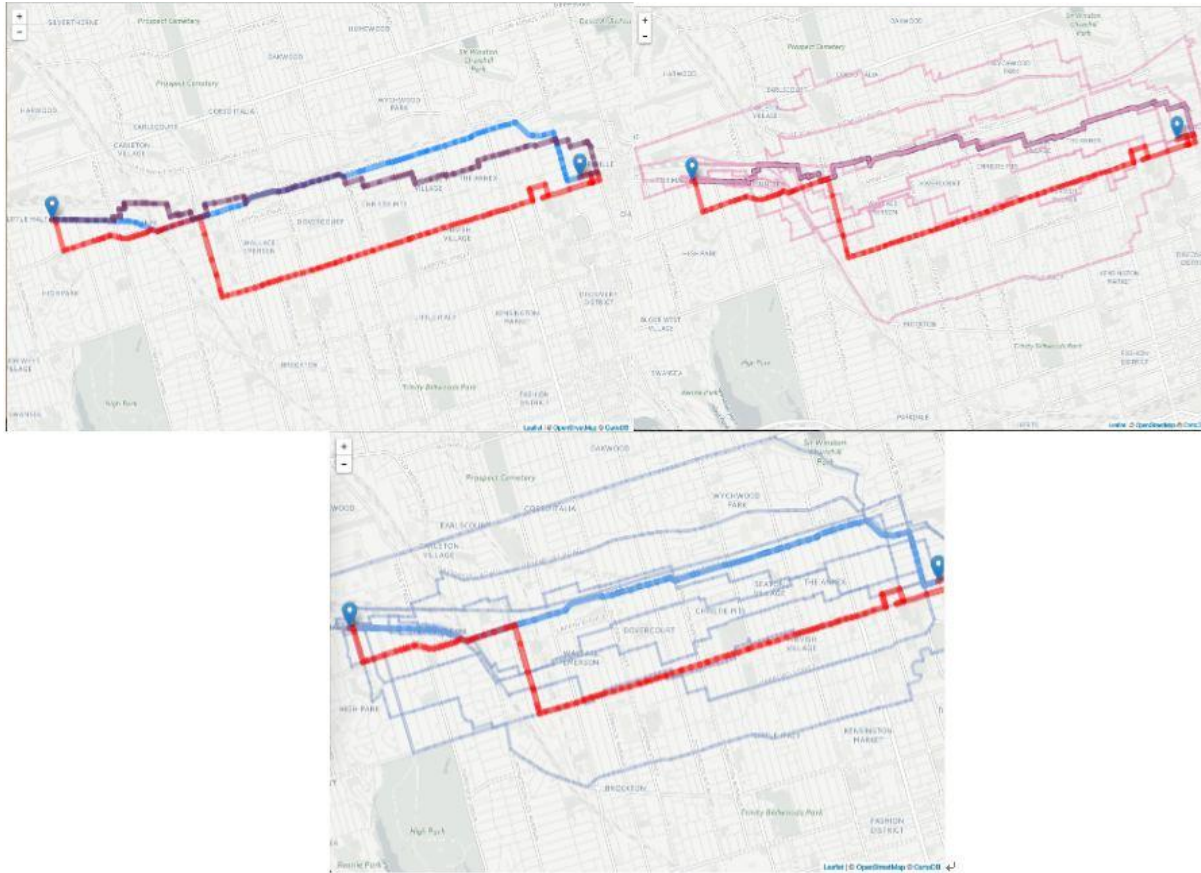


Figure 5-1 An Example of Observed Route and Alternatives

(left-top: observed route vs. shortest path vs. most energy saving path; right-top: observed route vs. shortest path vs. 10 shortest distance alternatives; bottom: observed route vs. most energy saving path vs. 10 energy saving alternatives)

Table 5-2 Detailed Information of the Sample Trip

Properties	Average Value for All Trips (Commuting Trips vs. Recreational Trips)	Observed Chosen Route	Shortest Path	Minimal Energy Consumption Path
Route ID	-	9727		
Date	-	2014-06-04		
User Age	-	25-34		
User Rider History	-	Since Childhood		
User Gender	-	No Selection		
Average Temperature	-	18.2		
Total Precipitation	-	0		
Total Snow	-	0		
Distance (m)	7,945 / 9,786	8,844	7,316	8,619
Energy Cost (J)	2,399,746 / 3,163,628	2,225,804	2,050,599	1,520,300
Number of Stops	33.853 / 33.244	67	59	24
Average Traffic Volume (AADT)	8,811 / 8,169	11,208	13,467	7,689

Proportion of Major Arterial	0.341 / 0.297	0.628	0.912	0.321
Proportion of Minor Arterial	0.252 / 0.213	0.161	0.066	0.010
Proportion of Collector	0.129 / 0.127	0.137	0.022	0.039
Proportion of Local Street	0.156 / 0.183	0.074	0	0.478
Proportion of Shared Lane	0.075 / 0.066	0.044	0	0
Proportion of Bike Lane	0.196 / 0.154	0.240	0.216	0.068
Proportion of Multi-use Pathway	0.070 / 0.102	0	0	0
Proportion of Cycle Track	0.017 / 0.011	0	0	0
Proportion of Roads Without Bicycle Facility	0.568 / 0.584	0.702	0.784	0.921
Proportion of Parallel Parking Streets	0.544 / 0.561	0.986	1	1
Overlapping Distance with Observed Route	-	-	7.72%	4.25%

Table 5-2 also describes the average values of all the attributes for both commuting and recreational trips. It can be seen that the values of the attributes are similar between commuting and recreational trips, while recreational trips have a higher value of average distance and energy consumption and lower values for both AADT and proportion of major streets (e.g., major, minor arterial, collector). Also, the routes used for recreational trips included more multi-use pathways (dedicated off-street bicycle facility) and local streets. In general, for recreational trips, cyclists tend to take routes that are longer but with less traffic. This makes intuitive sense as recreational trips are mostly for exercise and sightseeing.

Chosen routes vs. shortest distance path vs. minimal energy consumption path

For over 30,000 commuting trips covered in this study, the average overlapping percentage between observed routes and shortest paths is 28.67%, which means nearly one third of each chosen route is identical to the corresponding shortest path. This is consistent with the observations from the literature that many chosen routes are identical to or similar to the shortest path or close to the shortest path (Aultman-Hall et al., 1997). In addition, the average overlapping percentage between observed commuting routes and most energy saving routes (around 16.85%) is slightly lower than that between shortest paths. As energy consumption is a function of both distance and grades, this result suggests that cyclists tended to avoid roads with slopes, but place a higher consideration on travel distance.

For 9,611 recreational trips, the overlapping percentage between shortest path and observed routes is 24.59%, which is lower than that for commuting trips. The overlapping percentage between shortest path and minimal energy consumption path is slightly higher than that for commuting trips with a value over 17%. As the purpose of recreational trips is doing exercise or social events, cyclists usually consider energy consumption more important than distance.

In general, the shortest paths for all trips are more direct than the observed routes with the average proportion of major streets with high volumes in the shortest routes being much higher. The route with the minimal energy consumption is usually longer than the shortest path, but saves more energy.

Overlapping analysis among alternatives

The final alternative choice set excludes the path with the shortest distance and the path with the minimum energy consumption as they have a substantial large overlapping with the observed path. Exclusion of overlapping paths is important to reduce the IIA limitation (discussed in Section 5.2.3). Instead, the choice set contains the next 10 additional shortest distance paths and next 10 additional minimum energy consumption paths that were generated by successive iterations of the algorithm described earlier. As each additional iteration applies a penalty on that link's distance value, successive iterations result in larger detours. A total of 10 paths were considered to balance the need to include relevant choices while excluding paths that were unreasonably long (average 15% detour for commuting trips).

As shown in the figure (the 10 alternative shortest distance paths are shown in light pink, and the 10 alternative energy saving paths), overlapping among the 20 alternatives paths and the chosen routes still exists. The overlapping analysis between each alternative in the choice set was checked.

The overlapping was measured by the percentage ratio of common trip distance between two routes to the trip distance of the target route. The percentage results of overlapping among alternative routes for both commuting and recreational trips are low (ranging between 0.7% and 30%), suggesting the alternative route generation method used is effective. The results also show a higher average overlapping percentage in alternatives for recreational purposes (around 12.19%) than commuting (around 7.89%). One of the reasons for this may be the long distance of recreational trips, and the limited route options for cycling activities. Furthermore, the energy saving alternative routes have higher average overlapping with other routes comparing with shortest distance paths for both commuting (9.26% vs. 6.64%) and recreation (13.73% vs. 10.78%). The overlapping between shortest distance path and minimal energy consumption path is the highest among all the values. This is due to the fact that the road surface of Toronto is relatively flat when compared to other cities, so the difference between energy consumption and distance is relative small, especially for long trips.

Table 5-3 presents the overlapping analysis with the observed route as target route, showing the average overlapping percentage between each alternative and the observed route for both commuting and recreational trips. It can be seen that the overlapping between each alternative and observed route is low (between 2.94% and 7.61%). The energy saving routes have a higher value of overlapping percentage with observed route than shortest distance routes due to the consideration of grade when choosing paths. The alternative routes for recreational trips are closer to the observed routes than commuting trips. The complete overlapping analysis results estimating the percentage of overlapping between each two alternatives are specified in Appendix VI.

**Table 5-3 Overlapping Percentages for Alternative Routes and Observed Routes for
Commuting and Recreational Trips**

Alternatives	Commuting Trip\Recreational Trips (%)
Observed Route	1\1
Shortest Route 1	2.94\3.67
Shortest Route 2	4.01\5.17
Shortest Route 3	4.65\5.23
Shortest Route 4	5.07\5.59
Shortest Route 5	5.04\5.88
Shortest Route 6	5.37\5.96
Shortest Route 7	5.58\5.82
Shortest Route 8	5.74\6.02
Shortest Route 9	5.54\5.85
Shortest Route 10	5.8\5.92
Energy Saving Route 1	6.8\7.14
Energy Saving Route 2	7.58\7.65
Energy Saving Route 3	7.61\7.25
Energy Saving Route 4	7.18\7.36
Energy Saving Route 5	7.01\7.10
Energy Saving Route 6	6.97\6.94
Energy Saving Route 7	6.9\6.88
Energy Saving Route 8	6.85\6.71
Energy Saving Route 9	6.78\6.56
Energy Saving Route 10	6.78\6.56

5.4.2 Modeling Results

A number of Path size logit models were estimated using Biogeme based on 33,222 commuting trips and 9,611 recreational trips following various alternative settings. Table 5-4 (PSL1 to PSL5, MNL1) shows the estimation results of the commuting route choice behavior models, and Table 5-5 (PSL6 to PSL8) shows the estimation results of the recreational route choice behavior models based on the same scenarios as commuting route choice models except decreasing the size of alternative choice set from 20 alternatives (10 shortest distance routes, 10 minimal energy consumption routes) to 10 shortest distance routes. Table 5-6 (PSL9, 10, 11) presents the estimation results of models with different market segmentations by commuting trip length.

Table 5-4 Commuting Trip Route Choice Model Estimations for Different Scenarios

	PSL1 (n = 20)		MNL1 (n = 20)		PSL2 (n = 20)		PSL3 (n = 20)		PSL4 (n = 10)		PSL5 (n = 20)	
Variables	Coeff.	t-test	Coeff.	t-test	Coeff.	t-test	Coeff.	t-test	Coeff.	t-test	Coeff.	t-test
Distance (km)	-	-	-	-	-	-	-	-	-	-	-0.159	-20.26
LN(Distance (km))	-3.61	-50.16	0.154	2.53	-3.60	-49.96	-	-	-5.14	-45.62	-	-
Energy Consumption (1000s kj)	-0.151	-7.82	0.834	49.63	-0.142	-7.30	-0.301	-16.67	0.033	1.36	-0.196	-10.05
Number of Transit Stops	0.01	15.94	0.013	21.45	0.009 83	15.50	0.0010 1	1.64	0.012 2	17.57	0.004 79	7.61
Annual average daily traffic volumes (1000s veh/day)	-0.205	-12.00	-0.183	-11.01	-0.211	-12.28	-	-	-0.117	-6.80	-0.223	-13.09
Trip proportion (TP) with major arterial	17.8	112.62	19.3	125.82	17.8	112.67	19.4	131.25	15.9	94.06	19.0	119.5 3
TP with minor arterial	16.8	102.85	18.7	116.13	16.7	102.19	19.6	126.76	15.8	91.74	17.9	109.8 3
TP with collector	16.5	90.17	18.7	103.06	16.5	89.89	19.3	117.15	15.6	81.07	17.3	94.90
TP with local street	11.6	54.61	13.6	64.28	11.5	53.36	14.4	88.23	11.3	52.02	12.0	56.90
TP with shared lane	0.389	3.35	0.816	7.67	0.369	3.17	0.317	2.74	0.357	2.84	0.341	2.95
TP with bike lane	0.337	3.30	0.87	9.36	0.327	3.20	0.0379	0.37	0.523	4.69	0.174	1.72
TP with multiuse trail	17.9	68.23	19.3	74.03	17.9	68.05	21.3	107.63	17.3	64.48	18.3	70.12
TP with cycle track	2.57	11.68	4.63	21.75	2.49	11.19	2.60	11.90	3.20	12.69	2.54	11.72
TP without bicycle facilities	-3.00	-36.11	-2.89	-38.23	-3.02	-36.19	-3.09	-37.15	-2.53	-27.67	-3.04	-36.71
TP with parallel parking	-	-	-	-	187	3.14	-	-	-	-	-	-
Ln(Path Size Factor)	6.11	112.83	-	-	6.11	112.79	5.03	106.18	14.1	115.0 6	5.19	106.8 7
Number of Observations	33222		33222		33222		33222		33222		33222	
Final log-likelihood	-37698.968		-46594.145		-37694.020		-39151.393		-28383.455		-38857.988	
Rho-Square	0.627		0.539		0.627		0.613		0.644		0.616	

Table 5-5 Recreational Trips Route Choice Model Estimations for Different Scenarios

Variables	PSL6 (n = 20)		PSL7 (n = 20)		PSL8 (n = 20)	
	Coeff.	t-test	Coeff.	t-test	Coeff.	t-test
Distance (km)	-	-	0.235	14.89	-	-
LN(Distance (km))	1.57	16.44	-	-	-	-
Energy Consumption (1000s kj)	1.79	38.05	1.70	34.75	2.03	44.74
Number of Transit Stops	0.0261	14.81	0.0279	15.61	0.0356	21.36
Annual average daily traffic volumes (1000s veh/day)	-0.0759	-3.12	-0.0691	-2.84	-	-
Trip proportion (TP) with major arterial	13.9	59.45	13.7	59.01	12.9	58.58
TP with minor arterial	13.2	49.84	13.1	49.09	12.7	55.04
TP with collector	12.5	43.60	12.4	43.03	12.3	52.37
TP with local street	8.91	26.43	8.91	26.21	9.23	39.86
TP with shared lane	0.534	2.73	0.597	3.08	0.645	3.37
TP with bike lane	1.34	7.46	1.40	7.88	1.47	8.37
TP with multiuse trail	13.6	31.83	13.6	31.57	14.1	45.95
TP with cycle track	1.62	3.98	1.60	3.95	1.55	3.83
TP without bicycle facilities	-1.79	-12.69	-1.73	-12.34	-1.72	-12.49
Ln(Path Size Factor)	1.12	16.03	1.60	26.56	1.72	29.15
Number of Observations	9611		9611		9611	
Final log-likelihood	-10910.998		-10947.884		-11058.247	
Rho-Square	0.627		0.626		0.622	

Table 5-6 Model Estimations with Different Market Segmentations by Trip Length

	< 5 km (PSL9)		5 – 10 km (PSL10)		>10 km (PSL11)	
Variables	Coeff.	t-test	Coeff.	t-test	Coeff.	t-test
LN(Distance (km))	-6.08	-47.81	-2.64	-23.74	2.43	11.61
Energy Consumption (1000s KJ)	-2.84	-30.36	-1.62	-40.30	-0.236	-6.62
Number of Transit Stops	0.00169	0.66	0.0117	10.27	0.00984	7.68
Annual average daily traffic volumes (1000s veh/day)	-0.183	-8.93	-0.306	-10.20	-0.133	-1.86
Trip proportion (TP) with major arterial	12.9	56.96	20.8	76.58	23.0	36.89
TP with minor arterial	12.4	49.94	19.2	68.90	21.0	41.32
TP with collector	10.4	38.40	20.7	64.75	21.0	36.36
TP with local street	7.79	25.31	13.2	36.41	11.7	18.23
TP with shared lane	-1.32	-6.87	1.39	7.59	0.689	1.40
TP with bike lane	-0.0876	-0.58	0.363	2.08	0.372	0.88
TP with multiuse trail	8.89	21.76	18.8	40.93	25.6	30.62
TP with cycle track	-0.0734	-0.24	5.25	13.25	2.54	3.20
TP without bicycle facilities	-2.67	-23.54	-2.48	-16.46	-6.63	-19.26
Ln(Path Size Factor)	6.91	81.33	9.12	86.36	9.87	35.37
Number of Observations	8829		17049		7344	
Final log-likelihood	-12500.696		-15916.375		-3769.820	
Rho-Square	0.535		0.693		0.831	

5.4.3 Evaluation of Competing Model Structures

The alternative models are evaluated using log likelihood and Rho-square (ρ^2) values. It was found that ρ^2 values for all models are around 0.6, which indicates that the models performed well. The model considering a lower number of alternative routes ($n=10$) was found to have a slightly higher value of ρ^2 (0.644 vs. 0.627), but the alternatives only include 10 shortest distance routes generated by minimizing trip length; other factors were not considered, and the coefficient of energy consumption is positive, which is different than other commuting choice models using 20 alternatives (Table 5-4: PSL4 and PSL1). The positive coefficient does not make intuitive sense, as indicated in the literature (Hood et al., 2001; Menghini et al., 2009). The results with a higher number of alternative routes ($n=20$) (PSL1) were found to be more stable, yielding results that make more intuitive sense and more comparable with those from literature.

Transformation of distance into a logarithmic form improved the model fitting with an increase in ρ^2 value of 0.011 in commuting route choice models (PSL5 vs. PSL1) and 0.001 in recreational route choice models (PSL7 vs. PSL6). The use of the logarithmically transformed distance had reduced the bias caused by the large value of distance when compared to other variables. The improved model performance with a logarithmic form of distance suggests that relative route deviations are what cyclists most care about.

The coefficient associated with parallel on-street parking was unreasonably high (PSL2). This is because most of the chosen routes were within the central area of the city, which has a high proportion of roads with parallel on-street parking, thus leading to the high positive effect of parallel parking. Although adding a variable in the utility function makes the value of log-likelihood increase slightly, the unreasonably high coefficient for parallel parking does not improve the model performance; therefore, this variable was excluded from the models in later analysis.

In addition, Path-size logit models excluding trip distance and annual average daily traffic volume (AADT) were also performed to check the effects of those two variables on the model performance, as they have relatively high correlations with other variables (PSL3). Comparisons between model subsets were done according to the likelihood ratio test, which defines a test statistic according to the following equation:

$$LR = 2 * (\log L_A - \log L_B); \quad (5-9)$$

where Model B is a subset of Model A. The result of this test statistic is then compared to a value from a chi-squared distribution table for k degrees of freedom at a significance level of α , where k is the number of additional variables in model A. If the LR value is less than this statistic, then we can conclude that the variables should be excluded from the utility function; otherwise, Model A with all the variables should be used.

The likelihood ratio test statistic of deleting trip distance and annual average daily traffic volume (AADT) equals 294.5, which is much larger than the critical likelihood ratio value with 2 degrees of freedom (0.103). This means that trip distance and annual average traffic volume (AADT) should be included in both commuting and recreational route choice models.

In order to check potential non-linear effects of travel distance on route choice different route choice models were developed for different types of trips as defined by trip length. As shown in Table 5-6, the modeling results for trips of different distance ranges (except the short trips - those with distance less than 5 km) are similar to the one considering all trips together (PSL1). For the short trips, the effect of bike lanes, shared lanes and cycle tracks do not seem reasonable, as they showed negative effects on route choice. This may be due to the fact that shorter trips happened in neighbourhood areas where multiuse pathways are substantially more common than bike lanes and shared lanes. Furthermore, when the trip length is more than 10 km, the effect of distance becomes positive, and the sensitivity of energy consumption is not as obvious as it is in short trips. As more than 50% of the trips are between 5 - 10 km, the model for all trips performed similarly to the one using trips between 5 - 10 km.

Once a choice model is estimated, the resulting model could be interpreted qualitatively on the basis of the signs of the model coefficients, as discussed previously, and quantitatively on the basis of the relative magnitude of the coefficients. One of the quantitative interpretation analysis that are often performed in route choice modeling is the distance trade-off analysis, which shows the effect of a given factor (e.g., proportion of bicycle lane) on cyclists' choice decision with respect to distance. For this analysis, a measure called marginal rate of substitution is determined for a factor of interest based on the model estimations, which represents the perceived equivalent increase in distance for one unit change in that factor. In other words, this analysis estimated how much additional distance a cyclist is willing to accept for avoiding or obtaining specific attributions. Because of the use of the natural logarithmic transportation of distance, the marginal rates can be conveniently determined using Equation 5-10.

$$\text{Marginal Rate of Attribute } i = \left(\exp \left(\Delta \text{attribute} * \frac{b_{\text{attributes}}}{b_{\ln(\text{distance})}} \right) - 1 \right) * 100 \quad (5-10)$$

where $b_{\text{attributes}}$ is the coefficient of an attribute, and $b_{\ln(\text{distance})}$ is the coefficient of the logarithmic transformed of distance.

Table 5-7 presents the estimation results of marginal rates for both commuting and recreational trips.

Table 5-7 Marginal Rates of Substitution with Respect to the Natural Log of Distance

Attribute	Distance value for commuting trips (% dist.)	Distance value for recreational trips (% dist.)
Energy Consumption (1000s KJ)	4.272	212.72
Number of transit stops	-0.277	1.68
Annual average daily traffic volumes (1000s veh/day)	5.843	-4.720
Trip proportion (TP) with major arterial (1%)	-4.811	9.257
TP with minor arterial (1%)	-4.547	8.771
TP with collector (1%)	-4.468	8.287
TP with local street (1%)	-3.162	5.839

TP with shared lane (1%)	-0.108	0.341
TP with bike lane (1%)	-0.093	0.857
TP with multiuse trail (1%)	-4.838	9.049
TP with cycle track (1%)	-0.709	1.037
TP without bicycle facilities (1%)	0.834	-1.134

5.4.4 Effects of Related Factors

The coefficients of the models show the effects of factors on route choice decisions. In this research, this section describes the effects of alternative related characteristics and generic trip attributes including socio-demographic characteristics, trip purposes, weather (temperature) and month when the trip produced on route choices.

5.4.4.1 Effects of Alternative Route Dependent Characteristics

The model evaluation concludes that the PSL with 20 alternative routes (plus the chosen one) and logarithmic form of distance based on various characteristics including trip distance and network characteristics (e.g., number of transit stops, road types, bicycle facility types, AADT and energy consumption) shows the best performance in all models with various sample trips. A comparison of the sign and magnitude of the coefficients presents their relative effects. Although the values of the coefficients vary between different models and model settings, the direction of the effect of each variable and the magnitude of the effects were found to be similar, which confirms the stability of the modeling results.

Trip distance

As expected, the results of the models suggest that commuting cyclists prefer shorter routes, when all other attributes are held equal. A 10% increase in distance would decrease the utility of choosing the route by around 0.34 under commuting purpose. This is consistent with the observed data. Over half of the observed commuting trips were less than 20% longer than the shortest routes, and nearly 91% of the trips were less than 50% longer. On the other hand, cyclists hold diverse attitudes towards trip distance for recreational trips. The positive coefficients of trip distance suggest the preferences of longer routes in recreational route choices.

Energy consumption

Similarly, as energy consumption is a function related to both distance and slope, commuting cyclists also have a negative attitude towards routes requiring high energy consumption, which means they prefer not to travel a long distance on uphill roads. However, the model results show that, when compared to distance, cyclist commuters are less sensitive to energy consumption, with 1 KJ of additional energy consumed (e.g. from a hill) perceived as equivalent to a distance increase of 4.27% (Table 5-7). The preference on shorter and energy saving routes is more obvious in shorter trips (Table 5-6), as the sensitivity of trip distance and energy consumption is in proportion to the minimum distance from the origin to the destination. In contrast, the positive coefficients of energy consumption in recreational route choice models suggest that recreational cyclists perceived energy consumption as benefits, and 1 KJ of additional energy consumed is

equivalent to the exercise done by double distance. This result is consistent with the literature showing recreational cyclists are not sensitive to distance, but prefers moderate uphill (Antonakos, 1992; Broach et al., 2012).

Network characteristics related to perceived risk

The results also show the statistical significant effects of network characteristics related to perceived risk of cyclists on a route's desirability, which include the number of transit stops, annual average daily traffic volume (AADT), various types of roads and bicycle facilities. All of the network attributes were confirmed to have statistical significant effects on both commute and recreational route choice models with all kinds of generic trip attributes.

The coefficient estimated for the effect of the number of transit stops was found to be small in the models for both purposes with values of 0.01 for commute route choice model and 0.0261 for recreational route choice model. One additional transit stop only results in a 0.01 (commute route choice model) or 0.026 (recreational route choice model) increase in the utility of choosing the route, and is equivalent to a 0.2% (commute route choice model) or 1.68% (recreational route choice model) change in trip distance. The positive effects of transit stops confirmed the assumption that cyclists prefer to follow the routes of transit vehicles, which may be considered to be safer. As transit routes are often designed to fully cover residential areas, the possibility of a recreational route passing by more transit stops is higher than that for commuting trips because commuting trips consider directness the most important factor among all variables. However, these effects are not as obvious as other variables, and even lower in shorter trips. This means that although cyclists prefer routes with transit stops, they do not consider transit stops as a strong benefit for route choice behavior.

As expected, automobile traffic volume was confirmed to negatively affect the probability of a route being chosen, as shown by the models given in Table 5-4 (PSL1) and 5-5 (PSL6). Regardless of which bicycle facility the link has, more motorized vehicles with high speeds may make the route more dangerous, or at least make cyclists feel unsafe. This finding is consistent with the result in the literature (Howard & Burns, 2001; Menghini et al., 2009; Parkin et al., 2007). Cyclist commuters are more likely to make a 5.8% longer detour to avoid routes with high traffic volume. As recreational cyclists take benefits from long distance, the high traffic volume decreases the benefits by 4.72%.

Furthermore, the models show a preference for arterial roads followed by collectors and local streets. Cyclists were found to care about directness more than other factors. The coefficients of minor arterials and collectors are quite similar, suggesting there is no significant difference between these two types of roads, which is consistent with the classification system of Toronto. As local streets are the least preferred (coefficients from 7.79 to 13.2), a 1% trip proportion changes from major arterials to local streets would cause a decrease in the probability of choosing the route by 1.65% for commuting trips and 3.42% for recreational trips. Choosing local streets often involves long detours, lower speeds, and more bicycles, skaters, and pedestrians, which conflicts with the goal of travel in general: finding the fastest way to arrive the destination. The different preferences for various road types is more obvious for commuting trips and longer trips. For longer trips, the majority of the trips occurring on the major streets

instead of local streets to obtain higher speed and direct routes, while for recreational trips, cyclists do not care about directness in comparison to commuting trips.

Four types of bicycle facilities were included in the models: bike lanes, shared lanes, multi-use pathways and cycle tracks. The models, in addition, show an obvious preference for bicycle facilities when compared to links without bicycle facilities. Dedicated, off-street bicycle facilities (multi-use pathways and cycle tracks), separating the motor and non-motor traffic, are much more preferred than on-street facilities (bike lane and shared lane). This is consistent with the results from the literature (Abraham et al., 2002). For commuting trips, a 1% increase in trip proportion on multi-use pathways and a 1% increase in the trip proportion on cycle tracks is equivalent to decreasing distance by 4.8% and 0.7% respectively. This result is similar for all the models estimated, and when considering on-street facilities, the coefficient of shared lanes (0.389) is greater than bike lanes (0.337) in commute models, suggesting that cyclist commuters consider shared lanes faster than bike lanes, and they are comfortable sharing lanes together with motor traffic when given higher priority. Another reason for this is that shared lanes always appear on major streets where space is limited for separate bike lanes; commuting cyclists consider directness more important than safety, so major streets with shared lanes have a higher possibility of being chosen. Recreational cyclists also show obvious preference on off-street bicycle facilities. A 1% increase in recreational trip proportion on multi-use pathways is equivalent to the benefit made by 9.05% increase of distance. On the other hand, recreational cyclists consider safety more important than speed and directness, and recreational trips always happen on unpopular roads where bike lanes are able to be built, so the coefficient of bike lanes (1.34) is greater than shared lanes (0.534) suggesting a higher preference for safer facilities (bike lanes).

5.4.4.2 Effects of Generic Trip Attributes

Personal characteristics of cyclists may also have an effect on the route choices such as gender, age, and rider history as well as other trip-generic attributes such as weather factors and month of the trip (Antonakos, 1994; Hulse & Shafizadeh, 1996; Krizek, 2006; Sener et al., 2009). This section discusses the results of a market segmentation aligns for each of these attributes. Table 5-8, 5-9, 5-10 (PSL12 to PSL20) shows the estimation results of models with different market segmentations by socio-demographic characteristics of commuting cyclists (age, gender and rider history). Table 5-11 and 5-12 (PSL21 to PSL25) states the estimation results of models with different market segmentations by mean temperature (weather) and month periods.

Table 5-8 Model Estimations with Different Market Segmentations by Age

Age	18-24 (PSL12)		25-34 (PSL13)		35-49 (PSL14)		>50 (PSL15)	
Variables	Coeff.	t-test	Coeff.	t-test	Coeff.	t-test	Coeff.	t-test
LN(Distance (km))	-4.39	-7.17	-4.13	-36.48	-2.85	-23.56	-3.58	-17.56
Energy Consumption (1000s KJ)	0.223	1.26	-0.530	-15.01	-0.102	-3.04	0.199	4.18
Number of Transit Stops	0.0421	0.53	0.0142	11.91	0.01	10.58	0.00462	2.52
Annual average daily traffic volumes (1000s veh/day)	-0.239	-1.70	-0.135	-4.83	-0.273	-10.31	-0.0757	-1.4
Trip proportion (TP) with major arterial	18.6	14.55	17.0	69.53	19.5	71.06	17.4	37.59
TP with minor arterial	15.6	12.46	17.3	63.67	17.8	64.85	16.9	35.98
TP with collector	14.8	10.49	16.0	52.67	18.8	61.86	18.0	32.04
TP with local street	11.5	6.94	10.6	29.26	13.1	38.80	14.7	22.95
TP with shared lane	6.00	5.96	1.43	8.14	-0.537	-2.65	1.76	4.77
TP with bike lane	7.56	7.94	0.395	2.67	-0.475	-2.55	2.25	6.77
TP with multiuse trail	18.3	8.57	17.8	40.15	18.1	43.25	20.8	26.17
TP with cycle track	9.44	5.65	-0.0330	-0.12	4.69	10.85	10.6	11.41
TP without bicycle facilities	0.355	0.43	-2.81	-24.02	-3.59	-23.47	-1.64	-5.76
Ln(Path Size Factor)	5.99	14.35	6.32	76.78	6.15	66.01	7.10	38.93
Number of Observations	604		14467		11692		4030	
Final log-likelihood	-562.753		-15242.469		-13126.876		-4706.816	
Rho-Square	0.694		0.654		0.631		0.615	

Table 5-9 Model Estimations with Different Market Segmentations by Gender

Gender	Male (PSL16)		Female (PSL17)	
Variables	Coeff.	t-test	Coeff.	t-test
LN(Distance (km))	-5.57	-23.62	-4.01	-33.88
Energy Consumption (1000s KJ)	-0.596	-9.77	-0.138	-3.72
Number of Transit Stops	0.0176	6.92	0.0119	12.06
Annual average daily traffic volumes (1000s veh/day)	-0.361	-6.61	-0.262	-9.87
Trip proportion (TP) with major arterial	13.8	29.05	18.6	68.03
TP with minor arterial	11.9	25.10	18.3	67.45
TP with collector	11.0	20.55	17.7	58.47
TP with local street	5.40	8.57	13.0	37.72
TP with shared lane	2.66	6.66	1.95	10.31
TP with bike lane	0.258	0.79	0.737	4.35
TP with multiuse trail	10.6	13.07	19.3	46.83
TP with cycle track	4.19	5.35	0.809	2.44
TP without bicycle facilities	-5.03	-17.31	-2.09	-15.39
Ln(Path Size Factor)	7.27	39.91	6.55	70.76
Number of Observations	3676		12318	
Final log-likelihood	-3472.892		-13552.072	
Rho-Square	0.690		0.639	

Table 5-10 Model Estimations with Different Market Segmentations by Rider History

Rider History	Since Childhood (PSL18)		Several Years (PSL19)		One Year or Less (PSL20)	
Variables	Coeff.	t-test	Coeff.	t-test	Coeff.	t-test
LN(Distance (km))	-3.87	-40.24	-2.61	-16.36	-2.94	-15.42
Energy Consumption (1000s KJ)	0.0382	1.52	-0.557	-11.62	-0.368	-6.91
Number of Transit Stops	0.0110	13.19	-0.0029	-2.13	0.0212	10.37
Annual average daily traffic volumes (1000s veh/day)	-0.182	-8.35	-0.518	-11.74	0.271	5.72
Trip proportion (TP) with major arterial	17.6	84.60	21.3	53.95	17.8	43.61
TP with minor arterial	17.4	81.60	18.3	45.40	17.9	37.74
TP with collector	15.4	64.16	19.5	43.50	20.4	37.64
TP with local street	12.2	44.65	10.4	19.75	15.4	23.83
TP with shared lane	0.564	3.49	3.32	13.67	-2.79	-8.80
TP with bike lane	0.461	3.29	0.681	3.13	0.686	2.48
TP with multiuse trail	18.9	55.35	16.4	25.67	22.9	29.01
TP with cycle track	1.90	6.51	4.66	10.35	1.30	2.26
TP without bicycle facilities	-2.82	-24.37	-2.36	-13.98	-3.69	-16.04
Ln(Path Size Factor)	6.14	84.58	6.26	50.98	6.28	44.85
Number of Observations	18525		7202		5007	
Final log-likelihood	-20651.805		-7577.268		-5447.667	
Rho-Square	0.634		0.654		0.643	

Table 5-11 Model Estimations with Different Market Segmentations by Weather

Temperature	<= 5 degree (PSL21)		> 5 degree (PSL22)	
Variables	Coeff.	t-test	Coeff.	t-test
LN(Distance (km))	-4.55	-21.26	-3.49	-45.61
Energy Consumption (1000s KJ)	-0.224	-3.65	-0.141	-6.94
Number of Transit Stops	0.0179	8.67	0.0092	13.87
Annual average daily traffic volumes (1000s veh/day)	0.0851	1.75	-0.248	-13.56
Trip proportion (TP) with major arterial	15.9	36.12	18.0	106.61
TP with minor arterial	16.2	33.25	16.8	97.07
TP with collector	18.0	32.79	16.3	83.67
TP with local street	14.4	22.37	11.2	49.64
TP with shared lane	0.831	2.49	0.321	2.58
TP with bike lane	0.552	1.85	0.289	2.65
TP with multiuse trail	21.7	27.43	17.3	62.26
TP with cycle track	3.31	5.18	2.49	10.57
TP without bicycle facilities	-2.72	-11.74	-3.06	-34.21
Ln(Path Size Factor)	7.00	42.00	6.01	104.65
Number of Observations	3917		29305	
Final log-likelihood	-4594.767		-33004.643	
Rho-Square	0.615		0.630	

Table 5-12 Model Estimations with Different Market Segmentations by Month

Month	May – October (PSL23)		October – December (PSL24)		December – January (PSL25)	
Variables	Coeff.	t-test	Coeff.	t-test	Coeff.	t-test
LN(Distance (km))	-3.52	-41.79	-3.63	-23.51	-4.76	-14.52
Energy Consumption (1000s KJ)	-0.154	-6.86	-0.142	-3.33	-0.112	-1.23
Number of Transit Stops	0.00860	12.02	0.0120	8.07	0.0286	8.98
Annual average daily traffic volumes (1000s veh/day)	-0.287	-13.76	-0.0601	-1.75	0.210	2.81
Trip proportion (TP) with major arterial	18.3	98.20	16.4	49.22	15.0	22.91
TP with minor arterial	16.9	87.94	16.6	47.49	15.1	20.69
TP with collector	16.0	73.73	17.5	45.04	18.6	22.22
TP with local street	10.9	42.96	12.8	28.77	15.5	15.66
TP with shared lane	0.358	2.64	0.192	0.75	1.04	1.95
TP with bike lane	0.243	2.03	0.0712	0.31	1.29	2.78
TP with multiuse trail	16.8	53.59	20.0	36.13	23.3	19.09
TP with cycle track	2.24	8.74	3.32	6.91	4.08	3.90
TP without bicycle facilities	-3.14	-31.77	-2.75	-15.69	-3.11	-8.63
Ln(Path Size Factor)	5.98	95.55	6.38	52.79	7.40	27.40
Number of Observations	24915		6556		1751	
Final log-likelihood	-27610.141		-7883.094		-1983.559	
Rho-Square	0.636		0.605		0.628	

Socio-demographic characteristics (age, gender, rider history)

As shown in Table 5-8 to 5-10, in general, socio-demographic characteristics including age, gender and rider history do not have great influences on route choice behavior when comparing the estimation results with different kinds of cyclists and the results of the route choice model with all trips. The direction of the effect of each variable and the magnitude of the effects remain similar, which is consistent with the literature (Abraham et al., 2002; Hood et al., 2001). However, age was found to slightly influence the energy consumption and preferences towards bicycle facility type, which is similar with rider history.

Table 5-8 shows the estimation results based on cyclists in different age groups, which demonstrates that older cyclists care about more about energy consumption than younger cyclists. The absolute coefficient of energy consumption in cyclists between 25 and 34 years old was larger than cyclists between 18 and 24 years old (0.53 vs. 0.223). In contrast, the sensitivity of trip distance seems to decrease as age grows, which suggests older cyclists are more sensitive to slope than distance, and are willing to choose a longer trip to reduce their trip gradient. However, there is a counter-intuitive positive coefficient in energy consumption (0.199) for cyclists older than 50 years old, suggesting that cyclists older than 50 years old want to do more exercise. This may be also due to the trip they made are shorter trips, which are more sensitive to distance than energy consumption. Another reason may be the small trip samples produced by cyclists older than 50 years old. The sample size for older cyclists is much lower than the middle-aged cyclists (85% lower), which may also cause counter-intuitive results.

In addition, for cyclists older than 50 and between 18 and 24 years old, safety is more important than that for the middle-aged cyclists. The coefficients of bike lanes and dedicated off-street bicycle facilities (multiuse pathways and cycle tracks) are much larger than that of the shared lanes. The coefficient of shared lanes (-0.537) and bike lanes (-0.475) of cyclists between 35 and 49 even becomes negative. Similarly, the results also show a higher preference of low traffic volume local streets for these two groups than for middle-aged cyclists. The value of coefficients become around 27% higher than the commuting route choice model with all cyclists (Table 5-4) also suggests the awareness of safety for older cyclists.

Rider history also has influences on energy consumption and bicycle facility preferences. Rider history suggests how skillful a cyclist is with cycling. As cyclists become more skillful, they know the network better, and are therefore better able to identify the shortest route and avoid detours. The model shows experienced cyclists are less concerned with the energy consumption, which may be because they are accustomed to the energy consumption of cycling. In addition, cyclists with less cycling skill prefer perceived safer bicycle facilities: bike lanes, multiuse pathways and cycle tracks more than skillful cyclists (22.9 vs. 18.9 and 16.4).

Gender in Table 5-9 does not show obvious differences from the commuting route choice model with all trips. Both female and male cyclists prefer shorter trips with less energy consumption and traffic volume. They prefer major roads more than local streets for the directness from the origin to the destination. Dedicated off-street bicycle facilities are considered more preferred than shared lanes and bike lanes. Those conclusions are same with all commuting route choice models suggesting gender is not significant in cyclist's route choice behavior. Comparing

estimations with male cyclists and female cyclists, results show that male cyclists are more sensitive to distance, energy consumption as well as safety than female cyclists, which is inconsistent with the literature (Antonakos, 1994). This may be due to the lack of gender information, as many cyclists did not submit their gender, and the number of trip observations made by male cyclists is only 10% of the total commuting trips.

Weather characteristics (temperature)

In addition to socio-demographic characteristics, it is intuitively expected that weather characteristics such as temperature also affect cyclists' route choices. Table 5-11 shows that the model estimations based on trips produced in temperatures higher than 5 degree are similar with the commuting route choice model results based on all trips. However, in cold weather, cyclists are more sensitive to distance and energy consumption than warmer days (-4.55 vs. -3.49 for trip distance and -0.224 vs. -0.141 for energy consumption). Furthermore, the results show that in cold weather, cyclists seem more likely to choose a route with higher traffic volume (0.0851) and major streets (16.2 and 18). This may be because roads with more traffic volume in winter time are in better conditions than local streets from the perspective of winter road maintenance. Finally, for all kinds of weathers, cyclists prefer major roads with off-street bicycle facilities as specified.

Time (month) of the trip

The results of model estimation with different market segmentations by month are consistent with temperature, as winter months (December and January) generally have lower temperatures, and autumn months (October to December) have higher temperatures than winter months but lower temperatures than spring and summer months (May to October). In general, in spring, summer and autumn months, the month is not significant to route choice, while in the winter time, cyclists, care more about distance, which is consistent with the effects of weather characteristics found in the literature (B öcker et al., 2013).

5.4.4.3 Path-size Parameter

The Path size factor was estimated to have positive coefficients in all models, which is consistent with the theory of path size factor (see Section 5.3.4) (Table 5-4, 5-5, 5-6). As expected, they are significantly larger than zero, which captures the statistical errors caused by the IIA property as discussed previously. The log-likelihood and Rho-square of PSL are higher than MNL (0.627 vs. 0.539), suggesting better estimated results. Although the coefficients of travel distance and energy consumption in MNL are positive, this is not the case in the PSL model, which suggests that the IIA problem is important and ignoring it may lead to inaccurate results. The coefficients of the path size parameter in recreational models are smaller to those in the commuting models, which is consistent with the observed analysis that for the recreational trips, the alternative routes have more overlapping, as discussed in previous sections.

In conclusion, according to the results of both commuting and recreational route choice models (Table 5-4, 5-5) and the models with different market segmentations by trip length (Table 5-6)

and other generic trip attributes (Table 5-8 to Table 5-12), commuting cyclists prefer shorter routes with less energy consumption and motor traffic volume, while recreational cyclists prefer longer routes and routes with higher energy consumption. This result is consistent with most of the literature. In addition, both types of cyclists were found to show a strong preference toward major and minor arterials followed by local streets when compared to all other road types. All cyclists were confirmed to prefer on bicycle facilities, especially the dedicated off-street bicycle facilities such as multiuse pathway and cycle tracks. Bike lanes are more preferred than shared lanes for recreational trips, while commuting cyclists have opposite attitudes towards those two bicycle facilities. Cyclist commuters were found to consider directness more important than other factors such as type of roads, traffic volume and facility types. Socio-demographic characteristics have slight influences on route choice behavior, which is consistent with the results in the literature. Age was found to have influences on energy consumption as well as the preferences of bicycle facilities. Older people prefer routes with lower energy consumption and safer bicycle facilities such as dedicated off-street bicycle facilities. Rider history shows the level of skill of cycling, and more skillful cyclists prefer shorter, faster and more direct routes. They are not as sensitive as beginner cyclists to the dedicated bicycle facilities. Gender was found to not be a statistical significant factor in cyclists' route choice behavior. However, as the number of sample trips is limited in some models, there are some counter-intuitive results in route choice models with different socio-demographic characteristics. Weather characteristics (temperature) and month of the trip have high correlation. Lower temperature in winter were found to transfer some cyclists to major streets with higher traffic volume, due to their better winter road condition. They do not have statistical significant effects on other factors. The Path-size parameter has a positive effect, which agrees with theory and previous literature, and provides a corrective benefit on the errors caused by IIA property.

Chapter 6 Conclusions and Future Research

As the fifth largest city in North America, the City of Toronto is currently developing a new long term plan to revitalize its cycling network with more than 73 million dollars of investment. This research was motivated by the need to develop a comprehensive understanding of the cycling activities in the City, including origin-destination distribution, bicycle flows on road network and speed distributions. The research focuses particularly on understanding cyclists' route choice decisions and identifying the factors influencing their behavior. An extensive modeling analysis of route choice behavior was conducted using a unique data set featuring large sample size, complete road network representation, and detailed tracing of cycling routes. This chapter summarizes the main findings and contributions and highlight the future research directions.

6.1 Summary of Work and Findings

This research used cycling travel data collected from a group of cyclists in the City of Toronto who volunteered to participate in the study. Each cyclist used a smartphone application – CycleTrack to track their individual trips and submit all relevant data such as personal attributes and trip purpose. Additional data that are needed for route choice analysis were obtained, including road network, the locations of transit stops, road-side parking from Toronto Open Data repository, annual average daily traffic volumes (AADT) from Toronto Traffic Management Center, and weather data from Environment Canada. Furthermore, the slope of each link is calculated based on the digital elevation model (DEM) obtained from the Geospatial Center of University of Waterloo. A total of 4,556 cyclists participated in the research. However, because the participants are not required to submit the personal attributes, a large number of cyclists opted out in providing some of the personal attributes such as household income, comfort level, rider history, and gender.

The raw GPS readings were first preprocessed to remove the outliers and those that are not close to any of the roads. The individual GPS points were then matched to the specific road segments of the road network using a map matching algorithm. The Toronto network consists of 116,402 links and 40,808 nodes. The links are classified into five types, including major arterials, minor arterials, collectors, local streets and others. The road network also includes links representing various bicycle facilities, including bike lanes, shared lanes, multiuse pathways, cycle tracks and others. As most of the local streets do not have bicycle facilities, the coverage of the cycling network is limited with a coverage of around 12%. Among all the bicycle facilities, multiuse pathway takes the largest part of 34.2%, while bike lane and shared lane take similar proportions of 17.9%. Cycle tracks with a proportion of 0.3% is the smallest part of bicycle facilities. After map-matching, trips were generated with start and end points and time, the length of the trip, the average speed of the trip as well as a sequential links. Some validation was made to exclude these trips that had an abnormal high speed, extreme short length (less than 500 meters), or unreasonable long detours. A total of 51,449 valid trips were finally identified, among which 33,222 trips are for commute and the remaining 9611 trips are recreational. Most of the trips (68.62%) are in middle distance from 2 km to 9 km.

Discrete choice models were then developed for analyzing the route choice behavior of the cyclists using the collected cycling trip data. For each cycling trip, in addition to the route that

was actually chosen by the cyclist, a given number of possible alternative routes were generated by applying a shortest path algorithm iteratively. After each iteration, a cost penalty is added to the links on the shortest path to reduce their chance to be on the shortest path again on the next iteration of shortest path search. Distance and energy consumption are used in generating alternative routes. A shortest path algorithm was run for a pre-specified number of times based on each of the two items. After each run, links on the shortest path are penalized with a cost multiplier so that the new shortest path generated in the next iteration has minimum overlapping with the previous ones. This is intended to minimize the similarity between alternative routes that is needed to meet the Independence from Irrelevant Alternatives (IIA) assumption required by the commonly used multinomial logit choice model. The remaining correlation due to some unavoidable overlapping was further mitigated by using a variant of the multinomial logit model - path-size logit model (PSL) that was developed specifically to modeling route choices. A large number of models were developed and the following findings were reached:

- Both travel distance and annual average daily traffic volume were consistently shown to be statistical significant for both commute and recreation purposes.
- Commuting cyclists prefer shorter and direct routes with less traffic and more off-road facilities, which is consistent with the results in previous studies.
- Comparing with different types of facilities, off-road facilities such as multiuse trails and cycle tracks could increase the probability of choosing the route, while the shared lane are preferred than bike lane for the cyclists.
- Recreational cyclists have different attitude towards distance and energy consumption. Extra distance or energy consumption is consumed as benefits for exercise, so cyclists have a positive attitude towards distance and energy consumption. It was also found that for recreational cycling trips travelers tended to place heavier weight on safety consideration than that by those making commuting trips. They have a higher preference on streets with separate bike lanes than shared streets.
- Some socio-demographic characteristics such as gender do not have substantial effects on route choice behavior.
- Age influences cyclists' attitudes towards energy consumption and trip distance. Older cyclists prefer more energy saving routes even it makes a little detours. They also place heavier weight on safety consideration than middle-aged cyclists. They have a higher preference on streets with bike lanes and dedicated off-street bicycle facilities.
- Skillful cyclists with more cycling experience prefer shorter, faster and more direct routes, and do not care much about the energy consumption. Beginners show a high preference to streets with separate bike lanes and off-street bicycle facilities.
- Lower temperature and the winter months slightly increase the preferences of shorter and more energy saving routes. They also show a preference of high traffic volume due to the better winter road conditions in snow cleaning and drainage.

These findings shed lights on the importance of establishing a connected cycling network with dedicated cycling facilities such as bike lanes, cycling tracks and multiuse trails for improved cycling experience and ultimately increased share of active travel modes.

6.2 Contributions and Future Research

This research has conducted an extensive modeling analysis of the route choice behavior of a sample of cyclists in the City of Toronto using a unique trip dataset featuring large sample size, large road network, and detailed trip information. The study has made the following contributions:

- The first challenge in route choice studies is how to identify the travel paths based on GPS readings that contain positioning errors. A map-matching method needs to be used to link GPS coordinates to the underlying road network. The underlying process is however resource intensive, requiring large computer memory and processing time. In this research, we developed a new map-matching algorithm that makes use of various integral functionalities of a GIS platform (ArcGIS Desktop) and a GIS-enabled database (PostgreSQL), showing the advantages of both computational efficiency and high matching accuracy. The proposed method can also be used in other similar transportation applications such as vehicle or pedestrian route tracking and speed or delay calculation at the intersections.
- The second challenge in a route choice study is the generation of a route choice set that contains the possible routes that may have been considered by the cyclists when making their individual cycling trips. In most of the previous studies, the alternative routes were chosen on the basis of travel time or distance, ignoring many other factors such as collision risk, level of difficulties, and degree of similarity between alternative routes. The resulting choice set often contains overly similar routes with substantial overlapping. In this study, we have introduced a new choice set generation algorithm that includes multiple routing criteria for capturing the variety of factors that are often considered by cyclists in making their route choices, such as, overall energy consumption. In addition, the algorithm includes a link penalty function to penalize these links that have already appeared in the identified routes, thus increasing the variety of the alternative routes. This route choice set generation algorithm could also be applied in similar studies of other travel modes.
- A large number of discrete choice models have been developed and evaluated to investigate the sensitivity of the modeling results to both modeling options and route choice factors. The former includes the size of the route choice set (e.g., 10 alternatives vs. 20 alternatives), routing criteria (e.g., travel distance based vs. energy consumption based), and market segments by trip distance (e.g., < 5 km, 5 km to 10 km, and > 10 km). Route choice factors include operational, road network related, weather, and cyclist attributes. Few of the previous studies have involved such an extensive analysis in terms of the amount of data being used and the number of factors and modeling options being considered. The findings on modeling options are expected to have a long last value to the researchers in this field. The route choice modeling results have policy implications as they reveal the importance of developing connected, high-quality cycling infrastructure for attracting more people using cycling as their traveling mode. In addition, the comparison between model results for

commuting and recreational trips leads to effective policy initiatives targeted at particular group of cyclists (commuting and recreational cyclists). Finally, the marginal rates of substitution with respect to the natural log of distance have been estimated indicating how much additional travel distance cyclist would be willing to travel to avoid or obtain specific route attributions. This findings have great values in cost-benefit evaluation in route improvements.

While our research has contributed to the progress in improving our understanding of cyclists' route choice behavior, it can be further extended in several importation directions. Potential future research and development on this topic are identified as follows:

- The map-matching method proposed in this study is based on idea of assigning the closest link to the GPS points. This method may fail to yield one-to-one match for these GPS points that are new multiple streets (e.g., near intersections or parallel streets). For these non-unique matching cases, they were excluded from the trips or routes; instead, a short path algorithm was applied to connect the identified links along a route. The validity and performance of this method needs to be examined in future research and improved methods need to be developed if necessary.
- This research considered two cost measures, namely, distance and energy consumption, in generating alternative routes, which may not be sufficient in identifying all potential alternatives that cyclists consider. Future research should further examine additional route choice criteria in generating route choice set, such as those that were considered in the subsequent route choice modeling. In addition, larger route choice sets may be considered for each trip. It is possible that consideration of more alternatives may lead to higher accuracy of the modeling results.
- While the trip data set used in this research is arguably one of the largest used in studies on cyclist route choice, more data could be incorporated to strengthen the findings underlying a wider range of behavior. For example, a rich set of longitudinal data on a group of specific cyclists would allow in-depth investigation of cyclist behavior change and effect of learning. Furthermore, the study could be extended to other cities and regions to investigate the generalizability of the models and findings.
- The route choice models developed in this research presents an opportunity to develop customized route-finding tools and navigation applications. The utility functions in the route choice models could be used to define the generalized travel cost at the link level and then used in best path searching. There are an increasing number of similar tools to assist people finding routes for particular trips (Broach et al., 2012). The improved understanding of cyclists' preferences can make the route-finding tools more relevant and effective (Su et al., 2010).

References

- Abraham, J. E., McMillan, S., Brownlee, A. T., & Hunt, J. D. (2002, January). Investigation of cycling sensitivities. In *81st Annual Meeting of the Transportation Research Board. Transportation Research Board, Washington, DC.*
- Antonakos, C. L. (1994). Environmental and Travel Preferences of Cyclists. *Transportation Research Record*, 1438:25-33.
- Atkinson, M. D., Sack, J. R., Santoro, N., & Strothotte, T. (1986). Min-max heaps and generalized priority queues. *Communications of the ACM*, 29(10), 996-1000.
- Aultman-Hall, L., Hall, F., & Baetz, B. (1997). Analysis of bicycle commuter routes using geographic information systems: implications for bicycle planning. *Transportation Research Record: Journal of the Transportation Research Board*, (1578), 102-110.
- Aultman-Hall, L., & Kaltenecker, M. G. (1999). Toronto bicycle commuter safety rates. *Accident Analysis & Prevention*, 31(6), 675-686.
- ArcGIS Resource Center, Add Surface Information, 2016, <http://maxrice.com/excel-geocoding-tool/>. Accessed 11-08-2016.
- Bacchieri, G., Barros, A. J., Dos Santos, J. V., & Gigante, D. P. (2010). Cycling to work in Brazil: users profile, risk behaviors, and traffic accident occurrence. *Accident Analysis & Prevention*, 42(4), 1025-1030.
- Bacchieri, G., et al. (2005). Determinants and patterns of bicycle use and traffic accidents among bicycling workers in Pelotas, Rio Grande do Sul, Brazil. *Cad. Saude Publ.* 21, 1499–1508.
- Bellman, R. (1958). On a routing problem. *Quarterly of applied mathematics*, 87-90.
- Ben-Akiva, M. E., & Lerman, S. R. (1985). *Discrete choice analysis: theory and application to travel demand* (Vol. 9). MIT press.
- Ben-Akiva, M., and M. Bierlaire (1999). Discrete Choice Methods and Their Applications to Short-Term Travel Decisions. In *Handbook of Transportation Science* (R. W. Hall, ed.). Kluwer Academic, Boston, Mass.
- Bierlaire, M. (2003). BIOGEME: A free package for the estimation of discrete choice models. In *Proceedings of the 3rd Swiss Transportation Research Conference, Ascona, Switzerland.*
- Bierlaire, M., & Frejinger, E. (2005). Route choice models with subpath components. In *Swiss Transportation Research Conference* (No. TRANSP-OR-CONF-2006-032).
- Bierlaire, M., & Frejinger, E. (2008). Route choice modeling with network-free data. *Transportation Research Part C: Emerging Technologies*, 16(2), 187-198.

- Bliemer, M., & Bovy, P. (2008). Impact of route choice set on route choice probabilities. *Transportation Research Record: Journal of the Transportation Research Board*, (2076), 10-19.
- Böcker, L., Dijst, M., & Prillwitz, J. (2013). Impact of everyday weather on individual daily travel behaviours in perspective: a literature review. *Transport reviews*, 33(1), 71-91.
- Bovy, P. H., & Bradley, M. A. (1985). *Route choice analyzed with stated-preference approaches* (No. 1037).
- Bovy, P. H., & Fiorenzo-Catalano, S. (2007). Stochastic route choice set generation: behavioral and probabilistic foundations. *Transportmetrica*, 3(3), 173-189.
- Broach, J., Dill, J., & Gliebe, J. (2012). Where do cyclists ride? A route choice model developed with revealed preference GPS data. *Transportation Research Part A: Policy and Practice*, 46(10), 1730-1740.
- Brodal, G. S. (1996, January). Worst-Case Efficient Priority Queues. In *SODA* (Vol. 96, pp. 52-58).
- Cabbagetown Cycling Club (2015), *Ride Levels & Speeds*, <http://cabbagetowncyclingclub.ca/ride-levels-speeds/>.
- Cascetta, E., Nuzzolo, A., Russo, F., & Vitetta, A. (1996, July). A modified logit route choice model overcoming path overlapping problems: specification and some calibration results for interurban networks. In *Proceedings of the 13th International Symposium on Transportation and Traffic Theory* (pp. 697-711). Lyon, France: Pergamon.
- Chen, C. F., & Chen, P. C. (2010). Estimating recreational cyclists' preference on bicycle route facility-Evidence from Taiwan. In *12th World Conference on Transport Research*.
- City of Toronto (2010), Trinity Bellwoods Park - Proposed Multi-Use Path Upgrade, https://www1.toronto.ca/city_of_toronto/transportation_services/cycling/files/pdf/Trinity_Panel2_Bikeways.pdf, Accessed 11-08-2016.
- City of Toronto (2011), Chapter 4 Bicycle Friendly Streets, *Toronto Bike Plan*, 4-1-4-10.
- City of Toronto (2011), Chapter 5 Bikeway Network, *Toronto Bike Plan*, 5-1-5-14.
- City of Toronto (2016), *Toronto Cycling statistics*, 2016.
- City of Toronto (2016), *Toronto Cycling App*, 2016.
- City of Toronto (2016), *About the Road Classification System*, <http://www1.toronto.ca/wps/portal/contentonly?vgnextoid=8a8d2118b7412410VgnVCM1000071d60f89RCRD&vgnextchannel=6f2c4074781e1410VgnVCM10000071d60f89RCRD>.

- Davis, W. J. (1987). *Bicycle safety evaluation*.
- Daganzo, C. F., & Sheffi, Y. (1977). On stochastic models of traffic assignment. *Transportation science*, 11(3), 253-274.
- Delcan and Lura Consulting (2013), *City of Toronto Congestion Management Plan*.
- Dijkstra, E. W. (1959). A note on two problems in connexion with graphs. *Numerische mathematik*, 1(1), 269-271.
- Environment Canada (2013), Canada's Emission Trends.
- Epperson, B. (1994). *Evaluating suitability of roadways for bicycle use: Toward a cycling level-of-service standard* (No. HS-042 007).
- Eppstein, D. (1998). Finding the k shortest paths. *SIAM Journal on computing*, 28(2), 652-673.
- Fu, L., Sun, D., & Rilett, L. R. (2006). Heuristic shortest path algorithms for transportation applications: state of the art. *Computers & Operations Research*, 33(11), 3324-3343.
- Fuchs, F. (2010). On Preprocessing the ALT-Algorithm. *Student thesis, Faculty of Computer Science, Institut for Theoretical Informatics (ITI), Karlsruhe Institute of Technology (KIT)*.
- Fredman, M. L., & Tarjan, R. E. (1987). Fibonacci heaps and their uses in improved network optimization algorithms. *Journal of the ACM (JACM)*, 34(3), 596-615.
- Frejinger, E., & Bierlaire, M. (2007). Capturing correlation with subnetworks in route choice models. *Transportation Research Part B: Methodological*, 41(3), 363-378.
- Goldberg, A. V., & Harrelson, C. (2005, January). Computing the shortest path: A search meets graph theory. In *Proceedings of the sixteenth annual ACM-SIAM symposium on Discrete algorithms* (pp. 156-165). Society for Industrial and Applied Mathematics.
- Greenfeld, J. S. (2002). Matching GPS observations to locations on a digital map. In *Transportation Research Board 81st Annual Meeting*.
- Hood, J., Sall, E., & Charlton, B. (2013). A GPS-based bicycle route choice model for San Francisco, California. *Transportation letters*.
- Hoogendoorn-Lanser, S., van Nes, R., & Bovy, P. (2005). Path size modeling in multimodal route choice analysis. *Transportation Research Record: Journal of the Transportation Research Board*, (1921), 27-34.
- Hopkinson, P., & Wardman, M. (1996). Evaluating the demand for new cycle facilities. *Transport Policy*, 3(4), 241-249.
- Howard, C., & Burns, E. (2001). Cycling to work in Phoenix: route choice, travel behavior, and commuter characteristics. *Transportation Research Record: Journal of the Transportation Research Board*, (1773), 39-46.

- Hulse, L., & Shafizadeh, K. (1996, July). Bicycle Travel-to-Work Patterns. In *Proceedings from the National Conference on Undergraduate Research* (Vol. 2). University of North Carolina, Asheville.
- Hyodo, T., Suzuki, N., & Takahashi, K. (2000). Modeling of bicycle route and destination choice behavior for bicycle road network plan. *Transportation Research Record: Journal of the Transportation Research Board*, (1705), 70-76.
- Jan, O., Horowitz, A., & Peng, Z. R. (2000). Using global positioning system data to understand variations in path choice. *Transportation Research Record: Journal of the Transportation Research Board*, (1725), 37-44.
- Krizek, K. J. (2006). Two approaches to valuing some of bicycle facilities' presumed benefits: Propose a session for the 2007 national planning conference in the city of brotherly love. *Journal of the American Planning Association*, 72(3), 309-320.
- Li, H., Guensler, R., & Ogle, J. (2005). Analysis of morning commute route choice patterns using global positioning system-based vehicle activity data. *Transportation Research Record: Journal of the Transportation Research Board*, (1926), 162-170.
- Lim, Y., & Kim, H. (2005). A shortest path algorithm for real road network based on path overlap. *Journal of the Eastern Asia Society for Transportation Studies*, 6, 1426-1438.
- Marchal, F., Hackney, J., & Axhausen, K. W. (2004). Efficient map-matching of large GPS data sets-Tests on a speed monitoring experiment in Zurich. *Arbeitsbericht Verkehrs-und Raumplanung*, 244.
- Max Rice and Juice Analytics, Excel Geocoding Tool (2015), <http://maxrice.com/excel-geocoding-tool/>. Accessed 11-08-2016.
- Mehlhorn, K., & Sanders, P. (2008). *Algorithms and data structures: The basic toolbox*. Springer Science & Business Media.
- Menghini, G., Carrasco, N., Shussler, N., & Axhausen, K. (2009). Route Choice of Cyclists in Zurich: GPS-based discrete choice models. *Institute for Transport Planning and Systems*.
- Momjian, B. (2001). *PostgreSQL: introduction and concepts* (Vol. 192). New York: Addison-Wesley.
- National Travel Surveys Statistics (2013), *Development of Transport*, <https://www.gov.uk/government/collections/national-travel-survey-statistics/>. Accessed 11-08-2016.
- Nelson, A., & Allen, D. (1997). If you build them, commuters will use them: association between bicycle facilities and bicycle commute. *Transportation Research Record: Journal of the Transportation Research Board*, (1578), 79-83.

- NHTSA and Bureau of Transportation Statistics (2003), U.S. Department of Transportation. *National Survey of Pedestrian and Bicyclist Attitudes and Behaviors: Highlights Report*.
- Nielsen, O. A. (2000). A stochastic transit assignment model considering differences in passengers
- Nielsen, O. A., & Jørgensen, R. M. (2004). Map-matching algorithms for GPS-data-methodology and test on data from the AKTA road pricing experiment in Copenhagen. In *World Conference on Transport Research Society (WCTRS)*.
- OECD (2014), *OECD Economic Surveys: United States 2014*, OECD Publishing, Paris.
DOI: http://dx.doi.org/10.1787/eco_surveys-usa-2014-en
- Olds, T. S., Norton, K. I., & Craig, N. P. (1993). Mathematical model of cycling performance. *Journal of Applied Physiology*, 75(2), 730-737.
- Ontario Ministry of Finance (2016), *Ontario Population Projections Update, 2015–2041*, <http://www.fin.gov.on.ca/en/economy/demographics/projections>. Accessed 25-07-2016.
- Paek, J., Kim, J., & Govindan, R. (2010, June). Energy-efficient rate-adaptive GPS-based positioning for smartphones. In *Proceedings of the 8th international conference on Mobile systems, applications, and services* (pp. 299-314). ACM.
- Parkin, J., Wardman, M., & Page, M. (2007). Models of perceived cycling risk and route acceptability. *Accident Analysis & Prevention*, 39(2), 364-371.
- Prato, C., & Bekhor, S. (2007). Modeling route choice behavior: how relevant is the composition of choice set?. *Transportation Research Record: Journal of the Transportation Research Board*.
- Prato, C. G. (2009). Route choice modeling: past, present and future research directions. *Journal of Choice Modelling*, 2(1), 65-100.
- Pucher, J., & Dijkstra, L. (2003). Promoting safe walking and cycling to improve public health: lessons from the Netherlands and Germany. *American journal of public health*, 93(9), 1509-1516.
- Pucher, J., & Renne, J. L. (2003). Socioeconomics of urban travel: evidence from the 2001 NHTS. *Transportation Quarterly*, 57(3), 49-77.
- Pyo, J. S., Shin, D. H., & Sung, T. K. (2001). Development of a map matching method using the multiple hypothesis technique. In *Intelligent Transportation Systems, 2001. Proceedings. 2001 IEEE* (pp. 23-27). IEEE.
- Quddus, M. A., Ochieng, W. Y., Zhao, L., & Noland, R. B. (2003). A general map matching algorithm for transport telematics applications. *GPS solutions*, 7(3), 157-167.

- Ramming, M. S. (2001). *Network knowledge and route choice* (Doctoral dissertation, Massachusetts Institute of Technology).
- Region of Waterloo (2004), Regional of Waterloo Cycling Master Plan, [Online; accessed 5-July-2016].
- Rieser-Schüssler, N., Balmer, M., & Axhausen, K. W. (2013). Route choice sets for very high-resolution data. *Transportmetrica A: Transport Science*, 9(9), 825-845.
- Sener, I. N., Eluru, N., & Bhat, C. R. (2009). An analysis of bicycle route choice preferences in Texas, US. *Transportation*, 36(5), 511-539.
- Shafizadeh, K., & Niemeier, D. (1997). Bicycle journey-to-work: travel behavior characteristics and spatial attributes. *Transportation Research Record: Journal of the Transportation Research Board*, (1578), 84-90.
- Skiena, S. (1990). Dijkstra's algorithm. *Implementing Discrete Mathematics: Combinatorics and Graph Theory with Mathematica*, Reading, MA: Addison-Wesley, 225-227.
- Spitzer, B. L., Henderson, K. A., & Zivian, M. T. (1999). Gender differences in population versus media body sizes: A comparison over four decades. *Sex roles*, 40(7-8), 545-565.
- Su, J. G., Winters, M., Nunes, M., & Brauer, M. (2010). Designing a route planner to facilitate and promote cycling in Metro Vancouver, Canada. *Transportation research part A: policy and practice*, 44(7), 495-505.
- Sorton, A., & Walsh, T. (1994). Bicycle stress level as a tool to evaluate urban and suburban bicycle compatibility. *Transportation Research Record*, 17-17.
- Stinson, M., & Bhat, C. (2003). Commuter bicyclist route choice: Analysis using a stated preference survey. *Transportation Research Record: Journal of the Transportation Research Board*, (1828), 107-115.
- Strauss, J., Miranda-Moreno, L. F., & Morency, P. (2015). Mapping cyclist activity and injury risk in a network combining smartphone GPS data and bicycle counts. *Accident Analysis & Prevention*, 83, 132-142.
- Tilahun, N. Y., Levinson, D. M., & Krizek, K. J. (2007). Trails, lanes, or traffic: Valuing bicycle facilities with an adaptive stated preference survey. *Transportation Research Part A: Policy and Practice*, 41(4), 287-301.
- Usyukov, V. (2013). Development of a Cyclists' Route-Choice Model: An Ontario Case Study.
- Vandenbulcke, G., Thomas, I., de Geus, B., Degraeuwe, B., Torfs, R., Meeusen, R., & Panis, L. I. (2009). Mapping bicycle use and the risk of accidents for commuters who cycle to work in Belgium. *Transport Policy*, 16(2), 77-87.

- White, C. E., Bernstein, D., & Kornhauser, A. L. (2000). Some map matching algorithms for personal navigation assistants. *Transportation research part c: emerging technologies*, 8(1), 91-108.
- Winters, M., Babul, S., Becker, H. J. E. H., Brubacher, J. R., Chipman, M., Crompton, P., ... & Monro, M. (2012). Safe cycling: How do risk perceptions compare with observed risk. *Can J Public Health*, 103(9), S42-7.
- Yai, T., Iwakura, S., & Morichi, S. (1997). Multinomial probit with structured covariance for route choice behavior. *Transportation Research Part B: Methodological*, 31(3), 195-207.
- Yang, C., & Mesbah, M. (2013, October). Route choice behavior of cyclists by stated preference and revealed preference. In *Australasian Transport Research Forum 2013 Proceedings*.
- Yen, J. Y. (1971). Finding the k shortest loopless paths in a network. *Management Science*, 17(11), 712-716.
- Yiannakoulias, N., Bennet, S. A., & Scott, D. M. (2012). Mapping commuter cycling risk in urban areas. *Accident Analysis & Prevention*, 45, 164-172.
- Zeng, W., & Church, R. L. (2009). Finding shortest paths on real road networks: the case for A*. *International journal of geographical information science*, 23(4), 531-543.

Appendix

I Tools and Functions Used in Data Pre-Processing and Map-matching

Name of Tools of Functions	Description	Software	Processing Step
ST_MakePoint	This function is used to created 2D, 3DZ or 4D point geometry based on latitude, longitude, altitude.	PostGIS	Convert latitude, and longitude GPS observations into geometries
ST_StartPoint	Return the first point of a linestring.	PostGIS	Extract the start point as nodes from network
ST_EndPoint	Return the last point of a linestring	PostGIS	Extract the end point as nodes from network
ST_Transform	Change the coordinating system of a geometry by the srid specified in parameters.	PostGIS	Change SRID 4326 to 900913 to calculate distance or length in meter
ST_Length	Return the 2D length of a linestring with the unit of the coordinating system. SRID 900913 has the unit of meter.	PostGIS	Calculate length of a link in network
ST_Reverse	Return the geometry with vertex order reversed.	PostGIS	Create two-direction network.
ST_Distance_Sphere	Calculate linear distance in meters between two latitude/longitude points	PostGIS	Calculate the distance between transit stop points and link in the network.
Erase Point	Delete points outside or inside of the feature.	ArcGIS	Delete points outside the study area; delete points that are arbitrary
Add Surface Information	Add spatial information derived from a surface.	ArcGIS	Assign the elevation in DEM onto the links in the bikeway
Spatial Join	Join two features by spatial relationship	ArcGIS	Check whether parallel on-street

	specified one with one. If one feature can be joined more than one target features, the target features would be aggregated first.		parking is allowed on a link.
Buffer	Create round buffer polygons for geometry by specific distance.	ArcGIS	Create buffers for edges or intersections of network
Select Layer By Location	Select features in the layer by the spatial relationship specified by another layer.	ArcGIS	Select GPS observations inside the intersection buffers
Copy Feature	Copy the selected features in a layer to a new layer.	ArcGIS	Copy the selected observations inside the intersection buffers into another layer.
Select	Select features in a layer by SQL queries.	ArcGIS	Separate correctly map-matched observations and multiple joined observations.
Merge	Combine multiple datasets into one dataset.	ArcGIS	Combine the observations near intersections and the multiple joined observations together into one dataset.
ST_Azimuth	Return the radian angle from horizontal of the first point to the second point clock-wisely.	PostGIS	Calculate the direction of the network and separate the trips into two directions

II Statistic Summary for Commute Trips and Recreational Trips

Personal Information		Number of Commute Trips (33253)	Percentage of Commute Trips (%)	Number of Recreational Trips (9611)	Percentage of Recreational Trips (%)
Age	No selection	2429	7.30	823	8.56
	18 – 24	605	1.82	213	2.22
	25 – 34	14482	43.55	4278	44.51
	35 – 49	11703	35.19	2838	29.53
	50 – 64	3653	10.99	1329	13.83
	> 65	381	1.15	130	1.35
Household Income	No selection	5142	15.46	1667	17.34
	< \$20,000	735	2.21	387	4.03
	\$20 – 39,000	1670	5.02	677	7.04
	\$40 – 59,000	3701	11.13	1231	12.81
	\$60 – 79,000	3596	10.81	1192	12.40
	\$80 – 99,000	4693	14.11	1249	13.00
	> \$100,000	13716	41.25	3208	33.38
Gender	No selection	17231	51.82	5023	52.26
	Male	3680	11.07	1137	11.83
	Female	12342	37.12	3451	35.91
Rider History	no selection	2086	6.27	762	7.93
	Since Childhood	18537	55.75	5656	58.85
	Several Years	7220	21.71	1832	19.06
	One Year or Less	5008	15.06	1284	13.36
	Just Start	402	1.21	77	0.80
Comfort Level	no selection	2388	7.18	888	9.24
	Not comfortable sharing roadways with vehicles	5410	16.27	1275	13.27
	Only comfortable sharing roadways in clearly designated cycling facilities	3963	11.92	1264	13.15
	Comfortable cycling in most roadway conditions	21492	64.63	6184	64.34
Winter Cyclist	No selection	2533	7.62	832	8.66
	Yes	15874	47.74	4802	49.96
	No	14846	44.65	3977	41.38

III Map-matching Arbitrary Observation (PL/pgSQL Script)

--map matching the arbitrary point by estimating the nearby link to it

```
CREATE OR REPLACE FUNCTION map_matching_arbitrary(limit_s integer, offset_num
integer)
```

```
RETURNS void AS $$
```

```
DECLARE
```

```
coords numeric[];
```

```
tempcoord numeric;
```

```
temptripid numeric;
```

```
temptime timestamp;
```

```
templinksb numeric[];
```

```
templinksa numeric[];
```

```
templinks numeric[];
```

```
finallink numeric;
```

```
i integer;
```

```
BEGIN
```

```
    SELECT INTO coords
```

```
        array_agg(coord_id)
```

```
    FROM (SELECT DISTINCT coord_id FROM observation_arbitrary ORDER BY coord_id
```

```
LIMIT $1 OFFSET $2) AS subquery;
```

```
    RAISE NOTICE '%', array_length(coords,1);
```

```
    FOREACH tempcoord IN ARRAY coords LOOP
```

```
        RAISE NOTICE 'WOKRING ON %', tempcoord;
```

```
        SELECT INTO temptripid, temptime
```

```
            trip_id, recorded_at
```

```
        FROM observation_arbitrary
```

```
        WHERE coord_id = tempcoord;
```

```
        SELECT INTO templinksb
```

```
            array_agg(link_id)
```

```
        FROM
```

```
        (SELECT * FROM observation_correct
```

```

WHERE trip_id = temptripid
AND recorded_at < temptime
ORDER BY recorded_at desc
LIMIT 10) AS table1;
SELECT INTO templinksa
            array_agg(link_id)
FROM
(SELECT * FROM observation_correct
WHERE trip_id = temptripid
AND recorded_at > temptime
ORDER BY recorded_at asc
LIMIT 10) AS table1;
templinks = templinksb || templinksa;
SELECT INTO finallink
            table1.unnest
FROM (SELECT DISTINCT unnest, count(*) AS number
      FROM unnest(templinks)
      GROUP BY unnest
      ORDER BY number DESC LIMIT 1) AS table1;
RAISE NOTICE '%', finallink;
UPDATE observation_arbitrary SET link_id = finallink WHERE coord_id =
tempcoord;
END LOOP;
END;
$$ LANGUAGE plpgsql;

```

IV Trip Generation (PL/pgSQL Script)

```
-- generate raw trips table and network trips table by estimating the passing links
CREATE OR REPLACE FUNCTION generate_trips(limit_s integer, offset_num integer)
RETURNS void AS $$
DECLARE
trip_ids numeric[];
temp_trip_id numeric;
passing_links numeric[];
num_links integer;
i integer;
j integer;
start_time timestamp;
end_time timestamp;
max_direction float;
pass_link_dir_id integer;
distance_raw float;
velocity_raw float;
BEGIN
    SELECT INTO trip_ids
        array_agg(trip_id)
    FROM (SELECT DISTINCT trip_id FROM map_matching ORDER BY trip_id LIMIT $1
OFFSET $2) AS subquery;
    RAISE NOTICE '%', array_length(trip_ids,1);
    FOREACH temp_trip_id IN ARRAY trip_ids LOOP
        RAISE NOTICE 'WORKING ON %', temp_trip_id;
        SELECT INTO passing_geom
            array_agg(geom)
        FROM (SELECT DISTINCT geom, recorded_at FROM map_matching WHERE
trip_id = temp_trip_id ORDER BY recorded_at ASC) AS subquery1;
        SELECT INTO distance_raw
            sum(distance)
```

```

FROM map_matching WHERE trip_id = temp_trip_id;
SELECT INTO passing_links
        array_agg(link_id)
FROM (SELECT DISTINCT link_id, recorded_at FROM (SELECT DISTINCT
ON(link_id) link_id, recorded_at FROM may_map_matching WHERE trip_id = temp_trip_id
ORDER BY link_id, recorded_at ASC) AS table1 ORDER BY recorded_at ASC) AS
subquery2;
SELECT INTO start_time
        recorded_at
FROM map_matching WHERE trip_id = temp_trip_id ORDER BY recorded_at ASC
LIMIT 1;
SELECT INTO end_time
        recorded_at
FROM map_matching WHERE trip_id = temp_trip_id ORDER BY recorded_at
DESC LIMIT 1;
num_links = array_length(passing_links,1);
FOR m IN 1..(num_links) LOOP
    SELECT INTO max_direction
            MAX(direction)
    FROM map_matching
    WHERE trip_id = temp_trip_id
    AND link_id = passing_links[m];
    IF max_direction < 1.5708 THEN
        SELECT INTO pass_link_dir_id
                link_dir_id
        FROM network_directions
        WHERE link_id = passing_links[m]
        AND ((direction >= 0 AND direction <= max_direction + 1.5708) OR
(direction >= max_direction + 4.71239 AND direction <= 6.28319));
    ELSE
        IF max_direction > 4.71239 THEN

```

```

        SELECT INTO pass_link_dir_id
                link_dir_id
        FROM network_directions
        WHERE link_id = passing_links[m]
        AND ((direction >= 0 AND direction <= max_direction - 4.71239) OR
(direction >= max_direction - 1.5708 AND direction <= 6.28319));
    ELSE
        SELECT INTO pass_link_dir_id
                link_dir_id
        FROM network_directions
        WHERE link_id = passing_links[m]
        AND direction >= max_direction - 1.5708 AND direction <=
max_direction + 1.5708;
    END IF;
END IF;
passing_links[m] = pass_link_dir_id;
END LOOP;
distance_raw := 0;
distance_network := 0;
FOR i IN 1..(num_coords-1) LOOP
    distance_raw = distance_raw + ST_Distance_Sphere(passing_geom[i],
passing_geom[i+1]);
END LOOP;
FOR j IN 1..num_links LOOP
    distance_network = distance_network + ST_Length((SELECT geom::geometry FROM
network_directions WHERE link_dir_id = passing_links[j]),true);
END LOOP;
volectricity_raw = distance_raw/EXTRACT(EPOCH FROM (end_time - start_time));
volectricity_network = distance_network/EXTRACT(EPOCH FROM (end_time -
start_time));
RAISE NOTICE 'volectricity_network is %', volectricity_network;

```

```
        INSERT INTO network_trips(trip_id, start_time, end_time, distance_raw,  
velocity_raw, pass_links) VALUES (temp_trip_id, start_time, end_time, distance_raw,  
volectricity_raw, passing_links);  
    END LOOP;  
END;  
$$ LANGUAGE plpgsql;
```

V Dijkstra's Shortest Path Generation with Binary Heap (C# Codes)

//Dijkstra algorithm has 5 inputs: graph contains the edges and nodes of the network; start_node and end_node are the start and end points; generalized_cost is the string showing which column should be used as minimized cost; extra_params contains some extra information like gender, trip time and other information that may be used to calculate generalized cost for each trip.

```
public List<int> Dijkstra(Graph G, int start_node, int end_node, string generalized_cost, params
int[] extra_params)
```

```
{
```

```
    //dijkstra_result contains the result of the shortest path node id
```

```
    List<int> dijkstra_result = new List<int>();
```

```
    //previous_nodes contains the result of the previous node id and the current node id
```

```
    //the first int represents the current node id and the next int represents the previous node id
```

```
    Dictionary<int, int> previous_nodes = new Dictionary<int, int>();
```

```
    //Step 1: add the end node id into the dijkstra_result
```

```
    dijkstra_result.Add(end_node);
```

```
    //priority queue contains the queue in the Dijkstra Algorithm
```

```
    //priority queue use binary heap to increase efficiency
```

```
    PriorityQueue queue = new PriorityQueue();
```

```
    //the previous node of the start_node is -1 means invalid
```

```
    previous_nodes.Add(start_node, -1);
```

```
    //at first, the label of the first node is zero, so enter the first node into the queue with a
weight of zero
```

```
    Tuple<int, double> queue_item = new Tuple<int, double>(start_node, 0.0);
```

```
    G.vertices[start_node] = 0.0;
```

```
    queue.Insert(queue_item);
```

```
    //do the iteration until the queue is empty
```

```
    while (queue.Count != 0)
```

```
    {
```

```
        //selected the node with the minimum label value from the queue
```

```
        Tuple<int, double> selected_node = queue.Extract_Minimum();
```

```
        if (G.edges.ContainsKey(selected_node.Item1))
```

```

{
    //selected the neighbour vertices of the selected node
    Dictionary<int, double> end_nodes = G.edges[selected_node.Item1];
    //check the label of every neighbour vertices
    foreach (var reached_end_node in end_nodes)
    {
        //if the label of the neighbour vertices is larger than the calculated one, then
        //replace the label of the neighbour vertices
        //the label vertices should be changed in both G and queue
        //link_length as generalized cost use this
        double reached_end_node_value = reached_end_node.Value;
        if (generalized_cost == "Energy_Saving")
        {
            reached_end_node_value = reached_end_node.Value * (extra_params[0]
            == 1 ? 98.3 : (extra_params[0] == 2 ? 84.7 : 90.7));
        }
        Double    total_vertice_label    =    selected_node.Item2    +
        reached_end_node_value;
        if (G.vertices[reached_end_node.Key] > total_vertice_label)
        {

            if                (queue.Check_Exist(new                Tuple<int,
            double>(reached_end_node.Key, G.vertices[reached_end_node.Key])))
            {
                Tuple<int,    double>    old_node    =    new    Tuple<int,
                double>(reached_end_node.Key,
                G.vertices[reached_end_node.Key]);

                Tuple<int,    double>    new_node    =    new    Tuple<int,
                double>(reached_end_node.Key, total_vertice_label);
                queue.Update(old_node, new_node);
            }
        }
    }
}

```

```

    }
    else
    {
        queue.Insert(new Tuple<int, double>(reached_end_node.Key,
total_vertice_label));
    }
    G.vertices[reached_end_node.Key] = total_vertice_label;

    if (previous_nodes.ContainsKey(reached_end_node.Key))
    {
        previous_nodes[reached_end_node.Key] =
selected_node.Item1;
    }
    else
    {
        previous_nodes.Add(reached_end_node.Key,
selected_node.Item1);
    }
}

}

}

}

while (previous_nodes[end_node] != -1)
{
    dijkstra_result.Add(previous_nodes[end_node]);
    //G.edges[previous_nodes[end_node]][end_node] *= 1.5;
    end_node = previous_nodes[end_node];
}
dijkstra_result.Reverse();

```

```

        return dijkstra_result;
    }

```

Binary Heap

```

public class PriorityQueue
{
    //int is the node id and double is the label of the node
    List<Tuple<int, double>> queue = new List<Tuple<int, double>>();
    public int Count
    {
        get { return queue.Count; }
    }
    public Tuple<int, double> Minimum
    {
        get { return queue.First(); }
    }
    public void Clear()
    {
        queue.Clear();
    }
    public void Insert(Tuple<int, double> insert_item)
    {
        int position = Count;
        queue.Add(insert_item);
        while (position > 0 && queue[(position-1) / 2].Item2 > insert_item.Item2)
        {
            queue[position] = queue[(position-1) / 2];
            position = (position-1) / 2;
        }
        queue[position] = insert_item;
    }
}

```

```

public Tuple<int, double> Extract_Minimum()
{
    Tuple<int, double> result = Minimum;
    if (queue.Count == 1)
    {
        queue.Clear();
    }
    else {
        //remove the minimum root of the priority queue
        Tuple<int, double> last_item = queue[Count - 1];
        queue.RemoveAt(Count - 1);
        int i = 0;
        while (i < queue.Count / 2)
        {
            int j = (2 * i) + 1;
            if ((j < queue.Count - 1) && queue[j].Item2 > queue[j + 1].Item2)
            {
                ++j;
            }
            if (queue[j].Item2 >= last_item.Item2)
            {
                break;
            }
            queue[i] = queue[j];
            i = j;
        }
        queue[i] = last_item;
    }
    return result;
}

public bool Check_Exist(Tuple<int, double> check_item)

```

```

{
    return queue.Contains(check_item);
}

public void Update(Tuple<int, double> remove_node, Tuple<int, double> update_node)
{
    //Perform the decrease key of the min-heap
    int delete_index = queue.IndexOf(remove_node);
    queue[delete_index] = update_node;
    //Perform the heap filter up function
    while (delete_index > 0 && queue[(delete_index - 1) / 2].Item2 > update_node.Item2)
    {
        int parent = (delete_index - 1) / 2;
        queue[delete_index] = queue[parent];
        delete_index = parent;
    }
    queue[delete_index] = update_node;
}
}

```

VI Tables of Correlations among Candidate Factors for Commute and Recreational Trips

Correlations among Factors for Commute Trips

Trip Proportion with Parallel Parking			
Trip Proportion without Bicycle Facilities			
Trip Proportion with Cycle Tracks			
Trip Proportion with Multi-use Pathway			
Trip Proportion with Bike Lane			
Trip Proportion with Shared Lane			
Trip Proportion with Local Streets			
Trip Proportion with Collector			
Trip Proportion with Minor Arterial			
Trip Proportion with Major Arterial			
Average Traffic Volume			
Number of Transit Stops			
Energy Consumption			
Trip Distance			
Total Precipitation			
Mean Temperature			
Gender			1
Rider History		1	-0.03847
Age	1	0.24991	0.100855
Commuting Trips (33222)	Age	Rider History	Gender

					1
				1	0.057173
			1	0.186594	-0.15473
		1	0.145947	0.342803	-0.08505
		-0.23731	-0.20825	-0.10103	0.576333
1	-0.42446	-0.3511	-0.39236	-0.13028	-0.18616
0.871761	-0.05681	-0.31388	-0.50568	-0.15138	0.041946
0.409861	0.001224	-0.24309	-0.30282	-0.11081	-0.04964
-0.09222	-0.12532	-0.04157	-0.0663	-0.01481	-0.09804
-0.05788	-0.11933	-0.05579	-0.09469	0.017211	-0.09203
-0.0134	0.010513	-0.00044	0.006507	-0.00277	0.008683
0.034472	0.042533	-0.05972	-0.03906	-0.00956	0.072789
-0.08017	0.008036	0.067337	0.046157	0.087028	-0.03094
0.07615	-0.02266	0.043382	-0.05133	-0.01392	0.034008
-0.07103	0.014931	0.032483	0.067078	0.014039	0.006138
Trip Proportion with Major Arterial	Trip Proportion with Minor Arterial	Trip Proportion with Collector	Trip Proportion with Local Streets	Trip Proportion with Shared Lane	Trip Proportion with Bike Lane

			1
		1	-0.07717
	1	-0.19684	-0.00476
1	-0.08569	-0.3444	-0.27656
-0.22883	0.063335	-0.60924	0.288194
-0.06646	-0.05658	-0.41441	0.078283
-0.11922	-0.07525	0.037293	0.116681
-0.09664	-0.07705	-0.0979	-0.08519
-0.24773	0.235274	-0.25101	0.248554
-0.30214	-0.04962	0.43375	-0.04579
-0.5726	0.033712	0.421142	0.111575
-0.10755	0.10713	0.154922	-0.1474
0.412124	0.016275	-0.14473	-0.37413
0.401266	0.011457	-0.16217	-0.33293
-0.00015	0.003932	-0.00678	0.021065
-0.01599	0.006196	-0.03759	0.067354
0.030166	0.00634	-0.03535	-0.04339
-0.06156	0.018869	0.011183	-0.04433
0.00819	0.02008	-0.01153	-0.05991
Trip Proportion with Multi-use Pathway	Trip Proportion with Cycle Tracks	Trip Proportion without Bicycle Facilities	Trip Proportion with Parallel Parking

Correlations among Factors for Recreational Trips

Trip Proportion with Parallel Parking				
Trip Proportion without Bicycle Facilities				
Trip Proportion with Cycle Tracks				
Trip Proportion with Multi-use Pathway				
Trip Proportion with Bike Lane				
Trip Proportion with Shared Lane				
Trip Proportion with Local Streets				
Trip Proportion with Collector				
Trip Proportion with Minor Arterial				
Trip Proportion with Major Arterial				
Average Traffic Volume				
Number of Transit Stops				
Energy Consumption				
Trip Distance				
Total Precipitation				
Mean Temperature				1
Gender			1	-0.06133
Rider History		1	-0.07238	-0.0327
Age	1	0.211402	0.103056	-0.0672
Recreational Trips (9611)	Age	Rider History	Gender	Mean Temperature

					1
				1	-0.23818
			1	0.032715	-0.1254
		1	0.09144	-0.18254	-0.21047
	1	-0.00941	0.295614	0.003578	-0.1793
1	-0.17484	-0.2048	-0.04151	0.543944	-0.24742
-0.33861	-0.23762	-0.31464	-0.04292	-0.09876	-0.36254
0.037191	-0.1181	-0.36169	-0.02291	0.133562	-0.65072
0.071735	-0.11327	-0.24494	-0.01401	0.044375	-0.00639
-0.1181	-0.07555	-0.16351	-0.03249	-0.11222	0.482177
-0.10975	-0.07638	-0.17238	-0.02578	-0.09788	0.457068
-0.01024	0.00757	0.001727	0.016734	-0.01288	-0.0359
0.015219	-0.0315	-0.06452	0.015206	0.00476	0.095403
-0.04626	0.02712	0.068194	-0.005	-0.03116	-0.02073
0.046588	0.073264	-0.09033	0.022778	0.061916	-0.03942
0.020118	-0.00096	0.041903	-0.02383	-0.00105	-0.02195
Trip Proportion with Minor Arterial	Trip Proportion with Collector	Trip Proportion with Local Streets	Trip Proportion with Shared Lane	Trip Proportion with Bike Lane	Trip Proportion with Multi-use Pathway

		1
	1	0.158403
1	-0.14115	-0.00557
-0.04552	-0.49058	-0.41201
0.06519	-0.49462	0.167669
-0.05555	-0.31738	0.115812
-0.12468	0.223249	0.251203
-0.06921	-0.0608	-0.03051
0.255502	-0.16738	0.199315
-0.03574	0.411861	0.049049
0.044023	0.457284	0.230189
0.065106	-0.017	-0.16148
-0.00262	-0.28394	-0.37408
-0.00711	-0.28396	-0.35774
0.00303	0.036937	0.026194
0.018322	-0.08	-0.01849
-0.03463	0.047298	0.032857
0.02648	-0.04034	-0.04224
0.012687	0.020249	0.027528
Trip Proportion with Cycle Tracks	Trip Proportion without Bicycle Facilities	Trip Proportion with Parallel Parking