Development of a Cyclists' Route-Choice Model: An Ontario Case Study

by

Vladimir Usyukov

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

Waterloo, Ontario, Canada, 2013

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 presents the first North American route-choice model for cyclists developed from a large sample of GPS data. These findings should encourage all interested municipalities to implement cycling as part of their transportation planning by determining key designing and planning factors to encourage cycling. The analysis is based on processing *revealed preference* data obtained from 415 self-selected cyclists in Waterloo, Ontario, which corresponded to 2000 routes. Cyclists' route decisions were modeled using multinomial logit framework of discrete choice theory. The main finding involved in capturing two different behaviour groups, namely experienced and inexperienced cyclists. This was subsequently reflected in the two developed models. The key factors impacting route-choice were found to be trip length, speed, volume, bicycle lane presence and percent of uphill gradient that cyclists face. The predictive power of the best model was 65%. The outlier analysis found that the relative significance of uphill gradient coefficient in one circumstances and perhaps the exclusion of unobserved variables, in other circumstances could be the cause why probability of actual choice was not predicted by both models all the time.

In addition, this research involved in the development of a transferability study involving route-choice modeling for cyclists. The analysis is based on the revealed preference data obtained from 255 self-selected cyclists in Peel Region, Ontario, which corresponded to 425 unique routes. The choice set contained actual routes and a combination of alternatives obtained by labeling and impedance rules. The transferability of Waterloo's model to Peel Region was 37%. This means that cyclists behaviour in the Peel Region can be predicted correctly by travel length, bicycle lane presence and percent of uphill gradient for every third cyclist.

Acknowledgements

The author would like to express the warmest gratitude to the main research investigator, Professor Jeffrey Casello of University of Waterloo for the guidance of research and financial support during the research time.

In addition, the author would like to extend appreciation to Professor Jeffrey Newman of Northwestern University, for allowing to use choice modeling software Easy Logit, as well as Professor Susan Tighe and Professor Dipanjan Basu for their valuable contribution to improve the manuscript.

The author would like to thank to Kyrylo Rewa, Steve Xue and Erica Springate in their efforts during data collection stages.

Personally I would like to thank to Victor, Natasha, Maria M., Jason L., Aleli and Pancho, Philip, Daniel H., Tae, Debra, Samantha, Sergey K. and many others for helping out during the various stages of the study period at Waterloo. At last Olga G., for the mental inspiration of getting the study done rapidly.

In Dedication
To my family

Table of Contents

List of figures:

List of tables:

1 Chapter 1: Introduction to the Route-Choice Problem

1.1 Background

One of the fundamental reasons why people travel relates to the fact that different activities exist in different places. Quality of life directly relates to transportation. In economics, this type of demand is known as "derived demand" as people do not travel for the sake of traveling; rather, they travel to participate in activities such as work, school or leisure. Usually, the transportation network of a typical city provides a large number of choices for traveling between any two places. Since there is more than one route between an origin and destination, every trip involves a route, a choice and a route choice decision. This chapter borrows from the work of Bovy and Piet, which will be referenced throughout the chapter, accordingly.

One of the best ways to describe these terms is presented in Figure 1-1. The beginning of the node is known as the origin and the ending node as the destination. A *route* is a chain of consecutive nodes connected by road infrastructure segments, also known as "paths". *Choice set* can be defined as a list of all possible routes between an origin and destination that a traveler will evaluate before making a decision. In this example, our choice set consists of four routes, or paths.

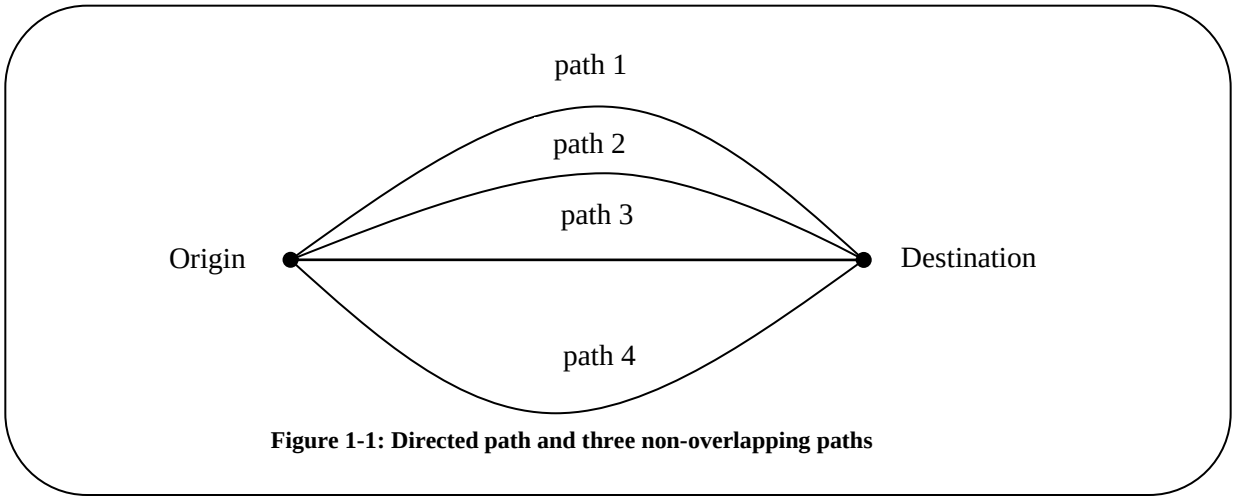


Figure 1-1: Directed path and three non-overlapping paths

In real life, some routes can be unique but the majority will have some kind of partial overlap with competing paths. The reason for having overlap is to provide redundancy in the number of road segments and node combinations a traveler can travel. In the context of this research, route-choice becomes a problem, as there are many alternatives between origins and destinations. According to Miller et al. (2001), "transportation is the aggregate of thousands, or in many cases millions of individual trip-making decisions". As a result, a traveller is faced with a multitude of route-choice decisions, taking the form of individual trips through a limited capacity transportation network. The best way to understand route-choice problem is by studying the traveler's behaviour in the network through the spatial choice they made (Bovy et al., 1990). This study focuses on route choices and attempts to answer the following questions:

1) What influences people in how they choose their routes?

2) What information do they have?

3) What road characteristics play a determinant role?

To answer these questions, some forms of quantitative models have been proposed and evaluated. Typically these models are aimed at predicting the use of routes, dependent on the routes' and travelers' characteristics (Bovy et al., 1990).

At the same time, route-choice is not an individual problem. In the context of a multi-modal transportation network, the public, in general, is interested in the best organization of all movements of people and goods in a transport system. At a multi-modal scale, route choice can be made by using all modes, including motorized and non-motorized, as they can be integrated into one transport system, as shown in Figure 1-2. On a smaller scale, cycling through the network can also be considered as a route-choice problem.

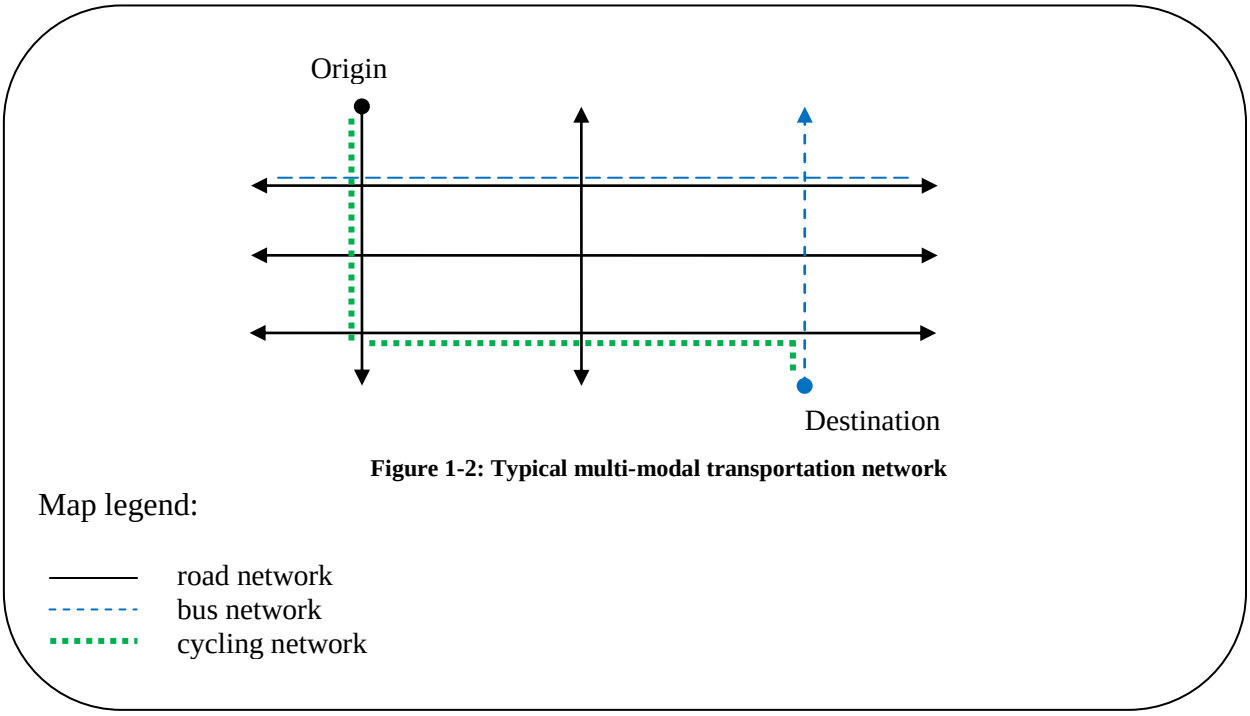


Figure 1-2: Typical multi-modal transportation network

Route-choice problems can be explained as the paradigm of complex human-environment relationships where route choices comprise only one component of a broader area of travel behaviour, which mostly depends on two factors:

- The traveler, with his or her subjective needs, experiences, preferences, perceptions, etc;
- The physical environment, with its objective opportunities and their characteristics (Bovy et al., 1990).

The process of making a route decision, with which any traveler is faced, can be structured rationally in a diagram, as shown in Figure 1-3. The complex human-environment interaction starts from the point when a traveler decides to take a trip between any two places. At the same time, the transportation system offers a large number of travel choices, but most of these alternatives may overlap with each other to a certain degree. The original choice set can be

regarded as all *possible* existing route alternatives. Usually, this choice set contains all possible alternatives, in spite of the fact that some of the alternatives are realistic and others are not.

Although the knowledge of a traveler is limited and the number of all possible routes is infinite, a traveler's choice set can be considered to begin from a set of *known* alternatives. A traveler will use previous experience, knowledge, and some specific constraints, and often the list of all known alternatives can be further refined to a set of *available* alternatives. A typical example of when a set of known alternatives is reduced to an available set, is when a traveler sets a certain cost constraints that preclude from accessing a destination for which travel costs are exceeded. Typical cost constraints are time, money, and distance. The list of all available alternatives will make up the choice set of a traveler. This list will contain a set of alternatives from which a traveler chooses according to particular circumstances. At this stage, the traveler will try to acquire all the required information about each route.

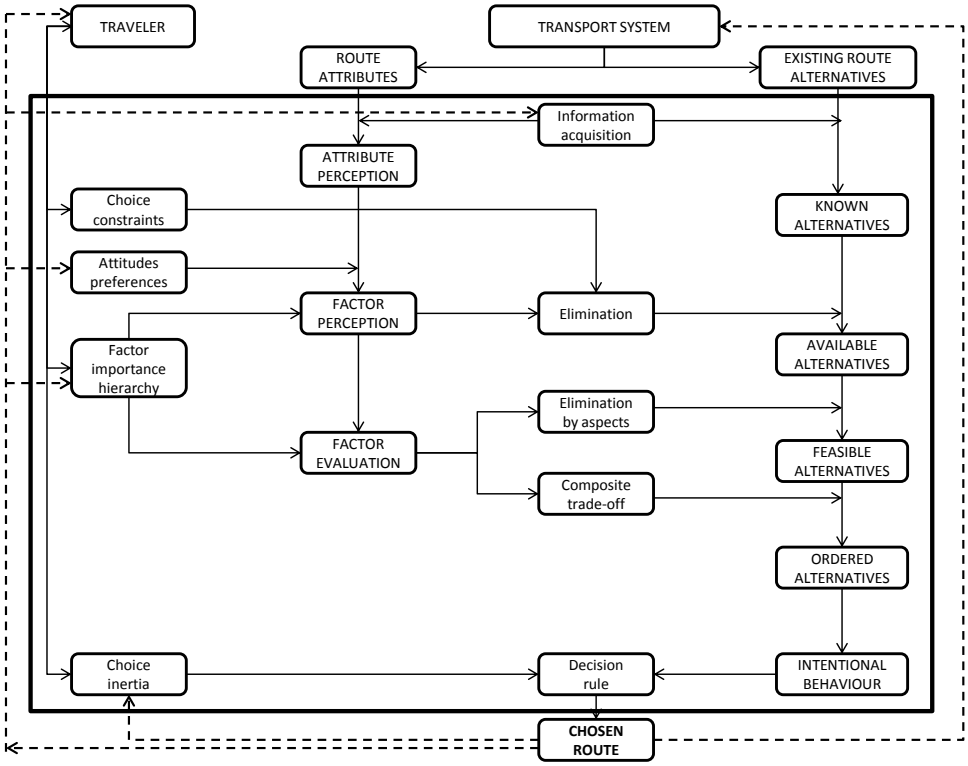


Figure 1-3: Elements of individual choice behaviour (Bovy et al, 1990)

Two assumptions are generally made about the traveler's behaviour. These assumptions are related to rationality and self-interest and are important in describing the decision-making process during the evaluation of alternatives. In regards to rationality, travelers in their own subjective opinion will perceive and measure the route characteristics of every path in the choice set. In regards to self-interest, travelers will try to optimize the route based on their constraints. However, not every route characteristic will be equally weighted by travelers, as their perception can be considered incomplete and inaccurate. Therefore from a traveler's perspective, some route characteristics will be more important or compensate for the others, depending on the relative importance that a traveler gives to them: high or low value. For example, a traveler who is under time constraints may consider it worthwhile to pay tolls in order to avoid possible time delays and thus getting to a meeting on time; another traveler may tolerate an inconvenience linked with transfers of using public transit because the other option is driving an automobile, which they may not have.

The choice set of all available alternatives will be further refined based on the *factor importance hierarchy*. A traveler will examine each alternative according to aspects in the available choice set and may find some alternatives to be unreasonable. After applying *elimination by an aspects filter*, a traveler will evaluate every remaining alternative and refine the choice to a group of *feasible* alternatives. At this stage in the process, the traveler will carry out a more thorough assessment in which a trade-off is made among the counter-balancing characteristics of every route. In econometric notation, a traveler applies a utility function, which reflects the relative importance of each aspect. Then, this traveler will put the feasible alternatives in the order of importance to create *ordered* alternatives. Once all the alternatives are ordered, a traveler, based on their reasoning, makes a final decision about which route to use. One example of such

reasoning can be to minimize the cost, time, distance, number of transfers; or maximize the number of sightseeing activities, if it is a trip for pleasure; other rules are possible too. It has to be noted that inertia will play an important role, as certain minimum tolerances need to be crossed before a traveler will change the habit of a certain type of behaviour. Furthermore, continuous feedback is gained through learning and using the system so that travelers' decisions can be assessed frequently.

The structure presented above is a simplified approach to describing highly complicated human-environment relationships in which every individual is unique. This difference can be explained through two filters: perception and evaluation (Bovy et al, 1990). Through the perception filter, each individual perceives subjectively the attributes of each route, and through the evaluation filter, each alternative is transformed into a desirability scale, which is linked to the travelers experiences, preferences, or constraints. The result of travellers, in this case cyclists, being affected by different perceptions, cognitions, emotions, learning, experiences, cognitive attitudes, and evaluation parameters when making a route choice suggests that route choice is a very individual matter and therefore cannot be reduced only to socio-economic and road characteristics. Each individual is different from others in regards to how these filters perceive and evaluate the problem, as "different individuals may make the same decisions, that is, choose the same route, though on different grounds" (Bovy et al, 1990), and this is what is observed and modeled using utility functions.

Route-choice models are considered of high importance and used in transportation to:

- Predict route choice dependent on the routes' and travelers' characteristics;
- Help in designing and re-designing transportation facilities;

- Assess travelers' reactions to proposed network changes;
- Assess the amount of excess travel, caused directly from route selection criteria.

In regards to the first area of application, utility models are developed for most significant modes in a city and used by every transportation agency to model various transportation scenarios. For example, utility functions can be used to find a modal split (the number of trips made by various modes) or to find the relative utility of selecting a particular transportation mode. Utility models can also be used to predict market segmentation, or the share of travelers selecting a particular route, based on road and traveler characteristics.

Route-choice models can also help in designing new and re-designing old transport facilities. Since route-choice models are quantitative models, that is based on the function of route and traveler characteristics, they can predict which road characteristics are significant. For example, do cyclists prefer a separated cycling facility? To what extent do cyclists tolerate high speed or high volume roads? Should the bicycle lanes be build along roads with a steep gradient? Are they going to be used? Overall, if the cyclists' preferences are known, planners can design cycling facilities in a way that attracts users.

In network analysis, route-choice models can be used to determine which routes travelers will choose if a road section is blocked or congested for a long time. In a similar manner, the sensitivity of travelers to road pricing through tolls on a road can be observed through spatial differences and flow distributions. Overall, these models provide the ability to evaluate traveler attitudes with respect to changes in a transportation system.

Finally the study of excess travel is also worthwhile. Excess travel can be defined as the difference in distance, time, or any other type of cost criteria between the chosen route and a route proposed by an optimal choice, like the shortest path. Excess travel causes inefficient use of limited natural resources, and individual and public time by not using optimal routes, and is a cause of congestion. According to one study, the cost of congestion is estimated to cost 12 trillion dollars globally (Rybarczyk et al., 2010).

1.2 Research Motivation

This research is motivated by the following:

1. The public interest as well as every level of government supports moving towards a more sustainable way of living. Cycling represents one of the most sustainable transportation modes offering numerous benefits. Currently, 48% of all trips in the USA are shorter than 4.5 km (3 miles) in length; some of these trips can be done by cycling (Pickrell and Schimek, 1998). Therefore the potential for cycling in urban environments is tremendous.
2. Currently route-choice models for cyclists are either at rudimentary stages of development or non-existent at all, which is the case in Canada and the USA. This thesis describes novel models from actual choices recorded by GPS tracks and it explains major determinants for cycling. These results can be used in strategic planning by transportation agencies.
3. From a scientific point of view, this research focuses on performing a transferability study and evaluating models derived in the Region of Waterloo for the neighbouring Region of Peel. According to the literature review, done up to December 2012, this kind of study has not been done before and is the first of its kind.

1.3 Research Goals and Objectives

There are two main goals:

- Propose and validate route-choice models, using GPS data obtained from cyclists in the Region of Waterloo, Ontario;
- Conduct a transferability study and evaluate the applicability of models to the Region of Peel.

More specific objectives are:

- To aggregate necessary socio-economic, GIS and GPS cycling track data in one database for the Region of Waterloo and Peel Region;
- To compare the characteristics of actual choice and the shortest path to determine key variables and relationship between variables;
- To propose and implement a method for generating a choice set with possible automation as a standalone program in GIS;
- To develop and automate a methodology for extracting road characteristics from a given dataset as a standalone program;
- To model cyclists choices using the logit framework, perform a thorough statistical validation of the results, and draw conclusions.

1.4 Thesis Outline

The thesis is structured in the following way. Chapter 1 reviews the theoretical background of the route choice problem and examines how it is applied in the transportation field. Chapter 2 reviews the literature related to the route choice of cyclists. Chapter 3 describes the study methodology and steps taken to develop route-choice models. Chapter 4 describes a first case study, that is to develop cyclist route-choice models in Waterloo Region. Chapter 5 describes a

second case study, evaluating the transferability of Waterloo's route-choice models to Peel Region. Chapter 6 summarizes the main findings and proposes recommendations.

2 Chapter 2: Literature Review

2.1 Introduction

In recent years noticeable changes have occurred in society, as people have become more concerned about the degradation of the environment due to human activity. As a result of these concerns, some people have become more aware of sustainable practices in their personal lives. In transportation, hybrid and electric-powered vehicles have appeared; airplanes, ships and rail vehicles have become more efficient; more sophisticated routings and deliveries of goods have been designed; and, a recurring interest in non-motorized modes, including cycling and walking, has risen. The literature review in this chapter will focus on cycling, with special attention paid to the development of route-choice models, as such models can explain and predict the number of users on a cycling facility as well as inform decisions about what type of infrastructure to build, which road characteristics play a role, and how travelers' behaviour can be altered to encourage cycling.

The popularization of active modes in North America can be attributed to two Acts passed in the United States in the 1990s: Intermodal Surface Transportation Efficiency Act (ISTEA) and the successor Transportation Equity Act (TEA-21). Both Acts were intended to allocate government funds towards non-highway projects, with a "clear motive to reduce congestion and improve air quality" (TEA-21) as most of the North American cities were known to be auto-dependent and could not be considered sustainable. The following facts highlight unsustainable living: in 2001 84% of trips in the United States are made by auto, which makes municipalities auto-dependent, not liveable (City of Toronto Bicycle Plan, 2001); 48% of all trips by all modes are shorter than five km, some of these trips can instead be carried out by non-motorized modes, but people still

prefer to drive (Pucher et al., 1999); the global cost of congestion is estimated to be 12 trillion dollars (Rybarczyk et al., 2010).

The benefits of cycling are numerous and some of them are described herein. In urban areas cycling is the fastest mode of transportation for distances up to ten kilometers; riding a bicycle for short trips can save riders 18-24 cents per km (City of Toronto Bicycle Plan, 2001). Also, it is a very energy-efficient and an inexpensive way to travel; there are no direct emissions, except at the manufacturing level (Heinen et al., 2010). The cost of infrastructure, in comparison to other modes, is comparatively low. For example, the cost range to design and build a road in Toronto is $350,000 - $500,000 per km; the cost of a 1.5 m bicycle lane per 1 km can fluctuate between $5,000 and $15,000 for restriping (City of Toronto Bicycle Plan, 2001). There are also numerous health benefits, as two-thirds of Canadians lead sedentary lifestyles and cycling could be part of daily exercise regimes (City of Toronto Bicycle Plan, 2001).

Most municipalities in Canada and the United States have become interested in supporting sustainable ways of living. Cycling in particular has become a part of many transportation master plans. For example, the Region of Waterloo and City of Toronto adopted cycling master plans back in the late 1990s, with a long-term vision of doubling the number of trips in a decade. The Region of Waterloo concentrated on building a 732 km cycling network and allocated 33M dollars to spend on the plan over the next 20 years (Region of Waterloo Cycling Master Plan, 2004). The City of Toronto plans to build a 1000 km cycling network, and for that purpose, allocated 73M to spend over a ten year period in order to double the number of trips by 2011 (City of Toronto Bicycle Plan, 2001). Despite these efforts, the cycling mode share is still very low and contributes to only 1.3% of total travel in Canada (City of Toronto Bicycle Plan, 2001)

and 0.9% in the US (Sener et al., 2009). One of the objectives of this research is to identify the key factors that determine cycling behaviour and cycling route choices, so that decision-makers can build infrastructure wisely and increase ridership. In the scientific context, an understanding of cycling route choices and cycling behaviour can be obtained through the use of utility models, which are also known as route-choice models. These models were described in Chapter 1. Models specific to cycling are explained in the following section.

2.2 Current and past practices used to estimate route-choice models: revealed and stated preference surveys

Modeling of route choice and travelers' behaviour has been explored extensively for automobile and public transit modes. However, these models cannot be considered relevant to describing the route choice behaviour of a cyclist, because certain factors, such as the presence of a cycling lane, or riding along a busy road or uphill, can strongly influence cyclists, but have no effect on automobile drivers.

Most observational studies which focus on modeling route choice behaviour can be categorized into one of two groups, depending on the data collection method. These groups are *stated* and *revealed preference* surveys. The stated preference type of survey is an approach used to obtain data from laboratory experiments, including computer simulation. The findings of revealed preference survey allows the researcher to use real-world observations collected from actual trips. The relative advantage of using a stated preference approach is in the controlled nature of the choice scenarios. The freedom of controlling choice scenarios allows the researcher to use a carefully designed structure for the experiments, where alternatives and their attributes can be predetermined and then compared across individuals (Bradley et al., 1986). This type of approach is fairly inexpensive, and even with a small group of individuals, researchers can obtain

multiple responses. However, results obtained with stated preference surveys largely depend on a defined set of possible alternatives and perceptions of attributes by individuals (Bovy et al., 1990). The main drawback of this type of survey is that it may not be known if responses from surveys correspond to actual choices of travelers in a similar situation.

Stated preference surveys were used before in a number of projects to determine the key factors that influence cycling route choice behaviour. One of the first comprehensive studies was done in the Netherlands (Ministry of Transport and Public Works, 1987). The study found cyclists to be most sensitive to distance and travel time (Bovy et al., 1990). A similar study from the US confirmed that travel time and distance were the most important factors for commuter cyclists (Stinson et al., 2003). Among other factors, the presence of a cycling facility and the presence of a cycling facility on a bridge received a lot of attention from researchers (Stinson et al., 2003). A group of scientists from Seattle confirmed that separated bicycle lanes, as well as socio-economic characteristics, like age, gender and income, were significant in explaining route-choice behaviour (Shafizadeh et al., 1993). At the same time, the above mentioned study from the Netherlands found, to a certain extent, contradictory results: cycling facility type was of a lesser importance, in comparison to the quality of paved surfaces (Bradley et al., 1986). A separate study evaluating bicycle-transit interface confirmed that experienced users in comparison to inexperienced cyclists were rather indifferent to the type of bicycle facility (Mahmassani et al., 1996). As cyclist's level of experience increases, cycling on a roadway becomes less burdensome (Hunt et al., 2006). In regards to the level of automobile traffic volume and safety, these factors were also considered significant in explaining cyclists route choice behaviour (Bovy et al., 1990). In separate studies done in Canada and the United States, the significance of traffic volume was confirmed (Stinson et al., 2003, Hunt et al., 2006). To some

extent, differing results were found in the Netherlands: traffic volume was found of a lesser importance, in comparison to other variables like travel distance or time (Bradley et al., 1986). Other factors like road gradient received a little attention, except for a study done in the United states (Stinson et al., 2003). At the same time *a priori* knowledge of cycling behaviour suggest that cyclists tend to prefer flatter rather than hilly roads and therefore gradient should be an important factor. Among other factors, pavement quality was considered significant (Bradley et al., 1986); socio-economic and demographic attributes generated very ambiguous results and therefore a recommendation was to avoid using them (Heinen et al., 2010). The complete list of factors found in the literature review is presented in Table 2-2, provided at the end of this chapter.

Revealed preference surveys offer the possibility of using actual trips to relate the most significant factors in travelers' route choice behaviour to road characteristics. This type of data collection is not considered new in transportation, as actual trips were recorded in the past, either by following a traveler, or by asking a traveler to draw a map with the taken path. In order to use revealed preference data, four challenges must be overcome. First is the generation of relative alternatives, despite the fact that relative alternatives are not always known. Ideally, one would need to generate a choice set distribution for each traveler, which is not practical for a real-life project. Second, alternatives in the choice set must be different from one another and satisfy independence from irrelevant alternatives condition of discrete-choice theory (IIA). For these reasons, adequate statistical verification is necessary to validate data. Third, researchers need to know the attributes of each route. Lastly, a large data sample is necessary. Most of these challenges were addressed in previous studies, allowing researchers to utilize revealed preference surveys.

Modern GPS technologies permit researchers to obtain actual trip data of great accuracy and volume, which has made data collection problems, like the cost or accuracy of a survey, problems of the past. Despite the ubiquitous proliferation of communication technologies into the daily lives of people, the application of communication technologies to bicycle route choices are found in a handful of studies only.

One of the pioneering studies using actual trips to model a cyclist's route-choice, was conducted by Altman-Hall (Altman-Hall, 1996). In the research, data was collected using hand-drawn maps and the generated choice set was then evaluated using a multinomial logit framework. The study attempted to relate a large number of variables to the actual route choice of travelers. Road variables, including type of road (arterial, collector, minor), type of cycling facility, speed limits, traffic volume, gradient, direction of travel, bridge and railway crossings, number of turns and turns at signals, as well as a range of socio-economic characteristics were examined. The major findings were the following: cyclists tend to avoid gradients, gradient-separated railway crossings and high-activity areas (Altman-Hall, 1996). In addition, the study was able to capture two types of behaviour: one group of cyclists (experienced) preferred to travel along the shortest route, even if it coincided with an arterial type of road; the second group (inexperienced) preferred travelling longer routes through residential neighbourhoods, which is consistent with the perception of safety. Since this study was one of the first to apply revealed preference to model cycling behaviour, certain limitations were found within it. The gradient variable did not stand out strongly because the direction of travel was not recorded during the survey; most of the calibrated models had weak statistical quality; and, most importantly, the predictive power of the models was not provided for peer review. Some of these limitations can be explained by the

novelty of the research, the small size of the sample and possibly the high variability within the sample itself.

A more recent study was performed by a group of researchers in Switzerland and is recognized as the "first route-choice model for bicyclists estimated from a large sample of GPS observations" (Axhausen et al., 2010). Cyclists were found to be very sensitive to trip length, the presence of a cycling facility and gradient. The study explored non-linearity in the parameters' of multinomial family of models, for which Box-Cox transformation was necessary. The significance of the model parameters was found to be quite high, and the elasticity of variables with respect to trip length was evaluated. However, certain data limitations precluded researchers from adding a road volume variable as a part of the model structure. The predictive power of the models was not provided for peer review. The transferability of models to other regions was not explored either.

A third study was performed in Portland, Oregon, and its findings are currently implemented into the region's travel forecasting system (Dill et al., 2011). Several findings were found to be consistent with those of other studies, as cyclists were found to be sensitive to trip length, gradient, traffic volume, the presence of a cycling facility and turn frequency. The predictive power of the models was not provided for peer review. The transferability of models to other regions was not explored either.

Lastly, a study carried out in Phoenix used hand-drawn maps, collected from a sample of cyclists in order to analyse commuter routes (Howard et al., 2001). The study was very restricted in its findings, as no statistical measures were provided for review. Most of the research was based on comparing actual routes to the shortest paths.

Our research uses a combination of the stated and revealed preference data, obtained from 415 self-selected cyclists in the Region of Waterloo, Canada. The evaluation of the stated preference data was carried out by a group of researchers in Waterloo, in 2011. The analysis of the stated preference data can be summarized in several findings, as follows: "cycling supports the travel demand of participants' in all age groups", with half of the travelers being older than 40; on average, cyclists' household income was found to be higher than the Regional average; 40 percent of cyclists indicated that when cycling was not an option, their primary mode was an auto (Casello et al., 2011). These results suggest that respondents are choice cyclists as they had an option to drive but opted for cycling.

With regard to the revealed preference data, GPS tracks of self-selected cyclists were recorded, and the initial analysis of tracks proposed an empirical relationship to differentiate recreational from utilitarian trips. In addition, several important findings were established between the design of the transportation network, the built environment and observed cycling patterns. It was found that a grid street network with the addition of trails offered a better directness of travel by minimizing the excess travel distance of cyclists. In particular, the presence of trails reduced the excess travel of cyclists by more than 18% (Casello et al., 2012). This research continues studying the nature of revealed preference data in order to develop models able to determine key factors in cyclists' route-choices.

2.3 Major factors influencing cycling route choice behaviour

In general, the main factors that influence route choice behaviour have been identified by Bovy and are summarized in Table 1. These factors can be categorized into four groups which describe: a traveller, in terms of socio-economic characteristics; a route, in terms of topological

attributes; a trip, in terms of purpose and mode selection, and other circumstantial parameters. Despite the great variability in the decision-making processes of travelers, route-choice modeling is mainly concerned with a route and a traveler's characteristics (Bovy et al., 1990). The most important factors influencing cycling route choice behaviour were explored in a number of stated and revealed preference studies, and the summary is provided in Table 2-1. Table 2-2 looks specifically at literature review and impacts on cycling, performed by two groups of researchers. Most of the factors can be found in their work (Hunt et al., (2006) and Heinen et al., (2010)).

The factors selected for evaluation in this study are trip length, traffic speed, traffic volume, presence of cycling infrastructure along the route and percentage of positive gradient that cyclists face. Individually these factors have received a lot of attention but together they have not. The only study that evaluated length, speed and gradient in the same modeling framework was done in Switzerland (Axhausen et al., 2010). However, none of the studies referenced in the literature review have a validation component stated explicitly for the peer review. As a result, there is no understanding of the predictive power of the obtained models. Finally, no studies have been found that perform transferability of their model parameters to other regions or countries.

Table 2-1: Factor attributes that influence route-choice (Bovy et al., 1989)

CHARACTERISTICS of the		ATTRIBUTES			
		General	**Effort-related**	**Comfort-related**	**Others**
TRAVELER		age, sex, life cycle, income level, education, household structure, Race, profession, length of residence, no. of drivers in family, years having driving license, no. of cars in family			
ROUTE	Road	type of road width, length, no. of lanes, angularity, intersections, bridges, slopes	travel time, travel cost	road surface, waiting time	speed limits, law enforcement
	Traffic	traffic composition, traffic density, in traffic flow, in counter flow, in cross flow, travel speed	congestion, access in/out, no. of turns, stop signs, traffic lights, pedestrian	noise nuisance, lighting, signposting, parking at destination	direct charges/toll, parking along the road, safety & probability of accident, reliability and variation in travel time
	Environment	aesthetics, building type, building density, land use along route, scenery	crossing, easy pick-up/drop-off		security, crowdness, privacy
TRIP		trip purpose, time budget, pressure, time of the trip, no. of travelers, mode used			
CIRCUMSTANCES		weather conditions, day/night, accident en route, emergencies, road and traffic information			

Table 2-2: Summary of factors compiled from two studies (Abraham and Hunt (2006), Heinen et al. (2010))

Factor	References
Facility Characteristics	
Type of facility (whether mixed with traffic, bicycle lane, or bicycle path)	Antonakos (1994); Aultman-Hall (1996); Axhausen and Smith (1986); Bradley and Bovy (1984); Calgary (1993); Copley and Pelz (1995); Goldsmith (1996); Guttenplan and Patten (1995); Harris and Associates (1991); Kroll and Ramey (1977); Kroll and Sommer (1976); Landis and Vattikuti (1996); Lott et al. (1978); Mars and Kyriakides (1986); Nelson and Allen (1997); Sacks (1994); Taylor and Mahmassani (1997)
Nature of shared roadway, including road class, sight distances, turning radii, lane/median configurations	Aultman-Hall (1996); Calgary (1993); Copley and Pelz (1995); Davis (1995); Denver (1993); Epperson (1994); Landis and Vattikuti (1996); Mars and Kyriakides (1986); Shepherd (1994); Sorton (1995); Sorton and Walsh (1994)
Existence of on-street parking	Davis (1995); Epperson (1994); Mars and Kyriakides (1986); Stinson and Bhat (2003, 2005);
Pavement surface type or/and quality	Antonakos (1994); Axhausen and Smith (1986); Bradley and Bovy (1984); Davis (1995); Epperson (1994); Landis and Vattikuti (1996)
Gradients	Antonakos (1994); Axhausen and Smith (1986); Davis (1995); Stinson and Bhat (2003); Rietveld and Daniel (2004); Rodriguez and Joo (2004); Parking et al. (2008); Hunt and Abraham (2007)
Intersection spacing and/or configuration	Aultman-Hall (1996); Davis (1995); Epperson (1994); Teichgraeber (1982)
Cycling treatments at signals, including timing and detection	Copley and Pelz (1995)
Completeness and directness of cycling infrastructure	Ambrosius (1984); Copley and Pelz (1995); Sacks (1994)
Availability of showers at origin or/and destination	Guttenplan and Patten (1995); Sacks (1994); Taylor and Mahmassani (1997)
Availability of secure parking for bicycle at origin or/and destination	Calgary (1993); Copley and Pelz (1995); Denver (1993); Guttenplan and Patten (1995); Mars and Kyriakides (1986 Sacks (1994); Taylor and Mahmassani (1997); Wynne (1992)
Network layout	Moudon et al. (2005), Zacharias (2005)
Continuity of cycling facilities	Stinson and Bhat (2003, 2005)
Non-cycle traffic characteristics	
Motor vehicle speeds and driver behaviour	Antonakos (1994); Davis (1995); Epperson (1994); Landis and Vattikuti (1996); Mars and Kyriakides (1986); Sorton (1995); Sorton and Walsh (1994)
Volume or mix of motor vehicle types, including proportion of trucks	Antonakos (1994); Axhausen and Smith (1986); Bradley and Bovy (1984); Calgary (1993); Davis (1995); Epperson (1994); Landis and Vattikuti (1996); Mars and Kyriakides (1986); Sorton and Walsh (1994)
Pedestrian interaction	Mars and Kyriakides (1986)
Age	Antonakos (1994); Aultman-Hall (1996); Sacks (1994); Taylor and Mahmassani (1997); Treadgold (1996); Pucher et al. (1999), Moudon et al. (2005), Zacharias (2005), Dill and Voros (2007)
Safety concerns	Antonakos (1994); Kroll and Ramey (1977); Kroll and Sommer (1976); Lott et al. (1978); Mars and Kyriakides (1986)
Level of cycling experience	Antonakos (1994); Axhausen and Smith (1986); Sorton and Walsh

Individual and trip characteristics	
Gender	Antonakos (1994); Aultman-Hall (1996); Sacks (1994); Taylor and Mahmassani (1997); Räsänen and Summala (1998), Banister and Gallant (1999), Pucher et al. (1999), Howard McDonald and Burns (2001), Dickinson et al. (2003), Krizek et al. (2004), Rietveld and Daniel (2004), Rodríguez and Joo (2004), Moudon et al. (2005), Plaut (2005), Ryley (2006), Dill and Voros, (2007)
Income	Taylor and Mahmassani (1997); Pucher et al. (1999), Stinson and Bhat (2005), Dill and Voros (2007); Witlox and Tindemans (2004), Plaut (2005), Schwanen and Mokhtarian (2005), Guo et al. (2007); Dill and Carr (2003), Zacharias (2005)
Private vehicle ownership	Sacks (1994); Cervero (1996), Kitamura et al. (1997), Banister and Gallant (1999), Stinson and Bhat (2004, 2005), Plaut (2005), Pucher and Buehler (2006), Dill and Voros (2007), Guo et al. (2007), Parkin et al. (2008), Stinson and Bhat (2004); Moudon et al. (2005)
Perceived social norm	Bruijn et al. (2005); Bamberg and Schmidt (1994)
Personal security concerns	Sacks (1994); Pucher et al. (1999), Rietveld and Daniel (2004), Lohmann and Rölle (2005), Southworth (2005)
Flexibility of work hours	Denver (1993); Sacks (1994)
Employment status	Boumans and Harms (2004)
Trip length by time or distance	Bradley and Bovy (1984); Calgary (1993); Guttenplan and Patten (1995); Parajuli 1996); Parajuli et al. 1996), Parkin et al. (2007), Timperio et al. (2006), Stinson and Bhat (2004), Dickinson et al. (2003); Hunt and Abraham (2007)
Environmental/situational characteristics	
Weather, season, temperature, rain	Calgary (1993); Stinson and Bhat (2004); Guo et al. (2007); Bergstrom and Magnussen (2003); Brandenburg et al. (2004); Nankervis (1999);
Sweeping/Snowplowing	Copley and Pelz (1995)
Nature of abutting land uses	Axhausen and Smith (1986); Davis (1995); Epperson (1994); Landis and Vattikuti (1996)
Aesthetics along route	Antonakos (1994); Sacks (1994)
Degree of political and public support for cycling	Clarke (1992); Copley and Pelz (1995); Wynne (1992)
Level of public assistance for cyclists, including maps, route advice and emergency aid	Denver (1993)
Education and enforcement regarding cycling	Antonakos (1994); Denver (1993); Wynne (1992)
Cost and other disincentives to use other modes	Moritz (1997); Sacks (1994); Taylor and Mahmassani (1997); Wynne (1992)

3 Chapter 3 - Methodology

3.1 Introduction

The main research goals were focused on determining, validating and assessing transferability of main factors to cycling using GPS data obtained from self-selected cyclists in the Regions of Waterloo and Peel using econometric modeling. Note that self-selected cyclists are cyclists who decided to participate in this research. These goals were achieved through a methodological implementation of more specific objectives, such as gathering of transportation and socio-economic data, generation of a feasible set of alternatives, automation data referencing of thousands of route segments to road attributes, performing rigorous statistical and predictive tests and lastly determination of main factors to cycling and drawing conclusions.

The process workflow diagram is presented in Figure 3-1, and each one of the components is described in the individual section.

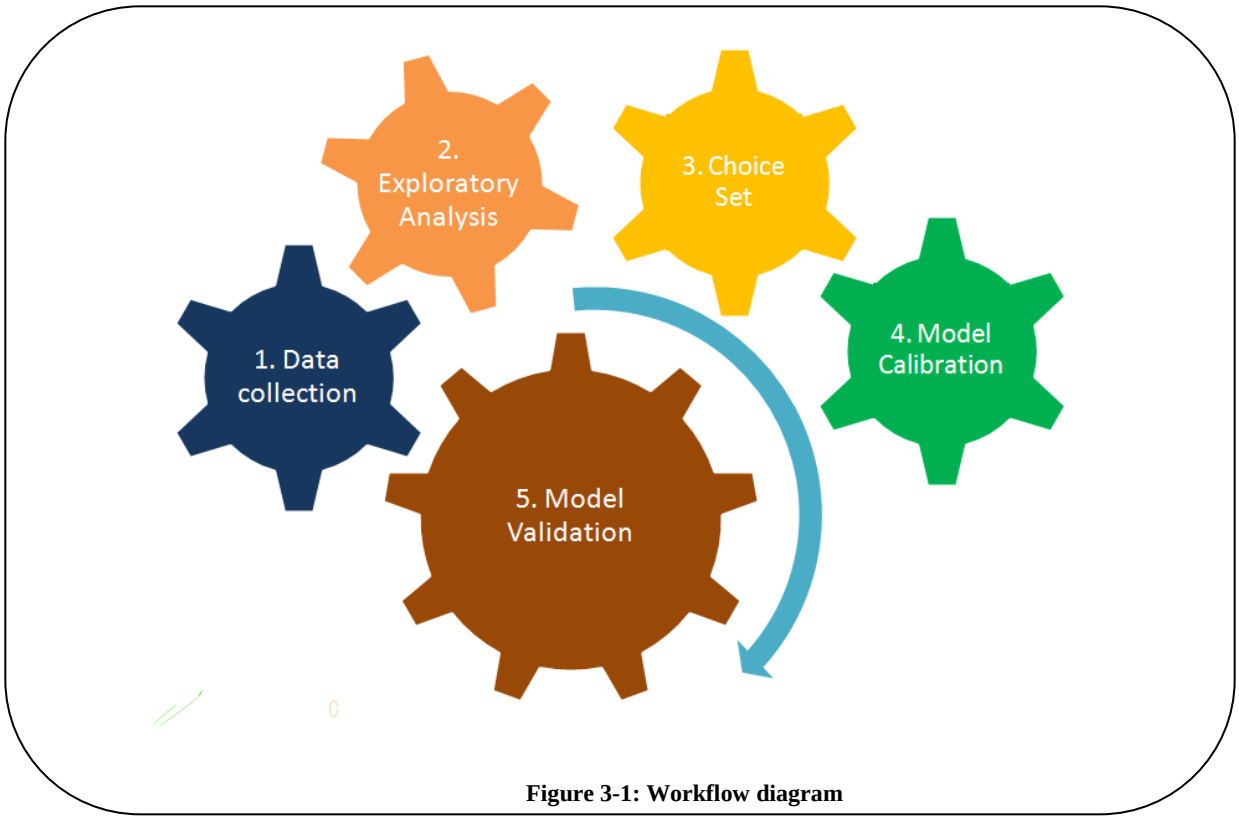


Figure 3-1: Workflow diagram

The entire route-choice workflow can be thought of as a sequential execution of the following steps: gather data, perform exploratory analysis of collected data, generate a choice set of feasible alternatives, perform model calibration and validate results. A brief explanation of each component follows, and the detailed explanation is presented in the individual sections of this chapter.

The first step in the process was gathering of the data. Data required for route-choice modeling included GPS routes obtained from 415 self-selected cyclists, and a range of attributes related to road links of transportation network. The following road attributes were considered significant in the choice selection of every cyclist: trip length, traffic volume, traffic speed, presence of cycling facilities and percent of uphill gradient.

The second step was to perform an exploratory analysis of gathered data. The most important part of this analysis was to build correlation matrices for each pair of data attributes. Exploratory analysis focused mainly on two objectives: first, was to validate the assumption of their linear relationship between selected road variables; second, was to validate that selected road attributes can enter into the constant part of a utility function. Most of this analysis was performed in the statistical package SPSS.

The third step was generation of a feasible choice set, which was the vital component of route-choice modeling. In total four alternatives were obtained for each observed route. This activity was performed in GIS using a set of heuristic rules. As a result, a table was obtained. The data, contained within the Table, needed to undergo a quality check and then aggregated into routes through a specially written computer program. The mathematical component of the algorithm used in the program is covered in section 3.4.2.

The fourth step was to model calibration using discrete choice theory: random utility approach. Multinomial logit was selected as the modeling framework to model cyclists' route choices. The calibration of model parameters was performed using a program called Easy Logit.

The fifth step was model validation, which included a range of intuitive, statistical and predictive tests. One of the most important components was to test the predictive power of each model on the dataset that was not seen by the model. Outlier analysis was also important as it allowed a better understanding of why the actual choice was not predicted by the model.

3.2 Data collection

The variables considered for the modeling of cyclists route-choice are presented in Table 3-1. The data describing OD and the actual path for each participating individual was gathered through the earlier research (Rewa, 2012). GIS data was available through the University of Waterloo, Canada. Subsequently, variables considered for the model estimation (spatial and descriptive), were combined into the one large geo-spatial database.

Table 3-1: Data inputs for the model

Data type	Description	Source
Actual path	The 'actual path' was obtained from GPS receivers installed on cyclists bicycles participated in the study, measured in meters.	Region of Waterloo
OD	Origin-destination pairs which represent the beginning and the ending nodes of a trip.	Region of Waterloo
Volume	Volume variable corresponds to PM peak hour traffic, measured in vehicles per hour.	Region of Waterloo
Speed	Speed variable corresponds to posted speed limit on each link, measured in km/hr	Region of Waterloo
Bicycle Lane	$\begin{cases} 1 & presence\ of\ bike\ facility, \\ 0 & otherwise \end{cases}$	Region of Waterloo
Length	Length of a route, measured in metres.	Region of Waterloo
% of Uphill Gradient	Gradient variable, measured in per cent (%). Data precision of DEM was specified at ± 0.5 meters	University of Waterloo

3.3 Exploratory data analysis

The initial exploratory data analysis was performed in 2012 and it should be noted a reader is referred at a reference section (Rewa, 2012). For our research, the main purpose of the additional exploratory analysis had two objectives:

a. Evaluate dependence between variables, selected for model building;
b. Establish threshold for recreational trips;

Evaluation of variables dependence was necessary to understand that a linear form of the parameters in the model configuration was suitable as the modeling framework. Most of the variables dependencies can be analysed through scatter plots matrices using the statistical software SPSS.

The purpose of establishing a threshold for recreational trips was also necessary in order to eliminate recreational trips from the further analysis. Recreational trips had very different characteristics from the utilitarian trips and this hypothesis was confirmed in the earlier study (Rewa, 2012). Recreational trips were considered to be trips when the ratio of the actual trip length to the shortest trip length exceeded certain established thresholds. Since most of the alternative generation algorithms are based on finding optimal paths, only the utilitarian trips were considered suitable for our research. The threshold for recreational trips is provided in Table 3-2.

Table 3-2: Recreational thresholds

Actual trip length	Ratio	% of trips
2 - 5 km	3.5	0.8
5 - 7.5 km	3.0	0.9
7.5 - 10.0 km	2.25	1.4
> 10 km	1.5	5.3

The recreational thresholds can be explained in the following way. If the shortest path distance between OD pairs was 2 km, for example, then any actual trip longer than 7 km [2 x 3.5 = 7 km] was considered to be a recreational trip. If the shortest path distance between the OD pair was 9 km, then any trip longer than 20 km [2.25 x 9 = 20.25 km] was also considered a recreational trip.

3.4 Modeling framework - discrete choice theory: random utility approach

This section is largely indebted to the fundamental concepts developed by Moshe Ben-Akiva and Steven Lerman (1985), and Daniel McFadden (1974, 1976, 1977). The theories behind discrete choice theory are well-developed in the field of microeconomics (consumer theory), where a choice made by a traveler reveals preferences towards a certain type of infrastructure or product. The framework to model the route-choice problem itself was performed through the application of *random utility approach*.

The choice theory can be regarded as a collection of procedures that defines the following four elements (Ben-Akiva et al., 1985):

- Decision-maker;
- Alternative generation;
- Attributes of alternatives;
- Decision rule;

A decision-maker is represented by an individual or a group, who is faced with different choice situations. Typically in the model, the decision-maker can be characterized by a vector of different socio-economic characteristics. Then, a decision maker is faced with a choice, and the choice can be made only from a set of alternatives, known as a *choice set*. Choice is defined through a set of procedures, known as *alternative generation*. Next, attributes of each alternative

need to be collected. Lastly, a decision maker evaluates alternatives according to some decision rule. There exists a variety of rules, including dominance, satisfaction, lexicographic and utility. This research is based on the utility rule.

The probabilistic choice theory is considered as the most appropriate method to model these processes. The reason for that is the following: it reflects the human behaviour and it explains behavioural inconsistencies better than deterministic utility approach, as it offers a more sensitive analysis to forecast the travel demand. For example, a probabilistic choice theory can model a scenario where two individuals are faced with an identical choice set, defined by the identical characteristics, can select different alternatives. Instead of selecting an alternative with the highest utility, the decision maker is assumed to behave with choice probabilities. Choice probabilities are defined by a probability distribution function over alternatives that include utilities as parameters (Ben-Akiva et al., 1985).

3.4.1 Generation of choice set

Ideally, it would be necessary to use a probability model that can generate a feasible choice set for every individual in the sample. The choice set of feasible alternatives would tell the likelihood of a particular route to be a part of the choice set, depending on the traveler's and transportation network's characteristics (Bovy et al., 1985). The size of the choice set, as well as the composition of attributes, need to be considered, because they affect model estimates (Bliemer et al., 2008). The choice set needs to have routes that are realistic as well as routes that are not realistic, as long as cyclists are aware of them. However, there is no benchmark with which to compare the amount of variation between alternatives (Altman-Hall, 1996).

Most of the procedures used to generate alternatives are based on heuristic rules, because there is no scientific basis to select one method over the other. The main idea behind all alternative generation methods is to propose the most feasible paths by meeting IIA property of the logit models. IIA property is concerned with duplication and correlation among alternatives. The following objectives needed to be met:

a. Alternatives needed to be based on sound heuristic rules;
b. Alternatives needed to be distinct (IIA property);

The choice set generation methods can be classified into *deterministic* shortest path-based algorithms, *stochastic* shortest path-based algorithms, *constrained enumeration* methods and *probabilistic* methods. An excellent literature review on this subject was performed by Carlo Prato (2009). Most of these methods are based on some type of constraint which needed to be optimized. The deterministic shortest path-based family can be classified into the following branches: *K shortest path, labeling approach, link elimination and link penalty*. *K shortest path* algorithm is based on the repeated generation of the shortest path method and is the most readily available to researchers through GIS software. *Labeling approach* was proposed by Ben-Akiva, and the objective of the labeling approach is to determine a set of criteria for a traveler's choice set. In mathematical notation, each label corresponds to an impedance function specified by each criterion. The possible list of criteria can be defined by minimum time, minimum distance, minimum number of traffic lights, most scenic route, most signposted route, minimum congestion, maximum road quality, etc. The closest study found to use this method was the research by Altman-Hall (Altman-Hall, 1996). In that study, it was proposed that a set of criteria be used to generate a choice set for cyclists. *Link elimination* method is based on the repetitive search for the shortest path after the removal of a part or of all the shortest path links from the

previous searches. The research conducted in Switzerland proposed to use link elimination method to generate alternatives (Axhausen et al., 2010). Lastly, *link penalty*, is based on the repetitive search of the shortest path, but a certain penalty is imposed on every link instead of a link removal. The information related to other classes can be found in Prato's literature review.

The following heuristic rules were considered to determine the choice set for model calibration. The chosen path must be in the choice set because it represented the chosen path of cyclists. The second alternative corresponded to the shortest path applied to the road network. This alternative represents a feasible alternative without cycling facilities. The third alternative corresponded to the shortest path within the transportation network that including cycling facilities. This alternative represented improvements done to the transportation network when cyclists have the opportunity to use available cycling facilities. The fourth alternative was created by imposing the distance penalty on paths created for the second alternative. In this way unique paths were found, which were distinct from paths created for the second alternative and which were a feasible option to cyclists. The fifth alternative was generated by imposing a distance penalty on the paths created for the third alternative. In this way unique paths were found which were distinct from the paths created for the third alternative, and which were a feasible option to cyclists.

Summary of generated alternatives:

- Actual path - is the actual path taken by a cyclist.
- Shortest path by the road network.
- Shortest path by the road and trails network.
- Second shortest path by the road network: impose a cost of 1.2
- Second shortest path by the road and trails network: impose a cost of 1.2

An example of O-D pair with five paths is presented in Chapter 4 (section 4.4.1).

3.4.2 Attributes of alternatives

GIS was used as one of the main tools to determine the characteristics of each route. In GIS, every link in a route is related to the table of attributes. The process of connecting individual road links to the table of attributes is known as *attribute relation*. Since most of the chosen paths had a "zigzag" travel pattern, a threshold for the search radius, within which attributes can be traced, had to be established.

The result of relation procedure created tables of data, also known as the turn tables. In order to aggregate individual links from the turn tables into routes, data was downloaded to ASCII file format and then summarized through a specially written computer program. The algorithm behind the program contained a set of instructions which is provided next.

Suppose that a cyclist's route can be represented schematically as the movement from "O" to "D", as demonstrated in Figure 3-2. Overall, the OD route consists of three road links 1-3 and each road link has a particular set of attributes. Segments OA and BD are uphill in the direction of travel; link AB is downhill and also has a bicycle lane. Additionally, for each link the following

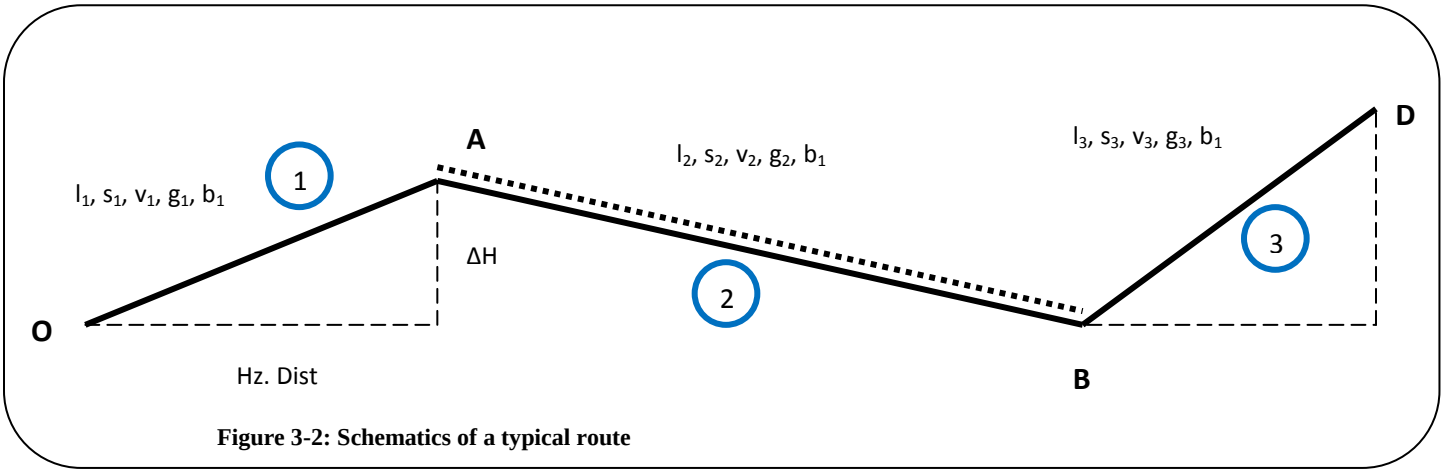


Figure 3-2: Schematics of a typical route

five attributes were obtained: trip length, traffic speed, traffic volume, percent of uphill gradient and a presence of bicycle lane. The objective was to aggregate the data from the individual links into OD route. Below is the legend that describes each component individually.

O,D - origin and destination nodes
A, B - intermediate nodes
l - length of a link, km
s - posted speed limit on a link, m/s
v - car volume on a link, veh/hr
g - percent of uphill gradient of a link, %
b - presence or absence of a bike lane, 0 or 1

The gradient variable for each road segment can be calculated using simple geometry. Since positive gradients were of interest only, a gradient for the link OA is obtained as follows:

- Change in elevation, ΔH, is given by:
  Elevation (point A) - Elevation (point O) eq.3.1
- The horizontal distance change over which ΔH occurs is given by:
  Hz. Dist = $(OA^2 - \Delta H^2)^{1/2}$ eq.3.2
- Finally, the percent of uphill gradient is given by:
  OA = 100x(ΔH/ Hz. Dist) eq.3.3

The instructions used to aggregate attributes along the length of a route are:

- $Total\ length\ of\ route\ OD\ in\ km, L = l_1 + l_2 + l_3 = \sum_{i=1}^{3} l_i$ eq.3.4
- $Weighted\ speed, W_{speed} = s_1 \times \frac{l_1}{L} + s_2 \times \frac{l_2}{L} + s_3 \times \frac{l_3}{L} = \frac{1}{L} \times \sum_{i=1}^{3} l_i \times s_i$ eq.3.5
- $Weighted\ volume, W_{volume} = v_1 \times \frac{l_1}{L} + v_2 \times \frac{l_2}{L} + v_3 \times \frac{l_3}{L} = \frac{1}{L} \times \sum_{i=1}^{3} l_i \times v_i$ eq.3.6
- $Weighted\ bike\ lane, W_{bike\ lane} = 0 \times b_1 \times \frac{l_1}{L} + 1 \times b_2 \times \frac{l_2}{L} + +0 \times b_3 \times \frac{l_3}{L} = b_2 \times \frac{l_2}{L}$ eq.3.7
- $Weighted\ gradient, W_{uphill\ gradient} = g_1 \times \frac{l_1}{L} + 0 \times g_2 \times \frac{l_2}{L} + g_3 \times \frac{l_3}{L} =$

$$= 100\% \times \frac{1}{L} \times (g_1 \times l_1 + g_2 \times l_2) \quad \text{eq.3.8}$$

Data aggregation can be presented using the following example, where trip length and traffic speed were known and are shown in Table 3-3.

Table 3-3: Example route summary

Link ID	Length	Speed
	[km]	[km/hr]
AB	1.8	60
BC	1.0	50
CD	2.4	40

$Total\ length\ of\ route\ OD, L = l_1 + l_2 + l_3 = \sum_{i=1}^{3} l_i$ = 1.8 + 1.0 + 2.4 = 5.2 km

$Weighted\ speed, W_{speed} = s_1 \times \frac{l_1}{L} + s_2 \times \frac{l_2}{L} + s_3 \times \frac{l_3}{L} = \frac{1}{L} \times \sum_{i=1}^{3} l_i \times s_i = \frac{1}{5.2} \times (1.8 \times 60 + 1.0 \times 50 + 2.4 \times 40) =$ 48.84 km/hr = 13.57 m/s

Each OD pair was systematically processed through a described set of instructions and data from the individual links was aggregated into routes.

3.4.3 Specification of deterministic and random utility components

The decision rule, selected to measure the "attractiveness" of each alternative in the feasible choice set, was based on the utility maximization idea. The utility term itself can be partitioned into deterministic and random components. The term “deterministic” can be thought of as a fixed scalar and random term which accounts for all the inconsistencies in the choice behaviour. On the example of a binary choice model, random utility terms U_{in} and U_{jn} are presented in the following notation:

$$U_{in} = V_{in} + \varepsilon_{in} \quad \text{eq.3.9}$$
$$U_{jn} = V_{jn} + \varepsilon_{jn} \quad \text{eq.3.10}$$

where V_{in} and V_{jn} are called the systematic or deterministic components of the utility of i and j; ε_{in} and ε_{jn} are the random terms which are also known as disturbances. In regards to the deterministic term of utility, it was important to answer what types of variables can enter these functions. In a general case, for any individual *n*, any alternative *i* can be characterized by a vector of attributes, like $\mathbf{z}_{in}$; socio-economic characteristics can be characterized by a vector of attributes, like $\mathbf{S}_n$. For convenience, a new vector can be specified, $\mathbf{x}_{in} = \mathbf{h}(\mathbf{z}_{in}, \mathbf{S}_n)$ and $\mathbf{x}_{jn} = \mathbf{h}(\mathbf{z}_{jn}, \mathbf{S}_n)$, where *h* is a function combining road and socio-economic vectors of characteristics. Therefore the deterministic component of utilities *i* and *j* can be presented as:

$$V_{in} = V(\mathbf{x}_{in}) \qquad \text{eq.3.11}$$

and

$$V_{jn} = V(\mathbf{x}_{jn}) \qquad \text{eq.3.12}$$

The functional form of equations 3.11 and 3.12 needed to meet two objectives. It had to reflect our understanding how various terms contribute to utility, and it had to be computationally simple to solve for unknown parameters. Let define $\boldsymbol{\beta} = [\beta_1, \beta_2, \beta_3, \cdots \beta_K]$, as the vector of *K* unknown parameters, so:

$$V_{in} = \beta_1 x_{in1} + \beta_2 x_{in2} + \beta_3 x_{in3} + \cdots \beta_K x_{inK} \qquad \text{eq.3.13}$$
$$V_{jn} = \beta_1 x_{jn1} + \beta_2 x_{jn2} + \beta_3 x_{jn3} + \cdots \beta_K x_{jnK}\,. \qquad \text{eq.3.14}$$

In order to meet the abovementioned objectives, the functional form for *V* is generally accepted as a *linear in the parameters* function for its computational simplicity and mathematical rigour. *Linear in the parameters model is not the same as linear in the attributes* of values stored in $\mathbf{z}_{in}$ and $\mathbf{S}_n$ vectors. It allows attributes in **h** to be presented in any mathematical form, including polynomial, exponential, logarithmic, ratio, piecewise, linear, etc. Therefore linear in the parameters model can capture a high degree of modeling complexities.

For the problem in this research, the functional form of utility, having a choice set of five alternatives, for cyclist №1 can be represented in the following mathematical form:

$$V_{alt1} = -\beta_1(Length_1) - \beta_2(Speed_1) - \beta_3(Volume_1) - \beta_4(Elev.Diff_1) + \beta_5(BikeLane_1) + \epsilon$$
$$V_{alt2} = -\beta_1(Length_2) - \beta_2(Speed_2) - \beta_3(Volume_2) - \beta_4(Elev.Diff_2) + \beta_5(BikeLane_2) + \epsilon$$
$$V_{alt3} = -\beta_1(Length_3) - \beta_2(Speed_3) - \beta_3(Volume_3) - \beta_4(Elev.Diff_3) + \beta_5(BikeLane_3) + \epsilon$$
$$V_{alt4} = -\beta_1(Length_4) - \beta_2(Speed_4) - \beta_3(Volume_4) - \beta_4(Elev.Diff_4) + \beta_5(BikeLane_4) + \epsilon$$
$$V_{alt5} = -\beta_1(Length_5) - \beta_2(Speed_5) - \beta_3(Volume_5) - \beta_4(Elev.Diff_5) + \beta_5(BikeLane_5) + \epsilon$$

where alternative 1 represents the chosen path, alternatives 2-5 represents a set of feasible paths for the same OD pair. Similar cases can be constructed for *n* number of individuals, *i* number of alternatives and *K* number of unknowns. The unknown parameters $\boldsymbol{\beta}$ are the only parameters which need to be calibrated. In this research, unknown parameters $\boldsymbol{\beta}$ were assumed to be of a generic type, which meant that values of parameters were common for each alternative. The calibration of the model was performed by a method of maximum likelihood.

The random component of the utility is introduced as ε, in the example of a simple binomial case with two alternatives only. The random utility approach is consistent with human behaviour, as it explains the behavioural inconsistencies between individuals. The basis for random assumption is due to effects of unobserved attributes, taste variations, measurement errors and imperfect information and instrumental variables (Ben-Akiva et al., 1985). Therefore utilities were considered to be random variables, because utilities were not known to the decision-maker with certainty. The choice probability that alternative *i* was chosen by a decision-maker *n* can be written in the following general form:

$$P_n(i) = \Pr[U_{in} \geq U_{in}]$$
$$= \Pr[V_{in} + \varepsilon_{in} \geq V_{jn} + \varepsilon_{jn}] \quad \text{eq.3.16}$$
$$= \Pr[\varepsilon_{jn} - \varepsilon_{in} \leq V_{in} - V_{jn}] \quad \text{eq.3.17}$$

However, in order to use a particular probability choice model, an assumption about the joint error distribution of error terms ($\varepsilon_{jn} - \varepsilon_{in}$) had to be specified. Once an assumption on the joint

error distribution is made, then the probability that the alternative *i* was chosen could have solved. Starting with an example of a binary probit model (two choices), the methodology is extended to the multinomial logit model. The main assumption related to the probit model exhibit of a normal distribution. Under this assumption, it is expected that the joint error terms $(\varepsilon_{jn} - \varepsilon_{in})$ are also normally distributed. As shown in eq.3.17, this relationship was used to solve for the choice probabilities. The result of eq. 3.17 indicate that if the difference in the utility terms $(V_{in} - V_{jn})$ is greater than the difference of error terms $(\varepsilon_{jn} - \varepsilon_{in})$, then the probability of alternative *i* is obtained by solving the integral, provided in eq.3.18.

$$P_n(i) = \int_{\varepsilon=-\infty}^{(V_{in}-V_{jn})} \frac{1}{\sigma\sqrt{2\pi}} exp\left[\frac{-1}{2}\left(\frac{\varepsilon}{\sigma}\right)^2\right] d\varepsilon, \sigma > 0 \qquad \text{eq.3.18}$$

The content of integral contains mathematical relationship, describing the normal probability distribution of the error terms. Therefore the probability $P_n(i)$ corresponded to an area under the cumulative distribution function (CDF) curve with the lower limit as infinity and an upper limit as $V_{in} - V_{jn}$. This is an analytical approach to understand how one solves for probabilities.

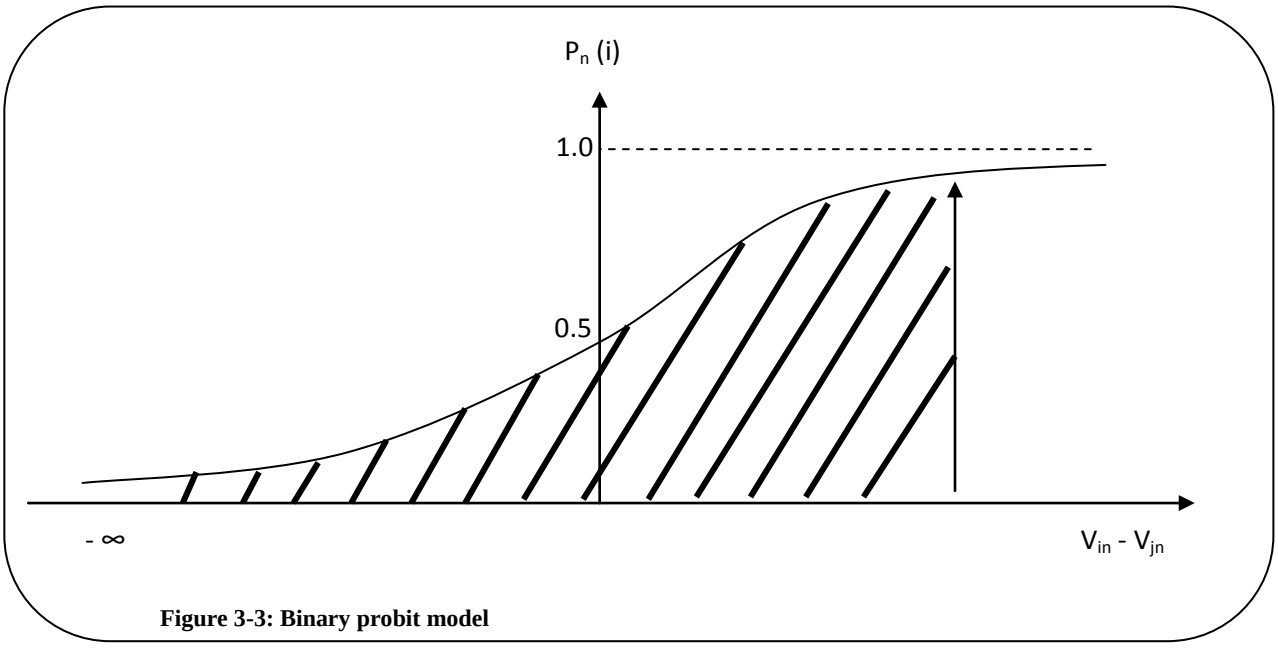


Figure 3-3: Binary probit model

The graphical method is described on Figure 3-3. The figure shows the differences in utility terms corresponding to the upper limit under the curve and the shaded area contained within the interval (-∞; $V_{in} - V_{jn}$] corresponds to probability of selecting alternative *i*.

3.4.4 Multinomial logit model and main properties of the model

In the previous section the binary probit model was introduced. The multinomial logit model is applied to situations when a choice set have more than two alternatives. In our case, the choice set consisted of five alternatives, out of which one was the chosen one and four others were generated additionally, and then added to the choice set as hypothetical paths for cyclists. The random utility theory presented for the binary probit example, is extended to a multinomial case, where the probability of alternative i in the choice set, C_n, selected by a decision maker n, is given by:

$$P(i|C_n) = \Pr[U_{in} \geq U_{in}, \forall\, j \in C_n,\ j \neq i] \quad \text{eq.3.19}$$

$$= \Pr\left[V_{in} + \varepsilon_{in} \geq V_{jn} + \varepsilon_{jn}, \forall\, j \in C_n, j \neq i\right] \quad \text{eq.3.20}$$

$$= \Pr\left[\varepsilon_{jn} \leq V_{in} - V_{jn} + \varepsilon_{in}, \forall\, j \in C_n, j \neq i\right] \quad \text{eq.3.21}$$

However, there is a family of various multinomial models. In order to use a particular multinomial model, an assumption about the error distribution of disturbance terms has to be made. For a multinomial model, the joint distribution of all the terms has to be specified, which added complexity, compared to a binary model where one dealt with the distribution of $(\varepsilon_{jn} - \varepsilon_{in})$ terms only. So, let $f(\varepsilon_{1n}, \varepsilon_{2n}, \cdots \varepsilon_{J_n n})$ to be the joint distribution function of the disturbance terms. The probability of choosing i from C_n is then given by equation:

$$P_n(1) = \int_{\varepsilon_{1n}=-\infty}^{\infty} \int_{\varepsilon_{2n}=-\infty}^{V_{1n}-V_{2n}+\epsilon_{1n}} \cdots \int_{\varepsilon_{J_n n}=-\infty}^{V_{1n}-V_{J_n n}+\epsilon_{1n}} f(\varepsilon_{1n}, \varepsilon_{2n}, \cdots \varepsilon_{J_n n})\, d\varepsilon_{J_n n} d\varepsilon_{J_n-1,n} \cdots d\varepsilon_{1n}$$

eq.3.22

To find the choice probability *i* for a multinomial model one would need to carry over multiple integration over the subspace of disturbances, presented in eq.3.22. In the previous section, an assumption was made that terms $(\varepsilon_{jn} - \varepsilon_{in})$ were normally distributed. A model known as binary *probit* was obtained. For a number of reasons, including not having a closed form and a computational complexity of working with integrals, it was proposed to use a "probit-like" model but which was analytically simpler. One such model is a *logit* model which corresponded to logistic distribution of disturbances. The logistic distribution approximates well normal distribution for the exception of having fatter "tails". The multinomial logit model is expressed as:

$$P_n(i) = \frac{e^{V_{in}}}{\sum_{j \in C_n} e^{V_{jn}}} \qquad \text{eq.3.23}$$

The mathematical expression presented in eq.3.23 is much simpler because it allows the researcher to find the choice probability while avoiding the rigour of integration. However, to use the logit model, the following assumptions about the disturbance terms need to be accepted:

- Identically and independently distributed (IIA)
- Each one of error terms followed Gumbel Type 1

The importance of IIA property and its ramifications are going to be presented in Section 3.4.5 because it was important for the model development.

3.4.5 Independence from irrelevant alternatives property (IIA)

The IIA property states that for any individual the ratio of choice probabilities for choosing two alternatives is independent of the availability of attributes of any other alternatives. The idea behind this statement is described in eq.3.24:

$$\frac{P_n(i)}{P_n(l)} = \frac{e^{V_{in}}/\sum_{j \in C_n} e^{V_{jn}}}{e^{V_{ln}}/\sum_{j \in C_n} e^{V_{jn}}} = \frac{e^{V_{in}}}{e^{V_{ln}}} = e^{V_{in}-V_{ln}} \quad \text{eq.3.24}$$

The IIA property has the following ramifications, which accompanied by examples with explanations:

- If the characteristics of choices/modes not included in the calculation of ratio are changed, the ratio of probabilities included in the calculation of the ratio remains to be unaffected. This can be viewed in the example, when there are three choices: a car, carpool and transit. The probabilities of selecting each mode are 0.38, 0.34 and 0.28, and these probabilities correspond to the base scenario. Now suppose that an improvement was made to transit and the probabilities of choosing a particular mode have changed to 0.37, 0.30 and 0.33, accordingly. Although the individual probabilities have changed, the ratio of auto and carpool probabilities remained unchanged (0.38/0.34 = 1.11 before and now 0.37/0.33 = 1.11). In reality, it would be expected that the probability of transit is increased at the expense of a less competitive mode, which is carpool.

- Alternatives must be distinct. This ramification received the most attention and can be explained through an example of a red and a blue bus. One assumes that initially there are two choices: a car and a red bus. The choice probabilities are given as one half for each mode. Then suppose that a blue bus was introduced and choice probabilities have changed accordingly to one third to each mode. Mathematically this statement is correct but it does not reflect the idea that in reality there are still two modes only: a car and a bus. When an individual makes a choice, they will choose between a car and a bus first, therefore the probability of each mode should be one half or 50% for each choice. Then, if the individual

selects a bus, they may choose either a red or a blue bus with a corresponding probability of one quarter or 25%. In these situations, a hierarchical structure like nested logit is adopted, which can better reflect the true relationship between a variety of choices and conditional probabilities.

3.4.6 Model calibration using maximum likelihood

Model calibration is a process of econometrically inferring the unknown parameters of $K = [\beta_1, \beta_2, \cdots \beta_K]$ from a sample of observations. There are two main approaches to find estimators, which are: the maximum likelihood and least squares. Maximum likelihood is the most commonly used method of inferring unknown parameters, as stated concisely: "maximum likelihood estimator is the value of the parameters for which the observed sample is most likely to have occurred" (Ben-Akiva et al., 1985). Most of the theory and mathematical notation in this section is from Daniel McFadden and presented here to make the methodology section complete.

Let N be the sample size, at which the following can be defined:

$$y_{in} = \begin{cases} 1 & if\ observation\ 'n'\ chose\ alternative\ i, \\ 0 & otherwise \end{cases} \quad \text{eq.3.25}$$

For a multinomial choice model, the maximum likelihood function is:

$$L^* = \prod_{n=1}^{N} \prod_{i \in C_n} P_n(i)^{y_{in}} \quad \text{eq.3.26}$$

For a linear in the parameters logit model

$$P_n(i) = \frac{e^{\beta' x_{in}}}{\sum_{j \in C_n} e^{\beta' x_{jn}}} \quad \text{eq.3.27}$$

In practice it is customary to work with the natural logarithm of L*, which is called the log-likelihood, and seek a maximum to:

$$L = \sum_{n=1}^{N} \sum_{i \in C_n} y_{in} \left(\beta' x_{in} - ln \sum_{j \in C_n} e^{\beta' x_{jn}} \right)$$

eq.3.28

Setting the first derivatives of L with respect to coefficients equal to zero, the necessary first-order conditions can be obtained:

$$\frac{dL}{d\widehat{\beta_k}} = \sum_{n=1}^{N} \sum_{i \in C_n} y_{in} \left(x_{in} - \frac{\sum_{j \in C_n} e^{\beta' x_{jn}} x_{jnk}}{\sum_{j \in C_n} e^{\beta' x_{jn}}} \right) = 0, \text{ for k = 1, ..., K}$$

eq.3.29

In more compact form it can be written as:

$$\sum_{n=1}^{N} \sum_{i \in C_n} [y_{in} - P_n(i)\,]\, x_{ink} = 0, \text{ for k = 1, ..., K}$$

eq.3.30

If the solution to eq.3.30 exists, it is unique which signifies that L is globally concave. The unknown parameters $\boldsymbol{\beta}$ can be mathematically solved through an intricate task. Therefore most of the procedures are now automated through a number of statistical packages like SPSS, or specialized packages like Easy Logit Modeler or Biogeme.

The method of maximum likelihood can be illustrated through an example below. This example presented to explain the theoretical concepts (Fu, 2012). Suppose the following travel time characteristics for three individuals is presented as shown in Table 3-4.

Table 3-4: Route choice characteristics

Individual №	Chosen mode	Travel time (car)	Travel time (bus)
1	Car	30	50
2	Car	20	10
3	Transit	40	30

In addition, assume that the functional form of the deterministic utility component can be described as a function of a travel time only. These functions are:

$V_{car} = aT_{car}$
$V_{transit} = aT_{transit}$

where *a* is the unknown parameter which in this case is generic for both modes, and *T* is a travel time by car and transit. The objective of the problem is to infer the unknown parameter *a* by using the method of maximum likelihood. The choice probability for both modes can be described in the following general form:

$$P(car) = \frac{e^{\beta' x_{in}}}{\sum_{j \in C_n} e^{\beta' x_{jn}}} = \frac{e^{aT_1}}{e^{aT_1} + e^{aT_2}} \quad \text{eq.3.31}$$

$$P(transit) = \frac{e^{\beta' x_{in}}}{\sum_{j \in C_n} e^{\beta' x_{jn}}} = \frac{e^{aT_2}}{e^{aT_1} + e^{aT_2}} \quad \text{eq.3.32}$$

Choice probabilities for each individual can be written as:

Individual 1: $P(car) = \frac{e^{aT_1}}{e^{aT_1} + e^{aT_2}} = \frac{e^{30a}}{e^{50a} + e^{30a}}$ eq.3.33

Individual 2: $P(car) = \frac{e^{aT_1}}{e^{aT_1} + e^{aT_2}} = \frac{e^{20a}}{e^{10a} + e^{20a}}$ eq.3.34

Individual 3: $P(transit) = \frac{e^{aT_1}}{e^{aT_1} + e^{aT_2}} = \frac{e^{30a}}{e^{40a} + e^{30a}}$ eq.3.35

The probability of the entire estimation sample is given by:

L = Pr (individual 1 chooses car) x Pr (individual 2 chooses car) x Pr (individual 3 chooses transit)

Therefore,

$$L = \left(\frac{e^{30a}}{e^{50a} + e^{30a}}\right) \times \left(\frac{e^{20a}}{e^{10a} + e^{20a}}\right) \times \left(\frac{e^{30a}}{e^{40a} + e^{30a}}\right)$$

$$\ln L = -\ln (1 + e^{20a}) - \ln (1 + e^{-10a}) - \ln (1 + e^{10a})$$

The only unknown parameter is *a*, and parameter *a* can be estimated by maximize L or log L. In this example, the search for the unknown parameter is done using the *golden section* method; the

final result is depicted on Figure 3-4. Note that the function is concave and the maximum occurs at a = 0.08 and the corresponding log-likelihood value equals to -1.7.

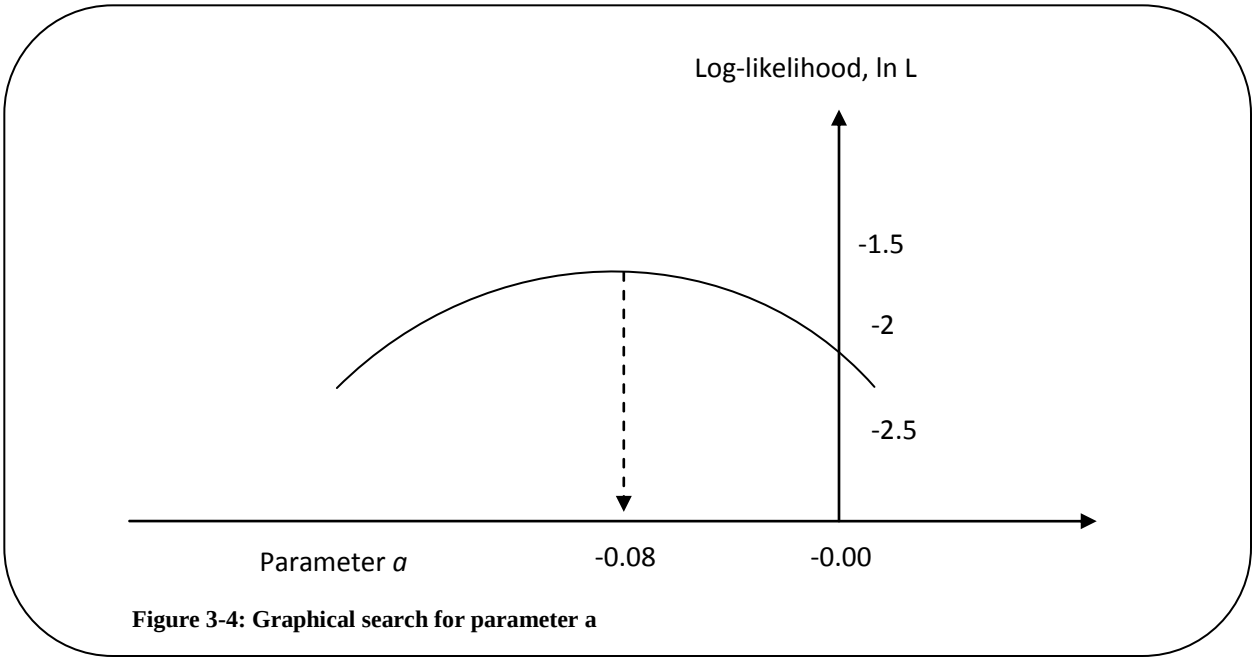


Figure 3-4: Graphical search for parameter a

With the estimated parameter value, the utility functions are fully calibrated and can be used to estimate the choice probabilities for a car and a transit, as long as travel times of their trips is known.

3.5 Verification and validation tests

The process of model building is considered to be a combination of science and engineering judgement. The reason it is challenging to find an ideal model is related to the fact there are several specifications of the model which fit the data well (Ben-Akiva et al., 1985). The science part of the model building process corresponds to the use of rigorous statistical tests. Scientific tests make inferences about the unknown parameters of mathematical models that *are* known to us. Mathematical models are considered to be known to us and are not questioned. At the same time a statistical goodness-of-fit results alone cannot always provide good explanation. There is a

need for some engineering judgement, as even good models may give poor prediction results. The judgement part of the model building process corresponds to *a priori* knowledge, expectations on the phenomenon being modeled.

3.5.1 Informal tests of the coefficient estimates

The examination of values and signs of model coefficients is considered to be as one of the most basic tests to be performed. There are certain *a priori* expectation towards the signs of trip length, traffic speed, traffic volume, percent of uphill gradient and presence of bicycle lane coefficients. In particular, negative signs are expected for the trip length, percent of uphill gradient and speed coefficients; a positive sign is expected for bicycle lane. No specific sign is expected for the volume coefficient. These statements can be explained through an example of looking at the percent of uphill gradient coefficient. Our *a priori* expectation towards percent of uphill gradient sign is negative because cycling uphill is considered a major physical effort that a traveler tries to avoid. The ratio of coefficients can provide valuable information also. Typically, they provide information on a trade-off or a marginal rate of substitution between two corresponding variables.

3.5.2 Statistical tests

The use of asymptotic *t*-test: this test is designed to test a hypothesis if a particular parameter in the model differs significantly from a critical value, known as a test statistic. Critical values are obtained from the standardized Student's *t* distribution for various confidence levels (90% or 95%) and for one or two-tailed tests. A set of conditions can be constructed to test each coefficient parameter in the model, using the notation provided below:

$$H_0 = \beta_K = 0$$
$$H_A = \beta_K \neq 0$$

eq.3.36

On the basis of the *t* statistic, the null hypothesis can be either rejected or accepted if individual parameters are equal to zero. Most often, additional insight into the data can be gained from finding a joint confidence region for several parameters.

The use of the likelihood ratio test: this test is designed to test a hypothesis that all coefficients are simultaneously equal to zero. A set of conditions can be constructed, using the following notation:

$$H_0 = \beta_1 = \beta_2 = \cdots = \beta_K = 0$$
$$H_A = \beta_1 = \beta_2 = \cdots = \beta_K \neq 0 \qquad \text{eq.3.37}$$

The test statistic is X^2 distributed with K degrees of freedom and can be obtained from:

$$-2\left(L(0) - L(\hat{\beta})\right) \qquad \text{eq.3.38}$$

where L(0) and L(β) correspond to the maximum likelihood values obtained for restricted and unrestricted forms of models. The test statistic is X^2 distributed with (K_U - K_R) degrees of freedom and it can be obtained from:

$$\left(-2L(\hat{\beta}_R) - L(\hat{\beta}_U)\right) \qquad \text{eq.3.39}$$

where 'R' and 'U' are the subscripts which correspond to the restricted and unrestricted models. The restricted model corresponds to a simpler model, as it requires less data than an unrestricted model.

The use of goodness of fit measure: this test is similar to the R^2 statistic found in linear regression analysis. This measure can be obtained using the following notation:

$$\rho^2 = 1 - \frac{L(\hat{\beta})}{L(0)} \qquad \text{eq.3.40}$$

Typically, everything else being equal, a model specification with the higher goodness-of-fit measure is considered better, as it better explains the data.

3.5.3 Test of the model structure: IIA assumption

This test is designed to test the assumption of accepting the multinomial logit structure. In practical terms, one would like to compare logit models estimated with subsets of alternatives from the universal choice set (Ben-Akiva et al., 1985). This test should answer the question if the model parameters are stable. If the IIA assumption holds for the full choice set, then the logit model should apply to any subset of alternatives. Typically, the coefficients from the full and subset choice sets are required to be within one standard error. Most of the mathematical notation can be found in the research papers by McFadden et al., 1974, 1976, 1977.

3.5.4 Prediction Tests

The tests presented in this section consist of performing outlier analysis and testing the predictive power of models. These tests are performed on the external data, meaning: the data which was 'not seen' by the model. The data are divided into two sets, where the first set uses 80% of data for model building, and the second set uses 20% of data for prediction.

Condition 1: Pr (Maximum) = Pr (Chosen Path)

The first prediction test is designed to find how often the model identifies the chosen paths as the highest probability paths. Route choice coefficients, obtained from the model calibration process, are substituted directly into the attributes of each alternative. If the probability of selecting the chosen path gives the largest probability among the choice set of alternatives, than a value 1 is assigned, otherwise 0. So,

$$\text{Let Z} = \begin{cases} 1 & \text{if Pr (Maximum)} = \text{Pr (Chosen Path)}, \\ 0 & otherwise \end{cases}$$

and then

$$100\% \times \frac{\sum Z}{N}, where\ N\ is\ the\ total\ number\ of\ routes$$

The success of prediction can be expressed in terms of the total number of matched records in relation to the total number of records, which gives the first measure of predictive power of a model.

Condition 2: $\mathbf{Pr\ (Maximum) - Pr\ (Chosen\ Path) < \epsilon, where\ \epsilon\ [0.01, 0.1]}$

The research also examined situations when the difference between the maximum and the actual probabilities is less than a certain minimum threshold. A typical range of examined threshold, ε, lies between 0.01 to 0.10. If the difference between the maximum and the chosen probabilities is less than a certain minimum threshold, than there is a likelihood that the actual route could have been chosen by a cyclist. In mathematical terms, it can be expressed in the following way,

$$\text{Pr (Maximum) - Pr (Chosen Path)} < \varepsilon \qquad \text{eq.3.42}$$

The models are tested for various values of threshold values, ε.

Condition 3: Outlier Test $\mathbf{Pr\ (Maximum) \gg Pr\ (Chosen\ Path)}$

Once the model parameters are substituted into the external dataset, unusually large deviations should be examined. Outliers can be located through a condition, when:

$$\text{Pr (Maximum) - Pr (Chosen Path)} > \varepsilon, \text{ where } \varepsilon \text{ » } 0.1 \qquad \text{eq.3.43}$$

When ε is greater than 0.1, then there is a strong evidence that the likelihood of predicting the chosen route with the calibrated model parameters is quite low, despite the fact this route was chosen by a traveler. These outliers should be examined carefully and an explanation should be given to the possible causes of observed effects.

Market segmentation prediction tests: this test is designed to demonstrate the ability of the proposed model to repeat correctly the observed shares of alternatives for a particular market segment. In mathematical terms, this test can be expressed simply as:

$$\sum_{n=1}^{N} P_n(i) = N_i, i \in C, \qquad \text{eq.3.44}$$

where N_i is the number of observations choosing alternative *i* from a choice set *C*.

Since the frequency of the actual choice is observed, this frequency is considered to be representative of the market segment.

3.6 Methodology summary

The entire route-choice methodology can be thought of a sequential execution of five processes as presented in Figure 3-5.

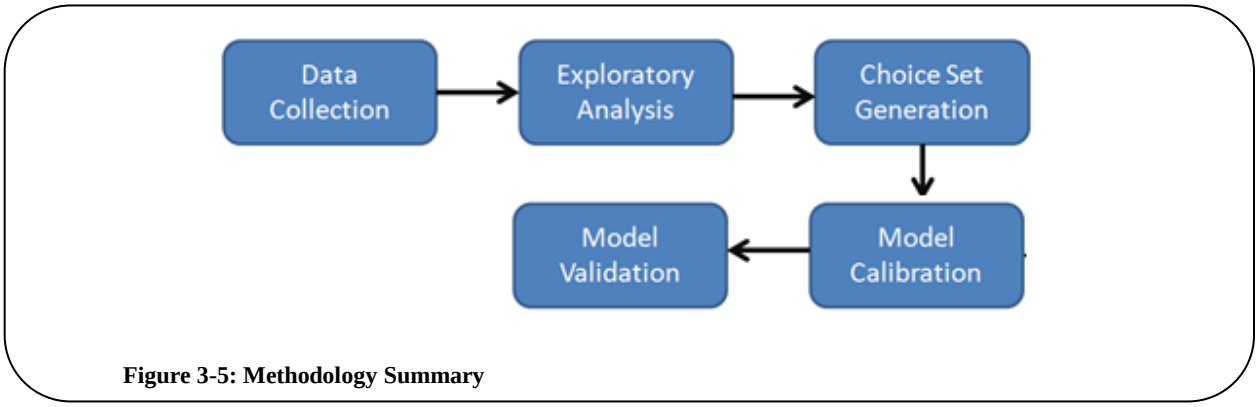


Figure 3-5: Methodology Summary

Data collection was the initial step in the methodology. Data required for route-choice modeling included GPS routes from individual cyclists and a range of attributes related to transportation network, such as trip length, traffic volume, traffic speed, gradient and presence of cycling infrastructure. Exploratory analysis of gathered data was performed after that. This analysis focused mainly on two objectives. Firstly, it involved the validation of the assumption of a linear relationship between the selected road variables; second, was to validate the presence or absence of confounded effects. The presence of the confounded effects can be verified in a statistical

package, such as SPSS. The third process involved a generation of a feasible choice set. This process was performed using a set of heuristic rules, which were based on finding optimal paths for cyclists. In total four alternatives were obtained for each observed route. Attributes were located in GIS, and data aggregation was performed through a specially written computer program. Afterwards these alternatives were verified for meeting the IIA assumption. The fourth process was to calibrate model parameters using discrete choice theory: random utility approach. Multinomial logit framework was selected as the modeling framework and the calibration of model parameters was performed using a program Easy Logit. At last, models were validated, which included a range of intuitive, statistical and predictive tests. Outlier analysis was also necessary because it allowed to have a better understanding of why the actual choice was not predicted by the model.

4 Chapter 4 - Application to the Region of Waterloo.

4.1 Introduction

The regional municipality of Waterloo is comprised of three cities - Waterloo, Kitchener and Cambridge, and four townships - North Dumfries, Wellesley, Wilmot and Woolwich. The Region of Waterloo is located 110 km west of Toronto and is situated within the boundaries of an area known as the Greater Golden Horseshoe. The Region has been designated by the Province of Ontario as "a place to grow" and is currently progressing towards intensification of its urban cores by building LRT and BRT transit systems along the central corridors (Places To Grow, 2006). The current population is estimated at 550,000 residents and is expected to grow to 730,000 by 2031; the number of jobs is also expected to grow by an additional 50% (Statistics Canada, 2006). As a result of growth, demand for housing, employment and transportation will add pressure on existing communities and municipal infrastructure.

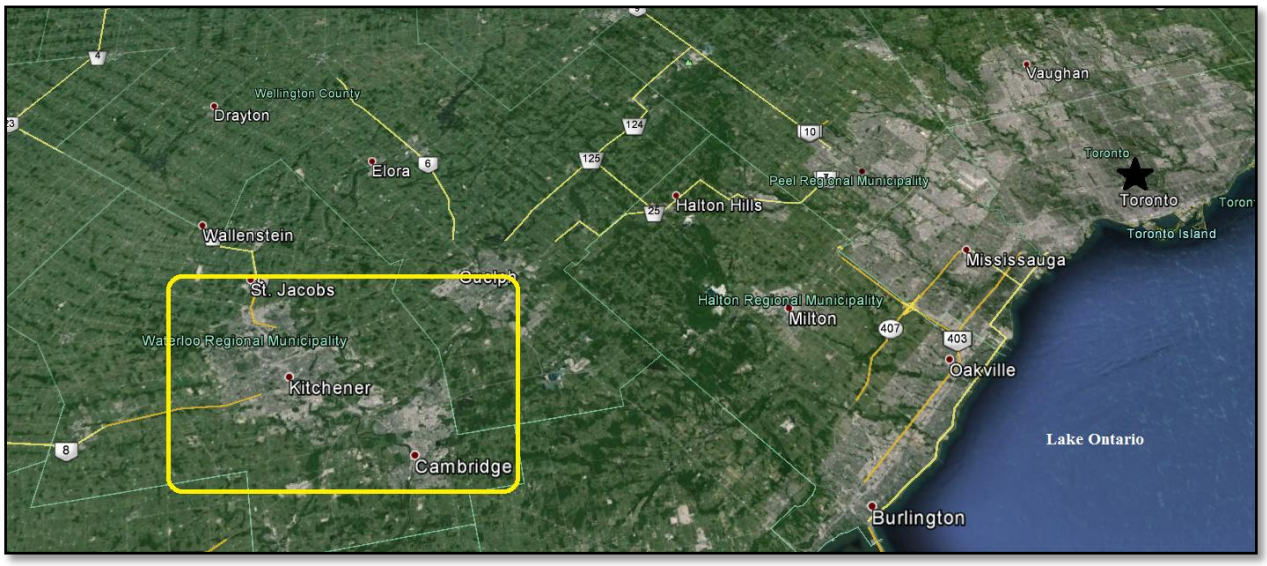


Figure 4-1: Region of Waterloo (Source: www.maps.google.ca)

In addition to building LRT and BRT transit systems, the Region of Waterloo has emphasized and promoted non-motorized modes. Cycling in particular, has received substantial attention.

The vision for developing cycling into a significant mode was introduced formally in the Cycling Master Plan. According to the Plan, the Region has allocated $33 million to build 730 km of cycling network and is expected to increase the mode share of cyclists to 2% by 2016 (Cycling Master Plan, 2006).

In order to assist the Region with the implementation of this plan to accommodate growth, the University of Waterloo partnered with the Region to collect, analyse and evaluate data on cyclists' behaviour and to determine key factors that may increase cycling. The modeling of cyclists' route choice, which is the focus of the thesis, can assist the Region in the strategic addition of new and modification of existing cycling infrastructure.

4.2 Data collection

As discussed earlier (Chapters 1, 2), in order to determine key factors that explain cycling route choice, data about travelers and route characteristics were required. Data describing travelers are typically collected in the form of socio-economic characteristics. Since 40% of these characteristics were missing from the survey, these variables were left out of the model building process.

Attributes of cyclist paths can be derived from several sources. Road attributes were obtained through a process of averaging and combining multiple GIS sources into one large geo-spatial transportation network database. The actual choice data were collected in the form of GPS data from small, inexpensive units carried by cyclists.

4.2.1 GPS data

The actual route choice of travelers was originally collected in 2011 by a University of Waterloo initiative through a year-long study involving 415 self-selected cyclists. Self-selected cyclists are

cyclists who decided voluntary to participate in this research. The study was conducted in several sessions by providing GPS tracking devices to individual cyclists. The result of the data collection provided the actual choice, origins and destinations, and routes themselves. Overall, 2000 routes were collected through the revealed preference survey, which translated into 4000 origin-destination pairs.

4.2.2 Transportation network

The transportation network consisted of topological elements, described by nodes, links and non-topological features, represented in the table of attributes. The table of attributes contained various road characteristics related to every link in the network. Most of the transportation attributes were obtained from the University of Waterloo Library. Gradient was obtained by performing a projection of a digital elevation model (DEM) onto the nodes of road network. The horizontal and vertical accuracy of DEM data was stated by the source at ±0.5 m, 1σ-level with density of 1 point per 10 m^2.

4.3 Exploratory data analysis

The original exploratory data analysis was carried out by Rewa (2012) and some of the main findings are summarized herein. The analysis of the socio-economic characteristics revealed that 76% of cyclists were male and 24% female; 97% of sampled cyclists were licensed drivers. The study looked at excess travel by comparing the chosen routes with the shortest paths obtained by the road and cycling networks. Based on the study, it was observed that 25% of the chosen trips had an excess travel of more than 40%, if only the road network was considered. The addition of a network of off-road cycling paths to the road network reduced the percentage of excess travel to 15%. The excess travel study found that cyclists were inclined to take longer routes when they had a choice. For example, they would travel longer distance to use a cycling facility with steep

gradients. The study also looked at motivation and obstacles to cycling and found that personal safety and high traffic volume were the most cited responses.

4.4 Modeling framework - discrete choice theory: random utility approach

This section proposes and evaluates several route-choice models. The theoretical background of necessary concepts was covered in Chapter 3.

4.4.1 Generation of choice set

In addition to the chosen path, four alternatives were generated, based on heuristic rules. These heuristic rules represented a combination of labeling and impedance methods, similar to those alternative generation methods described in the literature. Each of the alternative generation methods was based on the shortest-path deterministic family of algorithms. An example of alternatives obtained for one OD pair is presented in Figure 4-2.

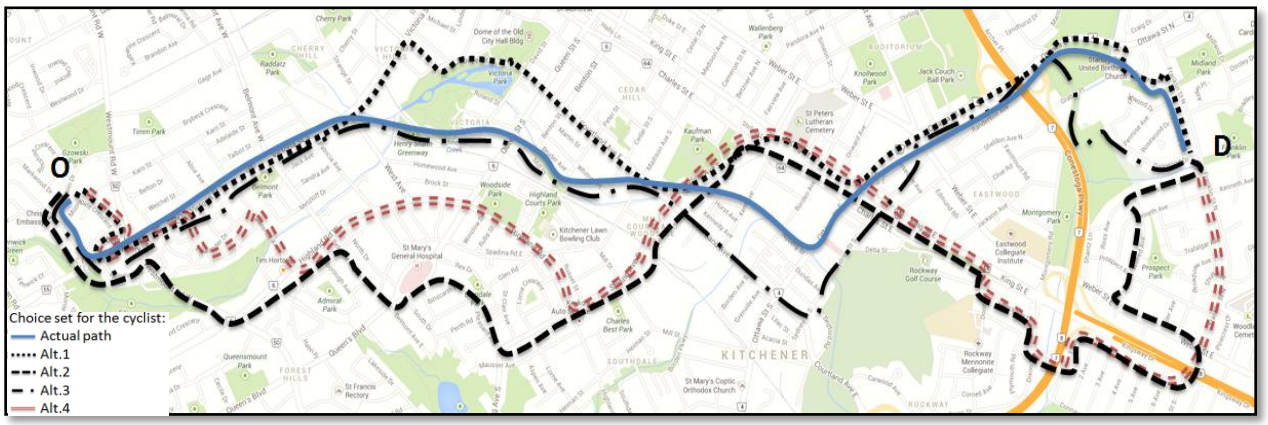


Figure 4-2: Origin-destination pair and generated set of alternatives for the Region of Waterloo.

4.4.2 Gathering of alternatives' attributes

Once alternatives were generated, it was necessary to ascertain the path attributes from the roadway network. The process of connecting topological features, represented by routes, to road attributes is done through feature relation. Feature relation is available in most GIS software packages. The main input to the feature relation process involved the search radius within which

attributes needed to be located. The search radius was selected to be 15 metres. As presented in Figure 4-3, the curve, constructed from the various search radii, becomes asymptotic to the x-axis at 15 metres. This means that most of the attributes can be located within a 15 metre radius.

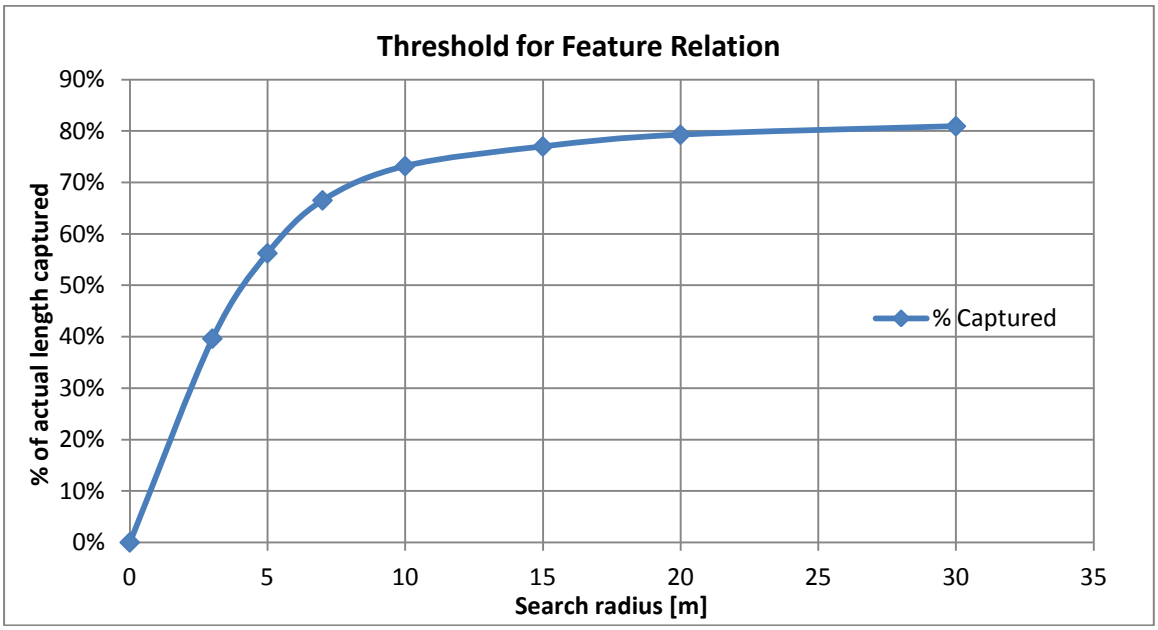


Figure 4-3: Threshold limit for linear referencing mechanism.

The ordinate axis in Figure 4-3, percent of the actual length captured, represents the percentage of the chosen length obtained from the feature relation process that was compared to the nominal length of a route. At a fifteen metre radius, one obtains 77% of the nominal length of a typical route, on average. Although 2000 routes were collected, only 905 of them were selected for the analysis. Recreational trips, trips shorter than 500 meters and trips whose length was missing for more than 40% of the actual length were removed from the analysis.

4.4.3 Exploratory analysis of alternatives

The available data were downloaded and then saved in ASCII file format. The processing stage involved the aggregation of individual links and attributes into routes. The method behind data aggregation was described in Section 3.4.2, and a typical output from the route aggregation

process is presented in Table 4-1. In this example, the chosen alternative was shorter in length and was observed on a roadway with lower traffic speed, lower amount of gradient and a greater amount of cycling facilities.

Table4-1: Data aggregation example (a chosen path and four alternatives)

Path Attribute	Alternative 1	Alternative 2	Chosen Alternative	Alternative 3	Alternative 4
Length (km)	7.2	7.4	**7.5**	9.2	9.9
Auto speed (km/h)	33.2	48.2	**35.8**	49.8	41.5
Auto volume (veh/h)	245.1	357.4	**240**	219.8	290.2
Grade	0.5	0.6	**0.4**	0.8	0.6
Presence of bicycle lane	0.6	0.2	**0.4**	0.02	0.2

The basic exploratory analysis of alternatives was necessary to start the model building process. Figure 4-5 represents scatter plot matrices for the chosen path and alternatives, obtained by finding the shortest path by road network and road and trails network. The inspection of individual relationships suggests that the dispersion of most of the variables was fairly narrow, and therefore a linear form can be used for the constant part of utility. The second observation was related to the plot of the speed - bicycle lane relationship. A correlation between independent variables was considered an indication of confounded effects, and therefore two models were necessary. One model had to include a bicycle lane coefficient and the second model had to include a speed coefficient. Figure 4-6 shows an additional set of scatter plot matrices. These matrices were obtained by differencing the matching attribute values of the chosen and the shortest path alternatives. The main observation confirmed that many cyclists traveled a longer path due to the advantages of a lower uphill gradient, lower traffic speed, volume and higher amount of cycling infrastructure. On the example of length-gradient relationship, as the difference in trip length increase the gradient goes down. This result is

consistent with our *a priori* knowledge of travelers' behaviour: cyclists opted for longer trip distances with higher amount of cycling facilities to avoid steep gradients.

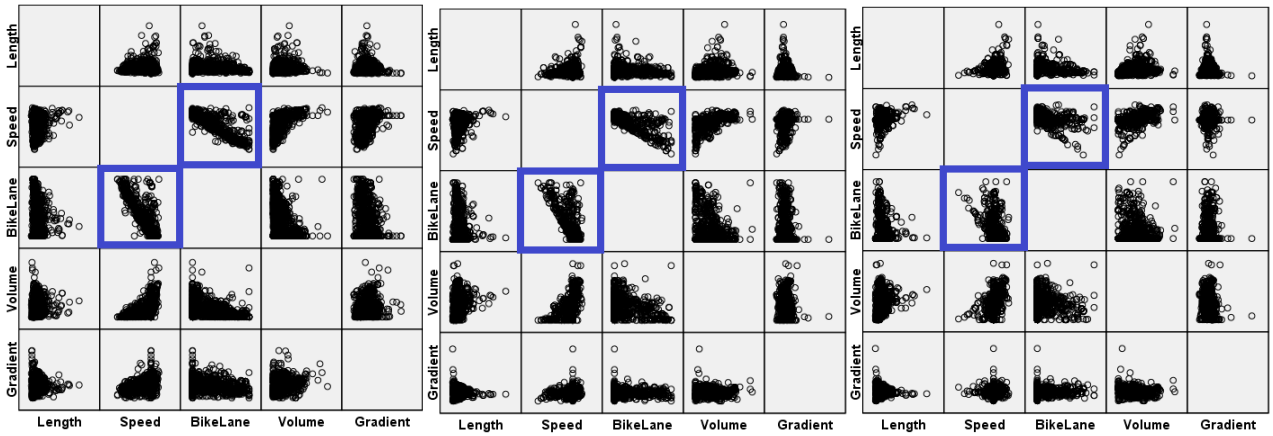


Figure 4-4: Scatter plots (from the left to the right: chosen paths; short path by the road network; short path by the road and trails network)

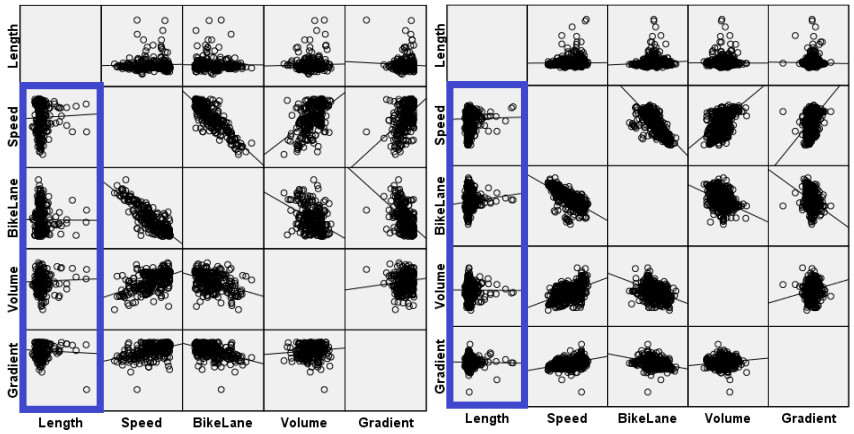


Figure 4-5: Scatter plots (from the left to the right: the difference between chosen and paths obtained by the road network; difference between chosen and paths obtained by the road and trails network).

4.4.4 Specification of route-choice model

Model building is regarded as a combination of science and engineering judgement. The judgement part of the model building consisted of the intelligent specification of the model form, as well as a priori signs of utility parameters. A *linear parameter* model was considered adequate to model the deterministic part of utility function. Multinomial logit structure was considered as the main framework to model the probabilistic nature of choice variables. The general specification of route-choice model is provided in eq.4.1:

$$U = -\beta_1 \times (Length) - \beta_2 \times (Speed) \pm \beta_3 \times (Volume) - \beta_4 \times (Elev.Diff) + \beta_5 \times (BLN) + \varepsilon \quad \text{eq.4.1}$$

The general form of eq.4.5 was specified in accordance with understanding of cyclists' behaviour which was supported by literature review. The expectations towards the signs of utility terms are provided in Table 4-2.

Length	Speed
"- ve", traveling longer paths is a cost for cyclists; cyclist may travel longer paths to compensate by other considerations, like safety;	"-ve", traveling along busy streets can be considered a safety issue for cyclists;
Bicycle Lane	**Elevation Difference/+ve Gradient**
"+ve", cyclist are expected to use dedicated cycling facility as they perceive them a safer travel;	"-ve", traveling uphill requires major physical effort which generally is tended to be avoided by most of the cyclists;
Volume	
"-ve", cyclists may not be attracted to travel along streets with heavy traffic volume;	

Table 4-2: Sign expectation of utility parameters.

Once the alternatives were aggregated, they were tested for IIA assumption of the multinomial logit framework. This test was designed to check for duplication among alternatives. The following rule was proposed:

IF {[Difference in length variable of any two alternatives is less than 500 meters]
AND
[Difference in any other variable of any two alternatives is equal to zero]}
THEN
"This is duplicate, remove from analysis!"

The rule was used to identify apparent duplicates and remove them from further analysis. Out of 965 routes - 60 were eliminated from the further analysis thus leaving with 905 routes for econometric analysis. The estimation of model parameters was performed using Easy Logit software.

4.4.5 Model Type 1

Work assumptions made on the general model form allowed the model building process to be undertaken. Since the bicycle lane variable was shown to have a strong correlation with the speed variable, model 1 excluded the speed parameter. The proposed model for evaluation is specified in eq.4.2:

$$Utility_{OD} = -\beta_1 \times (Length) + \beta_2 \times (Bike\ Lane) + \beta_3 \times (Volume) - \beta_4 \times (Elev.Diff) + \varepsilon \quad \text{eq.4.2}$$

The dataset was divided into portions of 80-20, where 80% (724 routes) of the data was used for estimation of the model parameters and 20% (181 routes) for validation. Parameter estimates and corresponding *t-Statistic* are presented in Table 4-3. All coefficients are of generic type, meaning that same coefficients apply to each alternative.

Generic Parameters	Estimated Value	t - Statistic
Length_KM	-0.1818	-5.1864
Bicycle Lane	4.3081	11.7626
Volume	0.0001	0.371
Gradient	-1.4864	-6.4719

Table 4-3: Estimated parameters of Model 1

The asymptotic *t*-test, carried out on trip length, bicycle lane, and percentage of uphill gradient parameters, was set at 95%; therefore, t-statistic exceeding ±1.96 was deemed significant,

assuming a two-tailed test. From Table 4-3, every coefficient, except for traffic volume had a high explanatory power and therefore had to be retained in the model. The volume coefficient had a very low statistic, which corresponded to little explanatory power and had to be tested for exclusion from the model. Cycling behaviour described by model 1 represents dedicated users of cycling infrastructure, or the so-called inexperienced group who prefer to travel along cycling infrastructure.

The inspection of trip length and percentage of uphill gradient coefficients suggested that they were negative. A negative sign for these parameters suggested that an increase in any of these variables for an alternative decreases cyclists' preference towards this alternative. The coefficient for the bicycle lane variable was positive, and an increase in this variable for any alternative causes an increase in the cyclists' preference towards this alternative. The interpretation of utility signs suggested that the presence of a bicycle lane provided value that offset the cost of taking a longer path.

The coefficient associated with each term represents the amount of change in utility or preference that would result from a unit change in that attribute. The utility equation is interpreted as follows: as the length of a trip increases by 1 km, the utility decreases by 0.18; presence of bicycle lane increases utility by 4.3; utility decreases by 1.48 for each percentage of uphill gradient that cyclists faced.

The relative importance of the trip length and bicycle lane parameters is evaluated by comparing the ratio of two coefficients, β_1 and β_2. This ratio is also known as the marginal rate of substitution. The ratio of $|\frac{\beta_2}{\beta_1}|$ = 24:1, suggests that 24 units of bicycle lane can be substituted for

1 km of the travel length. The ratio of $|\frac{\beta_2}{\beta_4}|$ = 3:1, suggests that 3 units of bicycle lane can be traded for 1 unit of percent of uphill gradient that cyclists faced.

The statistical summary of model type 1 is provided in Table 4-4. The adjusted ϱ^2 index, used to evaluate the overall quality of the model, is above 0.15. This indicates that the model is acceptable.

Table 4-4: The summary of model 1

Log Likelihood at Zero	-1160.4
Log Likelihood at Constants	0
Log Likelihood at Convergence	-972.46
Rho Squared w.r.t. Zero	0.162
Rho Squared w.r.t Constants	--
Adjusted Rho Squared w.r.t. Zero	0.1585

The exploratory power of the volume variable was tested through the application of the likelihood test. The value of Log L with volume term dropped, was -972.5266. The likelihood ratio test statistic is then LR = 2x[(-972.46) - (-972.5266)] = 0.13, which is significantly smaller than any chi-square statistic. Therefore the traffic volume term was dropped, and model 1 was simplified to the final form, provided in eq.4.3:

$$\boldsymbol{Utility_{OD}} = -0.1818 \times (\boldsymbol{Length}) + 4.3081 \times (\boldsymbol{Bike\ Lane}) - 1.4864 \times (\boldsymbol{Elev.Diff}) + \boldsymbol{\varepsilon} \quad \text{eq.4.3}$$

4.4.6 Model 1: prediction and outlier analysis

Parameters presented in Table 4-3 were used to test the predictive power of the model. Model parameters were applied to the external 20% (181 routes) of the data, the dataset which was not seen by the model. The predictive power of model 1 was found to be 65%. This result meant that 65% of routes were explained by the model.

An example of the predictive test is presented in Table 4.5. The probability for the chosen alternatives for trips #917, 918, 919 and 920 were the highest; the model predicts the chosen path as having the highest probability of occurrence. On the other hand, the probability for the chosen trip #921 was the lowest among all the alternatives.

Table 4-5: An example of alternatives' probabilites

	Trip_ID	Pr(actual)	Pr (alt.1)	Pr (alt.2)	Pr (alt.3)	Pr (alt.4)	Pr (sum)	Pr(best)
Model 1	917	**0.530**	0.093	0.146	0.114	0.117	1	0.530
	918	**0.513**	0.109	0.174	0.100	0.104	1	0.513
	919	**0.307**	0.288	0.154	0.207	0.045	1	0.307
	920	**0.857**	0.028	0.044	0.035	0.036	1	0.857
	921	0.043	0.193	**0.292**	0.228	0.244	1	0.292

The strength of the model can also be evaluated by estimating how often the chosen alternative is given each rank; the results are presented in Table 4.6. From best (1) to worst (5) amongst alternatives, the chosen path was ranked first for 65% of the time; second 10% of the time and third for 14% of the time. The chosen alternative was ranked 4th or 5th less than 10% of the time.

Table 4-6: Chosen alternative ranked by the frequency of occurrence

Rank	Frequency
1	65%
2	10%
3	14%
4	7%
5	3%

It is important to analyze situations where the chosen alternative was not predicted by the model.

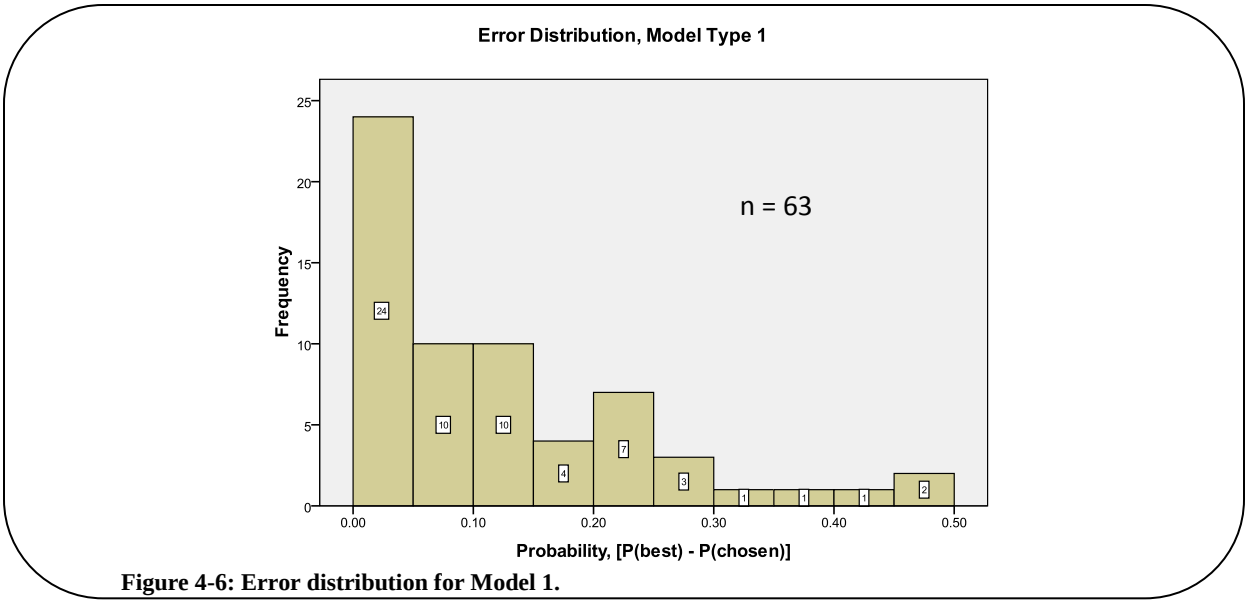


Figure 4-6: Error distribution for Model 1.

The probability differences between the model prediction of the highest probability route and the actual choice were calculated for each OD pair and then summarized on Figure 4-7 and Table 4-7. These results exclude cases where the model correctly predicts the chosen path. The results shown in Table 4-7 suggest that the model identifies the chosen path either correctly or within 5% of the maximum likelihood path 78% of the time.

Table 4-7: Percent of correctly forecasted values for various probability levels.

Probability Class	% Correctly Forecasted
0.00 - 0.05	78%
0.05 - 0.10	84%
0.10 - 0.15	90%
0.15 - 0.20	92%
greater than 0.2	100%

At the same time, it was critical to evaluate the situations when the probability of the chosen path was significantly smaller than the model's prediction. This consideration initiated the outlier analysis.

Table 4-8: Distribution of outliers classified by alternative type.

Alternative	Proportion	Decription
2	54%	Road and trails network
3	18%	Road network only
4	18%	Road and trails network (impedance rule)
5	11%	Road network (impedance rule)

Outliers were classified as values whose differences exceeded 0.1 level as specified in Table 4-7. The outliers were grouped by the alternative type; the results are presented in Table 4-8. Based on the provided exploratory summary, Alt.2 and Alt. 4 contributed to 72% of the outliers. In order to understand outliers better, one must answer two questions. What was common among the non-chosen alternatives that served to their advantage? Can outlier behaviour be explained by other theories?

Alt 2 vs. the Chosen Alternative:

- Alt 2. offered paths shorter by 462 meters, on average;
- Alt 2. offered more paths along bicycle lanes, on average ;
- Alt 2. offered fewer uphill gradient sections, on average.

Alt 3 vs. the Chosen Alternative:

- Alt 3. offered paths shorter by 215 meters, on average;
- Cycling facilities (lanes, trails) did not play a role;
- Alt 3. offered less percent uphill gradient sections, on average.

Alt 4 vs. the Chosen Alternative:

- Alt 4. offered paths longer by 1 km, on average;
- Alt 4. offered paths along the cycling facilities but differences were negligent;
- Alt 4. offered more paths with lower uphill gradient sections.

Alt 5 vs. the Chosen Alternative:

- Alt 5. offered paths longer by 900 meters, on average;
- The bicycle lane variable did not play a role;
- Alt 5. offered paths with lower uphill gradient sections, on average.

The summary of the outlier analysis is provided in Table 4-9.

Table 4-9: Summary of non-chosen alternatives vs. chosen

Alt. No.	Length	Bicycle lane	Gradient (%)
2	Shorter by 0.5 km	More	Less
3	Shorter by 0.2 km	Not significant	Less
4	Longer by 1 km	Not significant	Less
5	Longer by 0.9 km	Not significant	Less

The chosen alternative in comparison to Alt 2. and Alt 3. was described by longer paths, lower amounts of cycling facilities, and a larger percentage of uphill gradient, on average. The chosen type of behaviour can only be explained by habit, inertia, or an unknown to us variable(s). Chosen alternatives in comparison to Alt 4. and Alt 5., offered shorter trips but with larger amounts of uphill gradient sections; the cycling facility coefficient did not play a role. The chosen type of behaviour represented cyclists who were not opposed to the physical effort of taking steeper paths if they offered a shorter trip length.

4.4.7 Model Type 2

The assumption made about the general model form in Section 4.4.4, was necessary to continue the model building process. Since bicycle lane and speed variables could not be used in the same model, an alternative model type was specified in the form of eq.4.4:

$$Utility_{OD} = -\beta_1 \times (Length) - \beta_2 \times (Speed) - \beta_3 \times (Volume) - \beta_4 \times (Elev.Diff) + \varepsilon \quad \text{eq.4.4}$$

The entire dataset was divided into portions of 80-20, as 80% (724 routes) were used for estimation of model parameters and 20% (181 routes) for validation. Parameter estimates and the corresponding *t-Statistic* are presented in Table 4-10. All coefficients are of a generic type.

Table 4-10: Estimated parameters of model 2.

Generic Parameters	Estimated Value	t - Statistic
Length_KM	-0.083	-2.2037
Bicycle Lane	-0.7025	-15.466
Volume	0.0023	6.4307
Gradient	0.5009	-2.0703

The asymptotic *t*-test was carried out on the trip length, traffic speed, traffic volume and uphill gradient coefficients, and found them significant at 2σ-level. All coefficients had high explanatory power and therefore had to be retained. The behaviour described by model 2 indicates either experienced users who traveled along busy streets, or a lack of cycling infrastructure, which forced cyclists to use regular roads with high speed and traffic volume.

Coefficients associated with the trip length, traffic speed and percent of uphill gradient were negative, suggesting that an increase in any of these variables for an alternative causes a decrease in cyclists' preference towards this alternative. The coefficient for the traffic volume variable was positive, meaning an increase in the volume causes an increase in the cyclists' preference towards this alternative. The interpretation of utility signs suggested that a high traffic volume on a road compensated for taking a longer path with a lower amount of uphill gradient sections. Note that high traffic volume did not necessarily mean that speed is also high, as high volume could correspond to lower traffic speed which occurs due to congestion.

The parameters associated with each variable term represent the changes in utility (or preference) that would result from a unit change in that attribute. The utility equation can be interpreted as

follows. As the length of a trip increased by 1 km, the utility decreased by 0.083; utility decreased by 0.7025 for each increase in speed (m/s); utility increased by 0.0023 for each increase in the traffic volume; utility decreased by 0.5009 for each percent of uphill gradient cyclists faced.

The relative importance of trip length and traffic volume was evaluated by comparing the ratio of coefficients. The ratio of $|\frac{\beta_1}{\beta_3}|$ = 36:1, suggests that 36 units of travel length can be substituted for one unit of the volume. The ratio of $|\frac{\beta_2}{\beta_3}|$ = 305:1, suggests that 305 units of speed can be traded for a one unit of traffic volume.

Table 4-11: Summary of Model 2

Log Likelihood at Zero	-1160.4047
Log Likelihood at Constants	0
Log Likelihood at Convergence	-853.9279
Rho Squared w.r.t. Zero	0.2641
Rho Squared w.r.t Constants	--
Adjusted Rho Squared w.r.t. Zero	0.2607

The statistical summary for model 2 is presented in Table 4-11. The adjusted ϱ^2 index, used to evaluate the overall quality of the model, was 0.26 and this indicated that the model was acceptable. The final form of the model 2 is presented in eq.4.5:

$$\boldsymbol{U} = -0.083 \times (\boldsymbol{Length}) - 0.7025 \times (\boldsymbol{Speed}) + 0.0023 \times (\boldsymbol{Volume}) - 0.5009 \times (\boldsymbol{Elev.Diff}) + \boldsymbol{\varepsilon} \quad \text{eq.4.5}$$

4.4.8 Model Prediction of outlier test for Model 2

Parameters presented in eq.4.5 were used to test the predictive power of model 2. Model parameters were applied to the external 20% (181 routes) of data, the data which was not seen by

the model. The predictive power of the model was found to be 46%. This result meant that 46% of routes were explained by the model.

Table 4-12: Percent of correctly forecasted values for various probability levels.

Probability Class	% Correctly Forecasted
0.00 - 0.05	54%
0.05 - 0.10	62%
0.10 - 0.15	73%
0.15 - 0.20	77%
greater than 0.2	100%

The probability differences between the model's best prediction and the actual choice were calculated for each OD pair. The summary is presented in Figure 4-8 and Table 4-12. The results in Table 4-12 suggests that the model explains 54% of data at a 0.05 threshold. Model 1 appears to be a much better model.

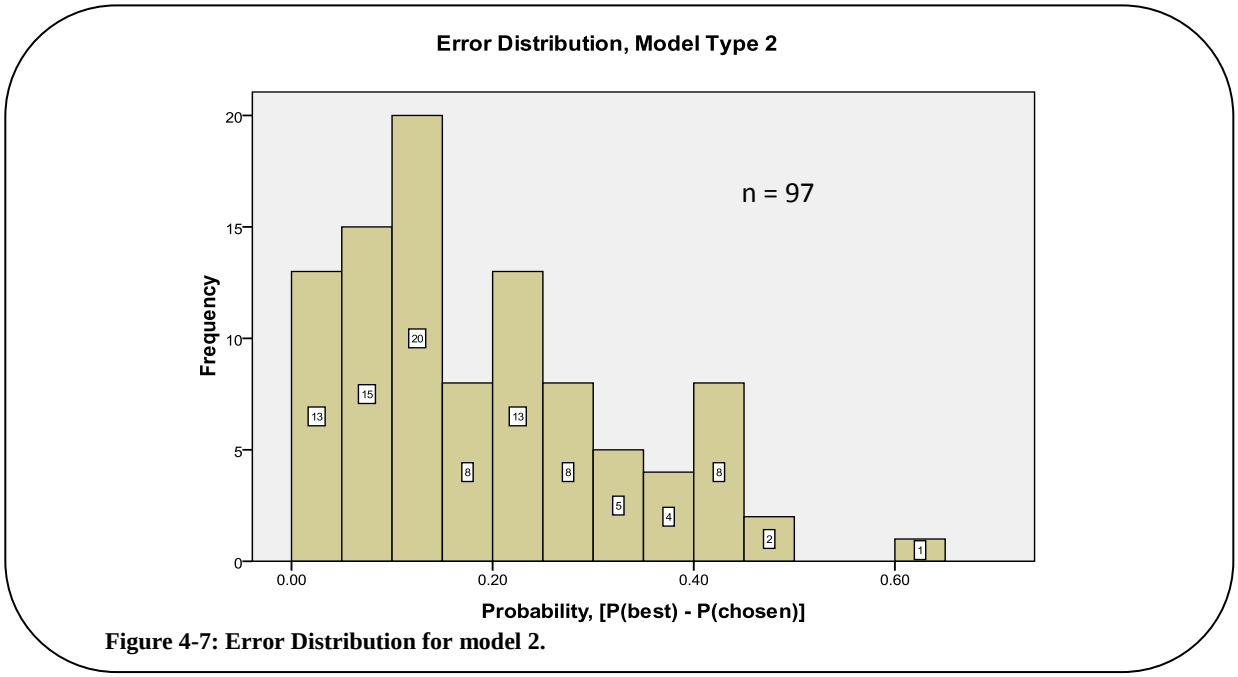


Figure 4-7: Error Distribution for model 2.

The chosen alternative was ranked by the frequency of occurrence and the results are presented in Table 4.13. According to this table nearly 17% of the time, the chosen alternative was ranked 4th or 5th. This further confirms the weaker performance of model 2.

Table 4-13: Chosen alternative by the frequency of occurrence

Rank	Frequency
1	46%
2	10%
3	26%
4	10%
5	7%

Outliers were classified by alternative type and the results are summarized in Table 4-14. Based on the provided summary, Alt.2 and Alt. 4 accounted for 89% of the found outliers.

Table 4-14: Distribution of outliers classified by alternative type.

Alternative	Proportion	Description
2	46%	Road and trails network
3	10%	Road network only
4	43%	Road and trails network (impedance rule)
5	0%	Road network only (impedance rule)

The common outliers are evaluated to the chosen routes.

Chosen Alternative vs. Alt 2.:

- Chosen Alt. was 1 km longer and had less traffic volume, on average;
- Speed and percentage of uphill gradient was not a factor.

Chosen Alternative vs. Alt 3.:

- Length, traffic volume and percentage of uphill gradient variables were not a factor;
- Cyclists chose paths along the roads with a higher speed, which decreased their utility;

Chosen Alternative vs. Alt 4.:

- Chosen Alt. offered routes shorter by 1 km but with drastically steeper sections of uphill gradient, on average;
- Traffic volume and speed variables were not a factor;

Chosen Alternative vs. Alt 5.: Not applicable.

The summary of outlier analysis is presented in Table 4-15.

Table 4-15: Summary of non-chosen alternatives vs. chosen

Alt. No.	Length	Speed	Volume	Gradient (%)
2	Shorter by 1 km	Not significant	Less	Not significant
3	Not significant	Less	Not significant	Not significant
4	Longer by 1 km	Not significant	Not significant	Less
5	No response	No response	No response	No response

89% of outliers were classified to belong to Alt.2 and Alt.4. Chosen routes in comparison to Alt.2 offered longer paths that were traversed along the roads with the lower traffic volume; other factors did not play a role. This kind of behavior can be explained by safety considerations. Chosen paths in comparison to Alt.3 were almost identical, except where cyclists traveled along higher traffic speed roads. The chosen alternatives compared to Alt.4 offered shorter paths but with sections of steep uphill gradient, on average. This behaviour was explained by a segment of respondents who minimize the trip length at the expense of the physical effort required to cover it.

4.5 Test of IIA assumption

This test validates the underlying assumptions of the multinomial logit framework: model structure itself. The assumption on the model framework was tested by randomly removing one of the alternatives from the choice set and then re-estimating the parameters. Alt.2 was removed and new model parameters for model 1 and 2 were obtained. The results are presented in Table 4-16. The changes in the parameters for both model types were within one standard error (Table 4-16). Since model parameters were not changed significantly in the statistical terms, the model's adequacy was considered valid.

Table 4-16: Newly estimated parameters of model 1 and 2.

	Model Type 1		Model Type 2	
Generic Parameters	Estimated Values	*t*-Statistic	Estimated Values	*t*-Statistic
Leng_KM	-0.3862	-9.6836	-0.258	-6.0158
Speed	X	X	-0.7818	-14.851
Volume	X	X	0.0032	8.1461
Elevation Differenc	-1.2661	12.8554	-0.3182	-1.255
Bike Lane	4.9281	-5.3207	X	X

The statistical summary of both model types is presented in Table 4-17. The adjusted ϱ^2 index, is above 0.15. This result indicated that both models were acceptable.

Table 4-17: Summary of model 1 and 2.

	Model Type 1	Model Type 2
Log Likelihood at Zero	-999.5182	-999.5182
Log Likelihood at Constants	0	0
Log Likelihood at Convergence	-760.5935	-649.476
Rho Squared w.r.t. Zero	0.239	0.3502
Rho Squared w.r.t Constants	--	--
Adjusted Rho Squared w.r.t. Zero	0.236	0.3462

4.6 Discussion of results

The results of route-choice modeling are summarized next. The data exploratory analysis and model building process found that bicycle lane and speed effects were confounded. Therefore two models were developed. Model 1 was described by trip length, bicycle lane presence and percent of uphill gradient. Model 2 includes trip length, traffic speed, traffic volume and percent of uphill gradient. Both models generated correct utility signs and the parameter statistics showed goodness-of-fit. Cycling behaviour described by model 1, represents dedicated users of cycling infrastructure. Therefore strategic investment in cycling infrastructure should minimize trip length, increase directness of travel and offer additional safety.

The behaviour described by model 2, describes either experienced users who traveled along busy streets, or a lack of cycling infrastructure that forced cyclists to use regular roads with high speed and traffic volume. The only common feature among the models was evidence that cyclists were minimizing the length and avoiding sections of steep uphill gradient. Therefore the building of new infrastructure should apply these findings.

The scientific part of the model building process included numerous statistical tests which were carried out to validate these models. Tests can be grouped into three categories. The first category took model structure as-is and performed formal and informal tests of utility parameters. The second category did not consider model structure as-given and tested the applicability of the model structure. The third category evaluated the predictive power of the models and outliers. In regards to the first category, rigorous statistical tests validated the model parameters and the overall fit. For the second category an IIA test was performed, and based on the performed statistical analysis, the change in the parameters of both model types was within one standard error. Therefore the multinomial logit framework is an appropriate choice for modeling the route-choice problem. In regards to the third category, predictive tests revealed an interesting pattern. Model 1 in comparison to model 2, with a slight adjustment predicted 78% of data while having lower statistical qualifications. This outcome confirmed that the model building process is as much an engineering judgement and science, and statistical qualifications alone cannot be used as a final judgement to test the quality of a model. Model 1 demonstrated superior predictive results. The outlier analysis was also performed. The exploratory analysis of outliers related to model 1, found that 72% of outliers were explained by either inertia or habit. No rational grounds provided a justification for the actual choice of cyclists. The remaining 28% represents travelers who did not oppose the physical effort of steep gradients, if it offered shorter

paths. The exploratory analysis of outliers related to model 2, found that that 57% of outliers were explained either by inertia or habit, as cyclists preferred longer routes with higher traffic speed. The remaining 43% of outliers were explained by cyclists who chose shorter paths at the expense of facing a steep uphill gradient.

5 Chapter 5 - Transferability of the Study to the Peel Region.

5.1 Introduction

North American cities are experiencing traffic congestion problems as the demand for transportation is increasing. The growing demand for transportation can be accommodated in a number of ways, including investing in public transit, as it offers scalability, and changing land uses. Planners and engineers put in place traffic demand management policies for the better organization of traffic flows, as well as the introduction of non-motorized modes and other policies and techniques. Cycling, being a part of the non-motorized modes, has received considerable attention from many municipalities due to the many advantages that it offers. Generally, municipalities which have higher rates of cycling, spend less on the transportation infrastructure, are economically more viable and have healthier communities.

Peel Region and University of Waterloo partnered on a joint project to gather data on cyclists to have a better understanding of them. Peel Region used the same method for data collection as was described for the Region of Waterloo. Due to the similarity of both regions, Waterloo and Peel, it was proposed to evaluate the applicability of Waterloo's models to the Peel Region's data. Both regions are very auto-oriented, with extensive low-density suburbs; both regions are allocating considerable financial resources to build a cycling network and to increase percentage of choice cyclists. The application of Waterloo's models to the Region of Peel data is regarded as a *transferability* study.

The Region of Peel is comprised of three cities - the City of Mississauga, the City of Brampton and the town of Caledon. The summary profile of Peel Region and cycling infrastructure is provided in Table 5-1.

Table 5-1: Summary profile for Peel Region.

Region of Peel	
Population (2011)	1,296,814
Employment (2006)	541,995
Area	1,246.89 km^2
Population density	1,040.0 prs / km^2 (10.4 prs / ha)
Cycling Data for the Region of Peel	
Bicycle Lane	27.4 km
Regional Trails (Hiking Trails)	208.5 km
Marked Bicycle Route	274.4 km
Paved Multi-Use Trail	573.2 km
Unmarked Dirt Trail	12.4 km
Unpaved Multi-Use Trail	94.3 km
Cycling mode share (commuting - 2006)	0.32%

The location of Region of Peel in relation to Toronto is provided on Figure 5-1. Peel Region introduced several programs to promote cycling, among them was a "Walk and Roll Peel Region" program. The program involved increasing the awareness of the existing infrastructure, to increase safety through education, to use best practices, to engage the public in discussions, and share news about active transportation projects.

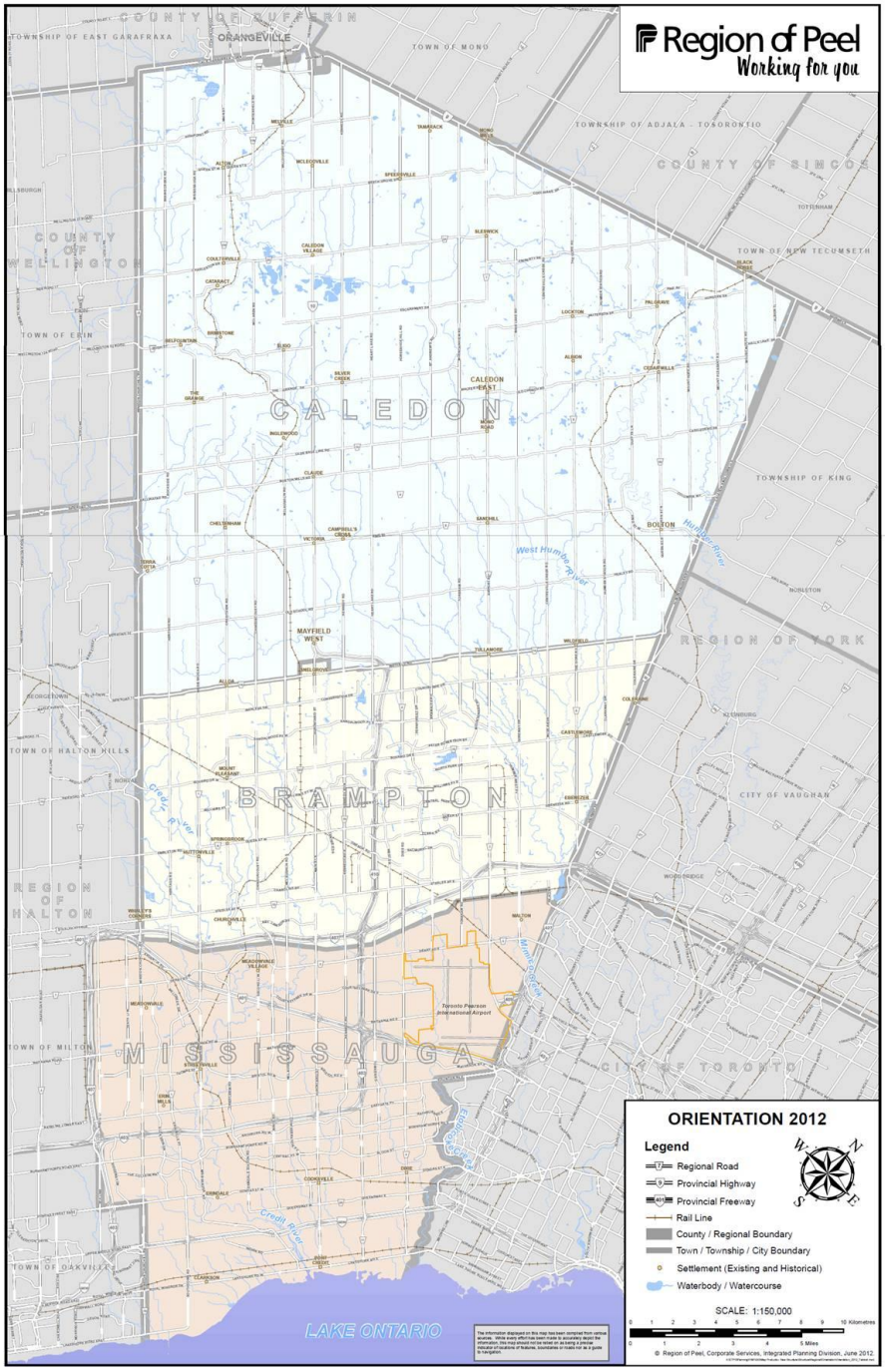


Figure 5-1: Peel Region: Mississauga, Brampton and Caledon.

5.2 Data collection and exploratory data analysis

The University of Waterloo research team designed two programs to collect data in Peel Region, which included a web-based survey and the GPS data on the chosen cyclists' routes. A total of

224 valid web-based answers were collected. This data included socio-economic, demographic, households composition, motivation and obstacles to cycling in the Region. The second part of the data collection included the use of GPS devices by individual cyclists. A total of 255 cyclists participated in the study which occurred between July and September, 2012. A total of 425 routes were selected for the transferability study.

The original exploratory data analysis are summarized here in brief. The analysis of socio-economic characteristics revealed that 77% of cyclists were male and 23% female; 32% of cyclists were travelers of 51 years or older which meant that age was not an obstacle to cycle; 28% of cyclists indicated that their average income was $100,000 or more, which suggested that financial constraints did not limit these cyclists in the mode selection. The number of cars per participant's household was 1.29 and the average for Ontario is 1.48. The main motivation for cycling was physical fitness. A survey question asked respondents to indicate if cycling was more convenient than other modes. The results were not supportive of this conclusion. This indicates that cycling does not compete with other modes at the moment. The main obstacles to cycling are mostly concerns related to the safety and vehicle traffic speed. This means that cycling ridership can be increased if safety is improved. Safety on the other hand can be increased by providing infrastructure specifically dedicated to cyclists, such as bicycle lanes, trails, etc. Cyclists were also asked about their alternatives if cycling was not an option and 68% of respondents said that an automobile could substitute these trips. This response indicated that respondents were choice-cyclists as they had other means to travel but they elected cycling. Excess travel was also evaluated and for 17% of the trips, travel along the roadway network was found longer by 50%, in comparison to what the direct distance was offering.

5.3 Method

The alternative generation methodology presented in Chapters 3 is also applicable to this study. Gathered OD pairs were used to obtain a set of feasible alternatives for every participating cyclist. These alternatives, plus the chosen routes, made up the choice set for travelers. Once alternatives were obtained, they were referenced to road attributes by means of the feature relation procedure. As a result, tables of data were downloaded and then saved in ASCII file format. Then individual links were averaged over the length and aggregated into routes. Although 425 routes were collected only 300 of them were selected for the further analysis. Recreational trips, trips shorter than 500 meters and trips whose length was missing for more than 40% of the actual length were removed from analysis.

In the research conducted for the Region of Waterloo, two route-choice models developed. The first model was specified in the following form:

$$\boldsymbol{Utility_{OD}} = -0.1818 \times (\boldsymbol{Length}) + 4.3081 \times (\boldsymbol{Bike\ Lane}) - 1.4864 \times (\boldsymbol{Elev.Diff}) + \boldsymbol{\varepsilon} \qquad \text{eq.5.1}$$

The coefficient associated with each term represented the amount of changes in utility or preference that would result from a unit change in that attribute. Thus the utility equation is interpreted as follows: as the length of a trip increased by 1 km, the utility decreased by 0.18; presence of bicycle lane increased utility by 4.3; utility decreased by 1.48 for each percent of uphill gradient cyclists faced. Due to the data scarcity, the second model was not verified.

5.4 Results and outlier analysis

The model presented in eq.5.1 was applied to the data gathered in Peel Region. When this model was applied, 37% of the trips were predicted correctly, meaning that 37% of data was explained by the model. It was important to analyze the situation when the chosen alternative was not

predicted by the model, despite the fact it was the observed choice of a traveler. The probability differences between the model's best prediction and the actual choice for each OD pair were calculated and summarized on Figure 5-2 and Table 5-2.

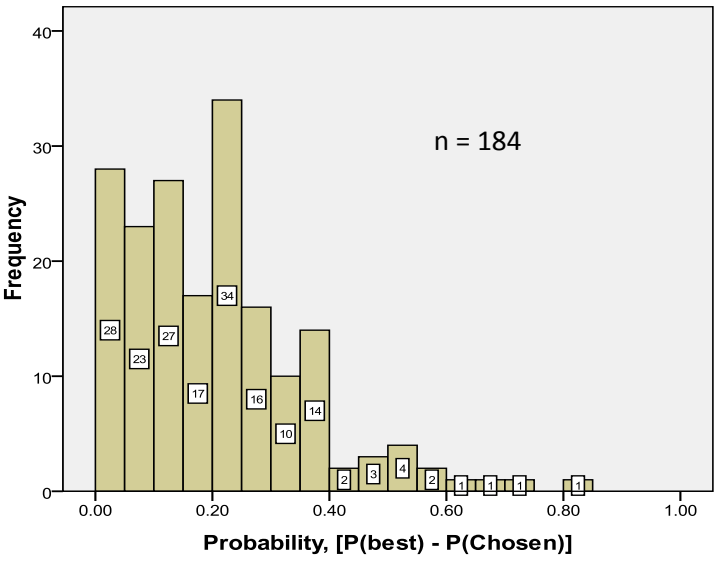


Figure 5-2: Error distribution for model 1.

Information provided in Table 5-2 suggests that if the acceptance threshold was increased to 0.05 level then model could have explained 48% of data.

Table 5-2: Percent of correctly forecasted values for various probability levels.

Probability Class	% Correctly Forecasted
0.00 - 0.05	48%
0.05 - 0.10	56%
0.10 - 0.15	65%
0.15 - 0.20	70%
greater than 0.2	100%

The chosen alternative was ranked by the frequency of occurrence and the results are presented in Table 5-3. From this table nearly 35% of the time, the chosen alternative was ranked 4th or 5th. This result signified that transferability of model coefficients was of a limited success in explaining the behaviour of cyclists in Peel Region.

Table 5-3: Chosen alternative by the frequency of occurrence

Rank	Frequency
1	37%
2	18%
3	17%
4	17%
5	18%

Outliers were classified by alternative type and the results are summarized in Table 5-4. Based on the provided summary, Alt.2 and Alt.4 contributed to 97% of the found outliers.

Table 5-4: Distribution of outliers classified by alternative type.

Alternative	Proportion	Description
2	32%	Road and trails network
3	1.5%	Road network only
4	65%	Road and trails network (impedance rule)
5	1.5%	Road network only (impedance rule)

The common outliers are compared to the chosen routes.

Alt 2 vs. Chosen Alternative:

- Alt.2 offered shorter by 1 km paths, on average;
- Alt.2 offered paths along bicycle lanes, on average;
- Percent of uphill gradient did not play a role;

Alt 3 vs. Chosen Alternative:

- Length and bicycle lane variable did not play a role;
- Alt.3 offered less steep paths, on average;

Alt 4 vs. Chosen Alternative:

- Alt.4 offered longer by 700 m paths, on average.
- Bicycle lane variable did not play a role;

- Alt. 4 offered less steeper paths, on average;

Alt 5 vs. Chosen Alternative:

- Trip length did not play a role;
- Alt.5 offered paths along bicycle lanes;
- Alt.5 offered less steep paths, on average;

The summary of outlier analysis is provided in Table 5-5.

Table 5-5: Summary of non-chosen and chosen alternatives

Alt. No.	Length	Bicycle lane	Gradient (%)
2	Shorter by 1 km	More	Not significant
3	Not significant	Not significant	Less
4	Longer by 0.7 km	Not significant	Less
5	Not significant	More	Less

Although alt.2 and 4 were both obtained using the road and trails network, these outliers have to be evaluated separately. Alt. 2 in comparison to the chosen routes offered shorter paths by 1km and most of these paths had a bicycle lane present. Chosen type of behaviour was explained by unawareness of travelers of the cycling infrastructure. It is evident that cycling infrastructure was available to cyclists but they did not opt to use it because they did not know it was available. Therefore efforts in promoting public awareness of cycling infrastructure should continue. Alt.4 in comparison to the chosen routes offered moderately longer paths with less steep sections and in these cases presence of a bicycle lane did not play a role. Chosen type of behaviour represented physically fit cyclists who were not opposed to more challenging trips if they were shorter. The advantage of Alt.3 and 5 combined, represented a marginal effect as they contributed to 3% of the total number of found outliers. The only common feature among these alternatives was that they offered paths with less steep sections. Other factors did not play a role.

5.5 Discussion of results

This chapter outlined the results of the transferability study which was the first of its kind. The model developed for the Region of Waterloo was applied to the data gathered for Peel Region. Based on the performed study, it was found that 37% of Peel routes could be explained by Waterloo's model, hence the outcome of this study was considered to be of a limited success. One of the most plausible reason for the observed results is the relative significance of parameters which were determined during the statistical analysis. Since the presence of cycling infrastructure in Peel Region was relatively low, the bicycle lane variable did not play a major role in explaining cyclists, as it was in Waterloo. In support of this statement, the goal of "Walk and Roll Peel Region" program should be to increase the public awareness of the existing cycling infrastructure. Another explanation echoing public unawareness of the cycling infrastructure comes from the outlier analysis. Based on the performed outlier analysis, it was found that 32% of the outliers were represented by Alt. 2, which offered shorter paths with higher frequency of bicycle lanes. It was evident that cycling infrastructure was available to cyclists, but they did not opt to use it. The analysis of the remaining outliers (Alt. 4, 65%) was described by alternatives which offered slightly longer paths with drastically less steep sections. Therefore the chosen trips were not predicted by the model.

6 Chapter 6 - Conclusions and Recommendations.

The thesis developed methodological contributions to model route-choice of cyclists to better understand their behaviour, to find key determinants to cycling and indicate what kind of networks to build, and thus where to make strategic investments. The proposed method should assist all interested municipalities in the implementation of cycling as part of bicycle transportation planning.

The ultimate output from the thesis included following components:

- A method to process GPS data and to generate a choice of feasible alternatives based on sound heuristic rules;
- Guidance on practical statistical tests necessary to validate and interpret a model throughout the model building stage;
- A program to assist with data filtering and aggregation of routes;
- A generation of a utility function that represents cyclists' path choice;
- A transferability study for neighbouring regions, first of its kind;

The key findings are summarized next. Both models fit data exceptionally well as utility signs and statistics were valid. Cyclists, described by model 1, represented dedicated users of cycling infrastructure. Therefore the strategic layout of cycling infrastructure should minimize trip length, increase directness of travel and offer additional safety through separate cycling lanes and trails. Cyclists described by model 2 acted in a twofold manner: either as experienced users who traveled along busy streets, or cyclists for whom a lack of cycling infrastructure forced from the use of regular roads with high traffic speed and traffic volume. The only common feature among

models was that the cyclists were minimizing trip length and percent of uphill gradient. Therefore building of new infrastructure should apply these results.

The prediction analysis discovered that model 1 was predicting 65% of the routes correctly, despite having lower statistical significance than model 2. The outlier analysis of both models provided additional insight into the data. Based on the performed analysis it was found that cyclists' behaviour was explained by a relative significance that each coefficient received. Thus cyclists who opted for a shorter trip length with steep road sections were classified as outliers. The remaining outliers presented behaviour which was explained by habit, inertia, or absence of some variables in the analysis (e.g. number of turns, pavement quality, etc.). This type of behaviour represented scenarios when the chosen routes were longer, had a lower presence of bicycle lanes and steep gradient sections.

The transferability study was performed and is recognized as the first of its kind. Based on the performed analysis it was found that model 1 explained 37% of cyclist route choice in Peel Region. This moderate result was attributed to a low public awareness and low presence of cycling infrastructure in the Region of Peel. The outlier analysis found that majority of outliers were explained by unawareness of travelers to the cycling infrastructure. The remaining part of outliers was explained by a relative scale of model coefficients in the utility model.

6.1 Future work and recommendations.

The future work and recommendations for further development can be suggested to proceed in several directions. First, the derived models should be evaluated for their use in the travel forecasting process by allowing to predict a mode split (e.g. auto, transit or bicycle). Although a number of utility models were developed for private and public modes, no models have been

developed to predict trips conducted by the bicycle mode. Second, different model structures need to be evaluated, including C-Logit. C-logit is considered to be the latest development in the route-choice theory. Other types of models, including non-linear, need to be evaluated as they may describe cycling behaviour much better than derived models. Third, more accurate means of generating a choice set, used to model route-choice behaviour, need to be investigated. In particular estimated model parameters (e.g. model 1, 2) should be evaluated for the purpose of developing more realistic choice of alternatives by substituting them into shortest path algorithm of GIS. Fourth, estimated parameters need to be evaluated for sensitivity towards the travel length. Fifth, current methodology should be assessed for its potential to be extended to model automobile or pedestrian route-choice behaviour. Sixth, other types of parameters can be evaluated such as the geometry of the road (e.g. number of turns), quality of pavement, clearly marked bicycle lanes (pavement with pigmentation), and other variables.

7 References

1. Altman-Hall, L.,M. (1996), Commuter bicycle route choice: analysis of major determinants and safety implications, (PhD Dissertation)
2. Ben-Akiva, M., Lerman, S. (1985), Discrete Choice Analysis: Theory and Applications to Travel Demand. MIT Press Series in Transportation Studies
3. Bliemer, M.C.J., Bovy, P.H.L. (2008), Impact of route choice set on route choice probabilities. Transportation Research Record, 2076, 10-19
4. Bovy, P.H.L., Stern, E. (1990) Route Choice: Way finding in Transport Networks. Kluwer Academic Publishers
5. Bradley, M., A., Bovy, P.H.L. (1984), A stated-preference analysis of bicyclist route choice, Proceedings - PTRC Annual Meeting, London, pp. 39-53
6. Broach, J., Gliebe, J., Dill, J. (2011), Bicycle route choice model using revealed preference GPS data. Transportation Research Board
7. Casello, J., Nour, A., Rewa, Ks., Hill, J. (2011), An analysis of stated preference and GPS data for bicycle travel forecasting. Transportation Research Board
8. Casello, J., Nour, A., Rewa, K., (2012), An analysis of empirical evidence of cyclists' route choice and their implications to planning. Transportation Research Board
9. City of Toronto Bicycle Plan: shifting gears (2001), http://www.toronto.ca/cycling/bicycleplan/
10. Fu, L. (2012), Urban Transportation Planning Modeling. Lecture Notes for course CIVE 444/640, University of Waterloo.
11. Greater Golden Horseshoe, https://www.placestogrow.ca/content/ggh/plan-cons-english-all-web.pdf

12. Heinen, E., Wee van Bert, Maat, K. (2010), Commuting by bicycle: an overview of the literature. Transport Reviews: A Transnational Transdisciplary Journal, 30:1, 59-96
13. Howard, C., Burns, E. (2001), Cycling to work in Phoenix: route choice, travel behaviour and route characteristics. Transportation Research Record 1773 Paper No. 01-2526
14. Hunt, J., D., Abraham., J., E. (2006), Influences on bicycle use. Transportation 2007 34:453-470
15. Hyodo, T., Suzuki, N., Takanashi, K. (1998), Modeling of bicycle route and destination choice behaviour for bicycle road plan. Transportation Research Record 1705, Paper No. 00-1434
16. ISTEA, http://ntl.bts.gov/DOCS/ste.html
17. McFadden. D (1976), The mathematical theory of demand models. In Behavioural Travel Demand Models . P. Stopher and A. Meyburg, North Holland, Amsterdam, pp. 75-96 (*Nobel Prize Laureate*)
18. McFadden. D (1974), Conditional logit analysis of qualitative choice behaviour. In Frontiers in Econometrics. P.Zarembka, ed. Academic Press, New York, pp. 105-142 (*Nobel Prize Laureate*)
19. McFadden. D, Tye, W., Train, K. (1977), An application of diagnostic tests for the irrelevant alternatives property of the multinomial logit model. Transportation Research Record 637: 39-46
20. Meyer, M. D., Miller, E. J. (2001) Urban Transportation Planning: decision oriented approach, 2nd Edition. McGraw-Hill Higher Education
21. Menghini, G., N. Carrasco, N., Schussler, N., Axhausen, K., W. (2010), Route choice of cyclists in Zurich. Transportation Research Part A 44 (2010) 754-765

22. Miller, H. J., and Shaw, S. L. (2001). Geographic information systems for transportations: principles and applications. USA: Oxford University Press.
23. Places to Grow. (2006), Planning for Growth: Understanding the Growth Plan
24. Pickrell, D., Schimek, P., (1998) Trends in personal motor vehicle ownership and use: evidence from the Nationwide Personal Transportation Survey. Federal Highway Administration. http://www-cta.ornl.gov/npts/1995/Doc/publications.html-ssi
25. Prato, C. (2009), Route Choice Modeling: past, present and future research directions. Journal of Choice Modeling, 2(1), pp. 65-100
26. Pucher, J., Komano, C., Schimek, P. (1999), Bicycling renaissance in North America? Recent trends and alternative policies to promote bicycling. Transportation Research Record Part A 33 (1999) 625-654
27. Region of Waterloo Cycling Master Plan (2004), http://www.regionofwaterloo.ca/en/gettingAround/resources/CYCLING_MASTER_PLAN_2004.pdf
28. Rewa, C., (2012), An analysis of stated and revealed preference cycling behaviour: a case study of the regional municipality of Waterloo, (MSc Thesis), http://hdl.handle.net.proxy.lib.uwaterloo.ca/10012/6910
29. Region of Waterloo Cycling Master Plan (2004), http://www.regionofwaterloo.ca/en/gettingAround/resources/CYCLING_MASTER_PLAN_2004.pdf
30. Rybarczyk, G., Wu, C. (2010) Bicycle facility planning using GIS and multi-criteria decision analysis. Applied Geography 30 (2010) 282-293

31. Sener, I., Eluru, N., Bhat, C. (2009), An analysis of bicyclists and bicycling characteristics: who, why and how much are they bicycling?
32. Shafizadeh, K., Niemeier, D. (1993), Bicycle Journey-to-Work: Travel Behavior characteristics and Spatial Attributes. Transportation Research Record 1578
33. Statistics Canada. (2006), Profile of Labour Market Activity, Industry, Occupation, Education, Language of Work, Place of Work and Mode of Transportation for Canada, Provinces, Territories, Census Divisions and Census Subdivisions, 2006 Census.
34. Stinson, M., Bhat, C. (2003), Commuter bicyclist route choice: analysis using a stated and preference survey. Transportation Research Record 1828, Paper No. 03-3301
35. Taylor, D., Mahmassani, H. (1996), Analysis of stated preferences for intermodal bicycle-transit interfaces. Transportation Research Record 1556
36. TEA-21, http://www.fhwa.dot.gov/tea21/sumcov.htm

8 Appendix A: Computer Code

```
using System;
using System.Collections.Generic;
using System.Linq;
using System.Text;
using System.IO;
using System.Collections;

// Program is written by Vladimir Usyukov (vusyukov@gmail.com), MASc candidate of University of Waterloo, Civil Engineering Department, May 2013.
// Program is used to filter and summarize road attribute data, necessary for route choice modeling.

namespace DataAggregation
{
   class Program
   {
      //Write Record into array;

      static void Main(string[] args)
      {
         //Declaration of variables;
         string strLine;
         int counter = 0;
         double sumSegments, sumSpeed, sumBicycle, sumVolume, sumGradient;
         bool flag = false;
         double[,] myArray;
         double[,] routeLength;
         double[,] cleanArray;
         double[,] averagedAttributes;
         double[,] summedAttributes;
         string[] strArray;
         List<int> uniqueRoute = new List<int>();
         char[] charArray = new char[] { '\t', ' ', ',' };

         try
         {
            FileStream aFile = new FileStream("ChoiceSet.txt", FileMode.Open);
            StreamReader sr = new StreamReader(aFile);

            strLine = sr.ReadLine();

            while (strLine != null)
            {
               counter = counter + 1;
               strLine = sr.ReadLine();
            }
            sr.Close();
         }
         catch (IOException e)
         {
            Console.WriteLine("Exception was thrown");
            Console.WriteLine(e.ToString());
            Console.ReadLine();
            return;
```

```
}
//Read data into 2D array


try
{
   FileStream aFile = new FileStream("ChoiceSet.txt", FileMode.Open);
   StreamReader sr = new StreamReader(aFile);

   myArray = new double[counter, 10];
   int removRec = 0;

   strLine = sr.ReadLine();
   for (int x = 0; x < counter; x++)
   {
      strArray = strLine.Split(charArray);

      for (int y = 0; y < strArray.Length; y++)
      {
         myArray[x, y] = Convert.ToDouble(strArray[y]);
      }

      if (flag)
      {
         if (myArray[x, 0] == myArray[x - 1, 0])
         {
            if (myArray[x, 1] >= myArray[x - 1, 2])
            {
               myArray[x, 9] = 1;
            }
            else
            {
               myArray[x, 9] = 0;
               removRec = removRec + 1;
            }
         }
         else
         {
            myArray[x, 9] = 1;
         }
      }
      flag = true;
      strLine = sr.ReadLine();
   }
   sr.Close();

   // Write data into clean array;
   cleanArray = new double[counter - removRec, 10];
   int newRow = 0;

   for (int i = 0; i < counter; i++)
   {
      if (myArray[i, 9] == 1)
      {
         for (int j = 0; j < 10; j++)
         {
```

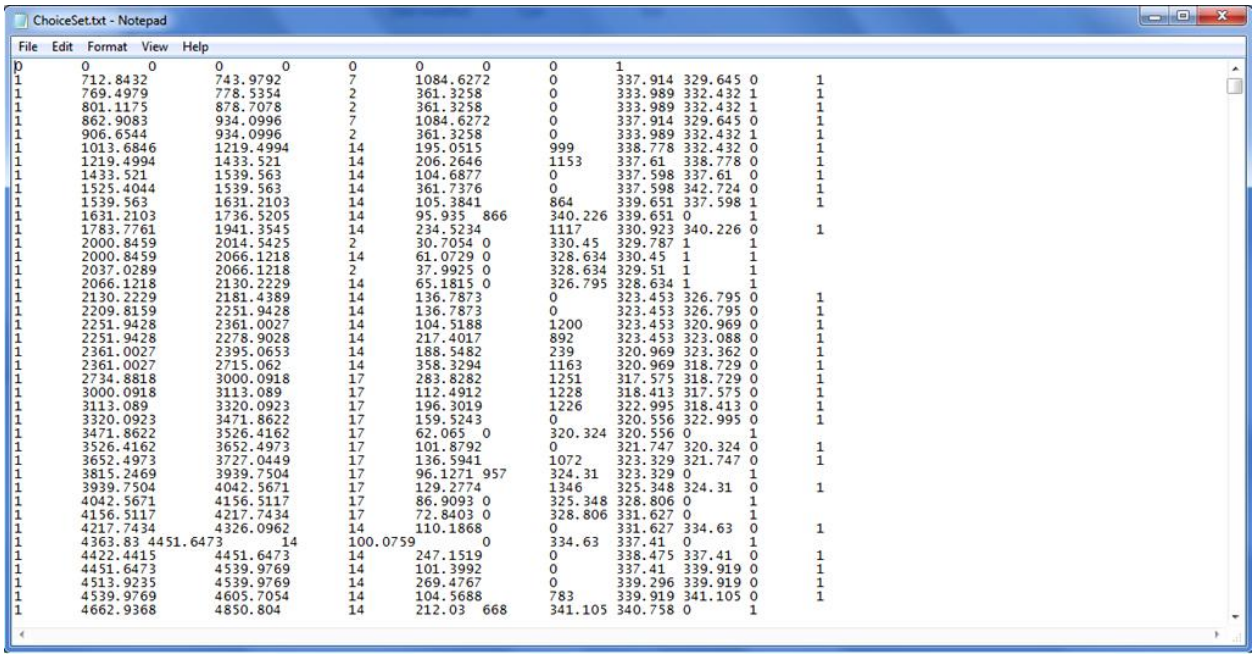
```
            cleanArray[newRow, j] = myArray[i, j];
        }
        newRow = newRow + 1;
    }
    else
        continue;
}

//Perform cleaning of file through N number of iterations;

for (int iteration = 1; iteration < 200; iteration++)
{
    double[,] tempArray;
    int removRec1 = 0;
    int newRow1 = 0;
    uniqueRoute.Clear();
    flag = false;

    //Console.WriteLine(cleanArray.Length/10);

    for (int x = 0; x < cleanArray.Length / 10; x++)
    {
        if (flag)
        {
            if (cleanArray[x, 0] == cleanArray[x - 1, 0])
            {
                if (cleanArray[x, 1] >= cleanArray[x - 1, 2])
                {
                    cleanArray[x, 9] = 1;
                }
                else
                {
                    cleanArray[x, 9] = 0;
                    removRec1 = removRec1 + 1;
                }
            }
            //Adding 1st new record of a new ID: call for a function
            else
            {
                cleanArray[x, 9] = 1;
                uniqueRoute.Add(Convert.ToInt32(cleanArray[x, 0]));
            }
        }
        flag = true;
    }

    tempArray = new double[cleanArray.Length / 10 - removRec1, 10];

    //Put cleanArray into tempArray, excluding flagged values;
    for (int k = 0; k < cleanArray.Length / 10; k++)
    {
        if (cleanArray[k, 9] == 1)
        {
            for (int j = 0; j < 10; j++)
            {
                tempArray[newRow1, j] = cleanArray[k, j];
```

```
                }
                newRow1 = newRow1 + 1;
            }
            else
                continue;
        }
        cleanArray = tempArray;
    }

    Console.WriteLine("Data summary:");
    Console.WriteLine("\tthe original number of records is - {0}, records removed - {1}", counter, counter -
cleanArray.Length / 10);

    //Find length of each route;
    routeLength = new double[uniqueRoute.Count, 2];

    for (int k = 0; k < uniqueRoute.Count; k++)
    {
        flag = true;
        sumSegments = 0;

        for (int j = 1; j < cleanArray.Length / 10; j++)
        {
            if ((cleanArray[j, 0] == cleanArray[j - 1, 0]) && (uniqueRoute[k] == cleanArray[j, 0]))
            {
                sumSegments = sumSegments + cleanArray[j, 4];
            }
            if ((flag) && (uniqueRoute[k] == cleanArray[j, 0]))
            {
                sumSegments = cleanArray[j, 4];
                flag = false;
            }
        }

        routeLength[k, 0] = uniqueRoute[k]; //ID of a route
        routeLength[k, 1] = sumSegments;
    }
    Console.WriteLine("\tnumber of unique routes is - {0}", uniqueRoute.Count);

    //Average road attributes over the length;

    averagedAttributes = new double[cleanArray.Length / 10, 5];
    double deltaElev;
    double hzDistance;

    for (int i = 0; i < cleanArray.Length / 10; i++)
    {
        deltaElev = 0;
        hzDistance = 0;

        deltaElev = cleanArray[i, 7] - cleanArray[i, 6];
        hzDistance = Math.Sqrt(Math.Pow(cleanArray[i, 4], 2) - Math.Pow(deltaElev, 2));

        averagedAttributes[i, 0] = cleanArray[i, 0];//ID of route
        averagedAttributes[i, 1] = cleanArray[i, 4] * cleanArray[i, 3];//Speed in m/s, conversion from km/hr ->
m/s
```

```
            averagedAttributes[i, 2] = cleanArray[i, 4] * cleanArray[i, 8];//Bicycle lane presence
            averagedAttributes[i, 3] = cleanArray[i, 4] * cleanArray[i, 5];//Volume in vph

            if (deltaElev > 0)
            {
               averagedAttributes[i, 4] = 100 * cleanArray[i, 4] * deltaElev / hzDistance;//% of uphill gradient
            }
            else
            {
               averagedAttributes[i, 4] = 0;//We are not interested in negative gradients
            }

         }

         //Sum values from averagedAttributes and average them over the length of a route
         summedAttributes = new double[uniqueRoute.Count, 6];

         for (int k = 0; k < uniqueRoute.Count; k++)
         {
            flag = true;
            sumSpeed = 0;
            sumBicycle = 0;
            sumVolume = 0;
            sumGradient = 0;

            for (int j = 1; j < cleanArray.Length / 10; j++)
            {
               if ((averagedAttributes[j, 0] == averagedAttributes[j - 1, 0]) && (uniqueRoute[k] ==
averagedAttributes[j, 0]))
               {
                  sumSpeed = sumSpeed + averagedAttributes[j, 1];
                  sumBicycle = sumBicycle + averagedAttributes[j, 2];
                  sumVolume = sumVolume + averagedAttributes[j, 3];
                  sumGradient = sumGradient + averagedAttributes[j, 4];
               }
               if ((flag) && (uniqueRoute[k] == averagedAttributes[j, 0]))
               {
                  sumSpeed = averagedAttributes[j, 1];
                  sumBicycle = averagedAttributes[j, 2];
                  sumVolume = averagedAttributes[j, 3];
                  sumGradient = averagedAttributes[j, 4];
                  flag = false;
               }
            }

            summedAttributes[k, 0] = uniqueRoute[k];
            summedAttributes[k, 1] = sumSpeed / routeLength[k, 1];
            summedAttributes[k, 2] = sumBicycle / routeLength[k, 1];
            summedAttributes[k, 3] = sumVolume / routeLength[k, 1];
            summedAttributes[k, 4] = sumGradient / routeLength[k, 1];
            summedAttributes[k, 5] = routeLength[k, 1];
         }
      }

      catch (IOException e)
      {
```

```
            Console.WriteLine("Exception was thrown");
            Console.WriteLine(e.ToString());
            Console.ReadLine();
            return;
        }


        // Writing the output file to dataOutput.txt

        try
        {

            FileStream bFile = new FileStream("dataOutput.txt", FileMode.Append);
            StreamWriter sw = new StreamWriter(bFile);

            sw.WriteLine("Route_ID\tSpeed\tBicyclelanePresence\tVolume\tUphillGradient\tLength");
            for (int k = 0; k < uniqueRoute.Count; k++)
            {
                for (int j = 0; j < 6; j++)
                {
                    sw.Write("{0}\t", summedAttributes[k, j]);
                }
                sw.WriteLine();
            }
            sw.Close();
        }

        catch (IOException e)
        {
            Console.WriteLine("Exception was thrown with output file");
            Console.WriteLine(e.ToString());
            Console.ReadLine();
            return;
        }

        Console.WriteLine("************************************************************");
        Console.WriteLine("Thank you. Result is located in file dataOutput.txt");
        Console.WriteLine("************************************************************");
        Console.WriteLine();
        Console.WriteLine("Press <Enter> key to exit");
        Console.ReadLine();

    }
  }
}
```

8.1 Typical input file




```
0       0       0       0       0       0       0       1
1       712.8432        743.9792        7       1084.6272       0       337.914 329.645 0       1
1       769.4979        778.5354        2       361.3258        0       333.989 332.432 1       1
1       801.1175        878.7078        2       361.3258        0       333.989 332.432 1       1
1       862.9083        934.0996        7       1084.6272       0       337.914 329.645 0       1
1       906.6544        934.0996        2       361.3258        0       333.989 332.432 1       1
1       1013.6846       1219.4994       14      195.0515        999     338.778 332.432 0       1
1       1219.4994       1433.521        14      206.2646        1153    337.61  338.778 0       1
1       1433.521        1539.563        14      104.6877        0       337.598 337.61  0       1
1       1525.4044       1539.563        14      361.7376        0       337.598 342.724 0       1
1       1539.563        1631.2103       14      105.3841        864     339.651 337.598 1       1
1       1631.2103       1736.5205       14      95.935  866     340.226 339.651 0       1
1       1783.7761       1941.3545       14      234.5234        1117    330.923 340.226 0       1
1       2000.8459       2014.5425       2       30.7054 0       330.45  329.787 1       1
1       2000.8459       2066.1218       14      61.0729 0       328.634 330.45  1       1
1       2037.0289       2066.1218       2       37.9925 0       328.634 329.51  1       1
1       2066.1218       2130.2229       14      65.1815 0       326.795 328.634 1       1
1       2130.2229       2181.4389       14      136.7873        0       323.453 326.795 0       1
1       2209.8159       2251.9428       14      136.7873        0       323.453 326.795 0       1
1       2251.9428       2361.0027       14      104.5188        1200    323.453 320.969 0       1
1       2251.9428       2278.9028       14      217.4017        892     323.453 323.088 0       1
1       2361.0027       2395.0653       14      188.5482        239     320.969 323.362 0       1
1       2361.0027       2715.062        14      358.3294        1163    320.969 318.729 0       1
1       2734.8818       3000.0918       17      283.8282        1251    317.575 318.729 0       1
1       3000.0918       3113.089        17      112.4912        1228    318.413 317.575 0       1
1       3113.089        3320.0923       17      196.3019        1226    322.995 318.413 0       1
1       3320.0923       3471.8622       17      159.5243        0       320.556 322.995 0       1
1       3471.8622       3526.4162       17      62.065  0       320.324 320.556 0       1
1       3526.4162       3652.4973       17      101.8792        0       321.747 320.324 0       1
1       3652.4973       3727.0449       17      136.5941        1072    323.329 321.747 0       1
1       3815.2469       3939.7504       17      96.1271 957     324.31  323.329 0       1
1       3939.7504       4042.5671       17      129.2774        1346    325.348 324.31  0       1
1       4042.5671       4156.5117       17      86.9093 0       325.348 328.806 0       1
1       4156.5117       4217.7434       17      72.8403 0       328.806 331.627 0       1
1       4217.7434       4326.0962       14      110.1868        0       331.627 334.63  0       1
1       4363.83 4451.6473       14      100.0759        0       334.63  337.41  0       1
1       4422.4415       4451.6473       14      247.1519        0       338.475 337.41  0       1
1       4451.6473       4539.9769       14      101.3992        0       337.41  339.919 0       1
1       4513.9235       4539.9769       14      269.4767        0       339.296 339.919 0       1
1       4539.9769       4605.7054       14      104.5688        783     339.919 341.105 0       1
1       4662.9368       4850.804        14      212.03  668     341.105 340.758 0       1
```

The correct specification of input file is described next. The first row of data needs to contain zeroes. The order of alternative attributes needs to be arranged in a way how it is presented below. Most of the data can be obtained from linear referencing procedure of GIS software. Then data can be arranged is a described fashion in MS Excel and saved in a text file with a name <**ChoiceSet.txt**> The program has no data limits for the number of records or the number of alternatives as several hundreds of thousands records can be processed with a several seconds.

Column No.	Description
1	is a route number
2	is a beginning node of a route segment
3	is an ending node of a route segment
4	is a speed
5	is a length of a link
6	is a volume on a road link
7	is an elevation of a beginning node
8	is an elevation of an ending node
9	is a bicycle lane indicator on a specific route segment
10	is a flag; initially all values are flagged at 1

8.2 Typical output file

Data processing is done through a computer program by double-clicking on it. At the end of execution, a an output file is created with a name <**dataOutput.txt**>.

Route_ID	Speed	BikelanePresence	Volume	UphillGradient	Length
1	12.94	0.08	437.57	0.55	11597.63
2	14.00	0.00	45.60	0.38	3693.59
3	14.00	0.01	66.15	0.43	3736.97
4	14.00	0.00	365.23	0.50	4604.33
5	14.00	0.00	323.24	1.45	322.06
6	12.22	0.07	454.88	0.73	6718.34
7	8.94	0.40	132.91	0.34	11882.56
8	13.98	0.11	155.14	0.82	4153.77
9	13.98	0.11	240.31	0.77	4947.51
10	10.41	0.42	222.63	0.47	8034.43
11	11.40	0.32	234.73	0.62	8415.92
12	14.00	0.00	36.35	0.51	2168.68
13	13.89	0.00	258.03	0.34	4970.04
14	13.88	0.01	234.20	0.34	4537.13
15	12.23	0.04	358.72	0.65	2102.08
16	14.00	0.00	0.00	0.63	202.37
17	14.00	0.05	44.68	0.23	3769.50
18	14.00	0.01	36.70	0.47	4200.35
19	13.81	0.02	380.37	0.49	4753.14
20	12.73	0.15	45.49	0.62	8704.31
21	13.44	0.01	385.07	0.47	11540.41