A Pedestrian Route Choice Model Concerning Quantified Built Environment Factors

Parham Hamouni

A Thesis

in

The Department

Of

Building, Civil and Environmental Engineering

Presented in Partial Fulfilment of the Requirements

for the Degree of

Master of Applied Science (Civil Engineering) at

Concordia University

Montreal, Quebec, Canada

Summer 2018

CONCORDIA UNIVERSITY

School of Graduate Studies

This is to certify that the thesis prepared

By: Parham Hamouni

Entitled: A Pedestrian Route Choice Model Concerning Quantified Built Environment Factors

and submitted in partial fulfillment of the requirements for the degree of

Master of Applied Science (Civil Engineering)

complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the final examining committee:

Dr. L. Wang Chair

Dr. O. Kuzgunkaya Examiner

Dr. B. Li Examiner

Dr. C. Alecsandru Supervisor

Dr. Z. Patterson Supervisor

Approved by

Chair of Department or Graduate Program Director

Dr.Amir Asif

Dean of Faculty

Date Summer, 2018

Abstract

This thesis builds on a growing body of research that seeks to understand how the built-in environmental attributes of the road network influence pedestrian route choice. Better understanding of these factors can help promotion of walkability. The thesis uses a high-quality GPS dataset of pedestrian trips recorded between October 17 to November 21, 2016, through the MTL Trajet app developed at Concordia University. Trip route characteristics are obtained by matching the GPS traces to a detailed GIS network dataset of road attributes. Additionally, built-in environment factors were captured by scenery quantification and micro-level land use analysis using Google Places API. Scenery was quantified by employing computer vision and machine learning techniques, with help of Google Street View API and deep learning frameworks. A path-size multinomial logit model is used to assess the utility of road and user features. Additionally, to improve prediction accuracy, a set of supervised learning classification techniques, including decision tree, random forest and gradient boosting tree were examined. The analysis of the results shows that the variation in scenery has a significant impact on pedestrians route choice. Additionally, machine learning classification techniques showed significant improvement of the accuracy ratio in comparison to discrete choice modeling framework.

Acknowledgment

I would first like to thank my supervisor Ciprian Alecsandru, for giving me the freedom to explore a topic for which I am passionate. His smartness and wisdom truly helped me in the process of this thesis. I would also like to thank him for the financial support providing during my studies.

I am thankful to Zachary Patterson, the second reader of my thesis and co-supervisor, for the suggestions and questions that strengthened the thesis, and for the provision of data and the infrastructure needed to complete this thesis. I am truly grateful that he believed in me.

I would like to thank Ali Yazdizadeh, for his friendship and mentorship. Without him, it was not possible to finish this thesis. I would also like to thank Kyle Fitzsimmons for helping me learning new skills in computer science. He was true a tutor and a friend for me.

Lastly, I would like to thank my parents for their tremendous support, love and encouragement to pursue my Master's degree, and to Atefeh for her patience, compassion and camaraderie to help me in the process of finishing this thesis. This accomplishment would not have been possible without their inspiration.

Contents

List of Figures

List of Tables

List of Abbreviation

Application Programming Interface	API
Classification and Regression Trees	CART
Classification tree analysis	CTA
Convolutional Neural network	CNN
Cross Nested Logit	CNL
Decision Tree	DT
Discrete Choice Model	DCM
Field of view	FOV
Generalized Extreme Value	GEV
Geographic Information System	GIS
Global Positioning System	GPS
Google Street View	GSV
Gradient Boosting	GBT
Gradient boosting tree	GBT
Independence of Irrelevant Alternatives	IIA
Independent and Identically Distributed	IID
Iterative Dichotomiser 3	ID3
Multinomial logit	MNL
Nested Logit	NL
neural network	NN
Paired Combinatorial Logit	PCL
Path Size	PSL
Path Size Logit	PSL
Random Forest	RF
San Francisco Bay Area Travel Survey	BATS
Support Vector Machine	SVM
United Kingdom	UK
United States of America	USA
Volunteered Geographic Information	VGI

1 Introduction

1.1.1 Thesis Overview

This thesis seeks to better understand the built environment factors and personal characteristics that influence pedestrian route choice in Montreal. The built environment is characterised by adjacent land use and scenery and their effects are investigated. The thesis analyses route trip data collected with a smartphone application and uses discrete choice models and supervised classification techniques to identify contributing factors and the best model in terms of prediction accuracy.

The objectives of this thesis are:

- To measure the effect of built-in environment factors, especially scenery, in a quantitative systematic approach in revealed preference setting by discrete choice modelling;
- To investigate the improvement of prediction accuracy of pedestrian route choice model in frameworks other than traditional multinomial logit;
- To investigate land uses in both macro and micro level and its effect on pedestrian route choice prediction accuracy.

1.1.2 Thesis Motivation

Pedestrian route choice is one of the new merging areas of transportation planning due to sustainable development organizing principles. Globally, huge increases in urbanization have caused a need for redefining urban planning principles. To address this issue, new urban planning theories, such as the New Urbanism have emerged in the literature (Leccese and McCormick 2000). New Urbanism is an urban design movement that supports the promotion of environmentally friendly habits. This end is achieved by multiple principles, including promotion of compact, pedestrian-friendly, and mixed-use neighborhoods (Leccese and McCormick 2000). Therefore; there is an understanding among researchers that facilitating walking in urban agglomerations is a significant task that leads to sustainable development (Talen and Koschinsky 2013). Quantifying factors that affect pedestrian perceptions of walking leads to understanding and promoting walking and walkability. This can be investigated by analyzing when people prefer

to walk instead of using other modes of transportation (mode choice) or analyzing factors affecting route choice (pedestrian route choice). Promoting walkability was one of the motivations to investigate pedestrian route choice.

Another motivation for studying pedestrian route choice is lack of thorough representation of walking trips in transportation demand models. Modeling route choice behavior is important to forecast traveler behavior, to predict future traffic assignment on transportation networks, to understand traveler reaction and adaptation to facilities and information, and to evaluate traveler perceptions of route characteristics (Prato 2009). To develop a model, it is important to first identify important factors and then find the modeling framework that has predictive power and interpretability. Pedestrian route choice models started to be addressed more extensively fairly recently (Broach and Dill 2015; Hintaran 2016; Hoogendoorn and Bovy 2004; Lue 2017). These studies have focused mostly on the impact of physical and geometrical factors of routes (e.g. the length of path, the gradient, number of turns, etc.). However, they mainly overlook other relevant factors that stated preference studies suggest that pedestrian consider attractive for a path. For example, it is stated that scenery plays an important role especially in recreational trips (Bovy and Stern 2012). These factors are mainly overlooked in studies due to the complexity of quantifying pedestrian path surroundings with a model-based approach or lack of data.

Final motivation of this study was to model pedestrian route choice in a revealed preference setting rather than stated. It is shown in the literature that people may behave in a different way than what they state (Wardman 1988). There are few studies shown in the literature that suggest using revealed route choice (Hintaran 2016; Lue and Miller 2018). This study was conducted to model pedestrian route choice in a revealed setting and investigate overlooked relevant factors with a predictive model.

1.1.3 Thesis Approach

This study used revealed preference Global Positioning System (GPS) data collected by a smartphone-based travel survey in Montreal to model pedestrian route choice via discrete choice modeling and machine learning supervised classification techniques. This data is further enhanced by adding geographical and land use information. Additionally, the aesthetic context of pedestrian route choice (i.e. scenery), is deeply investigated in a model-based, quantitative manner, by

employing recent developments in image processing (i.e. deep learning and machine learning). The data is analysed using both discrete choice modeling and supervised classification to analyze variables affecting route choice and additionally, acquire the highest prediction accuracy possible. The thesis involves the following steps:

- Build a network data set of important characteristics for pedestrians
- Convert the GPS traces into trips and match them to the network and develop a series of characteristics about the trips as well as the users;
- Create a choice set of feasible alternative routes;
- Use discrete choice modeling to estimate the relative utility of each road attribute on route choice.
- Use supervised learning models with the same variables as used in the discrete choice modeling setting to compare prediction accuracy of the different approaches.

The remaining sections of the thesis include the following: the literature review describes a background on past work and models related to this topic. The data collection and processing section describes the network, street attributes and smart-phone-based travel survey. The methodology describes the methods. Results chapter discusses the results obtained by the model.

2 Literature Review

This chapter reviews the literature on pedestrian route choice, including proposed route choice modeling methods and the findings of previous pedestrian route choice studies. The first section provides an overview of choice models, pedestrian route choice and factors that have been found to influence pedestrian route choice. Additionally, machine learning frameworks are introduced and their previous application in transportation and route choice problem is discussed.

2.1 Choice models

Discrete choice models are designed to model behavioral processes that lead to a subject's choice. There are many different approaches to capture human choice behavior, and they range from deterministic theories in economics to probabilistic or stochastic models in psychology. In the psychological views of decision making, alternatives are viewed as a set of known aspects. The randomness in choice comes from the decision rule. On the other hand, the economic view of decision making is based on the notion of precedence of desirability over availability. The expressed preferences are functions of the consumer's taste template, experience and personal characteristics. The economic approach, based on the theory of Random Utility Maximization, has been used in transportation as well, (Ben-Akiva and Lerman 1985; Cascetta 2001; McFadden 1986; Simon 1959).

Discrete choice models (DCM) assume that each alternative in a choice experiment can be associated with a latent quantity, a utility. The utility of each alternative is based on multiple aspects (Schüssler 2010) , including:

1- The attributes of the alternative
2- Individual preferences captured via socio-economic proxies
3- The choice situation and its similarities with other available alternatives

Based on the concept of utility-maximization, the individual is assumed to select the alternative with the highest utility, given constraints from his or her activity agenda and risks involved in their decisions.

Representing route choice behavior is modeling the choice of a given route within a set of alternative routes. A route choice model associates a probability to each alternative, and the one with the highest probability is considered chosen (Bierlaire and Frejinger 2008).

There are four elements in each discrete choice model: the choice set, attributes or factors of each member of choice set (alternatives), socio-economic factors to describe the decision-maker and finally, a random term, capturing unobserved error and uncertainties of the choice process (Antonini 2005). In the route choice context, the individual choosing a route is the decision-maker, the choice set is the list of plausible routes, alternative attributes to quantify characteristic of each alternative and finally, socio-economic attributes describe the decision-maker quantitatively. The random term is presented in random utility maximization theory to identify unobserved alternative attributes, unobserved socio-economic characteristics, measurements errors and instrumental variables. For each alternative in the choice set, a utility function consists of two components. These components are V, deterministic part of utility and ε representing the random part of the choice. The formulation is as follows (Manski 1977):

$$U_{in} = V_{in} + \varepsilon_{in} \quad \textbf{[1]}$$

Where V, the deterministic component of the utility, is a function of socio-economic characteristics of the decision-maker and alternatives' attributes. It is defined as $V_{in} = f(\beta, x_{in})$ where β is a vector of coefficients and x_{in} is a vector of attributes of alternative i when n is the individual choosing (Schüssler 2010). Within the random utility model framework travelers are assumed to maximize utility. There are several types of model formulations to solve this probabilistic setting, due to different assumption on the random term.

2.1.1 Multinomial Logit

The Multinomial Logit (MNL) model associates a probability to each alternative of a route based on its corresponding utility (Train 2009). The model has a logit structure, which assumes that the perceived attractiveness of the alternatives is mutually independent. The ratio of the choice probabilities for two alternatives is not affected by the systematic utilities of the other alternatives (Antonini 2005). However, this assumption is not always true for route choice where alternative routes can have correlation due to overlapping paths. Since the error terms in the MNL model are

independently distributed, no correlations are included in the model. Because of the high likelihood of overlap of alternatives in real network, the MNL model is not the best modeling framework. Since model alternatives overlap, they do not hold the property of Independence of Irrelevant Alternatives (IIA). Additionally, it is plausible that different decision-makers have heterogeneous preferences, which is not reflected in MNL (Bliemer and Rose 2010).

To allow the correlation among alternatives, models such as multinomial probit model was introduced (Bouthelier and Daganzo 1979). However, these models need extensive computational effort. To use MNL models which have good computational efficiency, there are some approaches introduced which are discussed briefly.

Overcoming the IIA property is a major research issue in the field of discrete choice modelling. There are various model structures in use to overcome the overlap problem. These model structures can be classified as (Schüssler 2010):

- Introducing adjustment terms in the deterministic part of the utility function (category 1)
- Imposing a nesting structure (category 2)
- Explicitly modeling the correlation using multivariate error terms (category 3)

The first category of models consists of modifications of the Logit structure. These models assume that the utility of an alternative is influenced by its level of similarity with other alternatives and that it can be corrected accordingly (Schüssler 2010). They address similarities by correcting the systematic component of the utility function (V), by adding a deterministic adjustment term that measures the similarity (similarity attribute) to the utility function. The formulation is presented as follows:

$$U_{in} = V_{in} + f(A_{in}) + \varepsilon_{in}$$

A_{in}: adjustment term that measures the similarity between alternative i and all other alternatives j≠i

$f()$: transformation of A_{in}

Using the first class has the advantage of maintaining MNL structure model, which is applicable with reasonable computational difficulty. On the other hand, finding the right transformation of A, (i.e. f) is not straightforward. The C-logit and Path-size Logit (PSL) are suggested to address transformation function f(). They assume that if an alternative is similar to other alternative, its

utility should be reduced and therefore, the probability assigned to this alternative should be adjusted accordingly. These models will be introduced later in this section.

The second class includes generalizations of the Logit structure. Generalizations of the Logit structure have a more complex error structure and are members of the Generalized Extreme Value (GEV) model family. Models of the GEV family take correlation patterns in the choice set into account. The unobserved portions of utility for all alternatives are jointly distributed as a generalized extreme value. This distribution allows for correlations over alternatives (Train 2009). Detailed theory about GEV models can be found in (McFadden 1978). Models derived from the GEV formulation include the MNL (when all correlations are zero), the Nested Logit (NL), Cross Nested Logit (CNL) model and the Paired Combinatorial Logit (PCL). In these models, alternatives of the choice set are subdivided into nests, where alternatives belonging to the same nest are correlated to each other.

Modifications of the logit structure addressed correlation among factors, but they incorporated random taste heterogeneity appropriately. The third category of models handle limitations of the MNL model. The probit model assumes that the unobserved attributes are multivariate normal distributed. In comparison, MNL and other GEV models error terms are assumed to be IID Gumble distributed (Aldrich and Nelson 1984).

Assumptions of the probit model is a limitation as well, since in different setups it may cause the assumption of normal distributions be inappropriate. For example, in transportation context, considering positive coefficient for distance traveled is not intuitive, so assigning a normal distribution to this coefficient with a zero mean is a strong assumption. The mixed logit (logit kernel) model tries to capture properties of both logit and probit model error terms. This model's error terms includes both multivariate randomly distributed portion to account for unobserved attributes (Walker 2001). The reason for this error function is that the probit-portion in the utility function captures the correlation between alternatives. When the cross-alternative correlations in these models are not present, the model reduces to MNL (Bekhor, Ben-Akiva, and Ramming 2006).

As discussed in this section, the first class accounts for correlation between alternatives through a transformation function f(). C-logit, a model of this category, introduces a "Commonality Factor" to correct for the overlap. The commonality factor is proportional to the overlap each alternative to other members of the choice set (Cascetta et al. 1996). The lack of theoretical guidance of this model is an obstacle to apply this model on choice problem (Frejinger 2008).

The Path-size logit was first introduced by (Ben-Akiva and Bierlaire 1999). The utility function of path *i* for a decision-maker *n* is defined as follows and the probability of choosing a path are respectively presented as:

$$U_{in} = V_{in} + \beta_{PS}\ln(PS_{in}) + \varepsilon_{in} \quad [2]$$

$$P(i|C_n) = \frac{e^{\mu(V_{in}+\ln(PS_{in}))}}{\sum_{j\in C_n} e^{\mu(V_{in}+\ln(PS_{in}))}} \quad [3]$$

Where:

C_n: the choice set for user n (includes chosen route)

μ: the logit scale term

V_{in}: systematic utility for alternative i for user n

PS_{in}: the path size factor for alternative i for user n

$$PS_{in} = \sum_{a\in\Gamma_i} \frac{L_a}{L_i} \frac{1}{\sum_{j\in C_n} \delta_{aj}} \quad [4]$$

Where:

Γ_i: the set of links in path i

L_a: the length of link a

L_i: the length of path i

δ_{aj}: a dummy variable, equals 1 if link a is on path j and 0 otherwise

$\sum_{j\in C_n} \delta_{aj}$: the number of paths in choice set C_n sharing link a

The correction factor aims to penalize the paths that overlap one another. If none of the links are present in other alternatives, the PS factor will be 1 and its natural logarithm zero, thus the formulation would be the same as the MNL. However, if the paths overlap, this value will have a value lower than 1 and therefore, its logarithm would be negative, so it will decrease the utility of the corresponding alternative. However, if an unlikely long path has an overlap with a likely one, it would decrease the utility for the likely one, thus being unrepresentative.

There are other formulations that represent the same concept. Equation [5] presents the formulation suggested by (Ben-Akiva and Bierlaire 1999).

$$PS_{in} = \sum_{a \in \Gamma_i} \frac{L_a}{L_i} \frac{1}{\sum_{j \in C_n} \frac{L^*_{C_n}}{L_j} \delta_{aj}} \quad \textbf{[5]}$$

Where C_n is the length of the shortest path in the choice set.

Another formulation was introduced by (Ramming 2002) to account for the impact of unrealistic long paths in the choice set. It is called the 'Generalized PS' which is described in Equation [6].

$$PS_{in} = \sum_{a \in \Gamma_i} \frac{L_a}{L_i} \frac{1}{\sum_{j \in C_n} (\frac{L_i}{L_j})^{\varphi} \delta_{aj}} \quad \textbf{[6]}$$

Where φ is an arbitrary parameter, controlling how route length would impact the correction factor. There are studies that show how to set this factor. It was concluded that φ=14 would provide the best fit (Hoogendoorn and Bovy 2004). If the φ is set to zero, the formula would be similar to [4].

There is another PS factor, introduced by (Bovy, Bekhor, and Prato 2008). It is represented in [7]. The main difference with [4] is the placement of logarithm in the formula. The authors showed it has more theoretical ground to weight it in this way.

$$PS_{in} = \sum_{a \in \Gamma_i} \frac{L_a}{L_i} ln(\frac{1}{\sum_{j \in C_n} \delta_{aj}}) \quad \textbf{[7]}$$

Additionally, there are other types of PS factors such as Expanded Path Size Logit. This is used when choice set is generated stochastically (Frejinger, Bierlaire, and Ben-Akiva 2009). In this thesis, it is not investigated because choice set generation is not stochastic.

2.2 Choice modeling factors

Choices in transportation can be categorized into strategic-, tactical- and operational levels. The strategic level refers to departure time and activity pattern choice. The tactical level relates to activity scheduling, activity area choice and route choice to reach activity areas; and operational level to walking behavior. Thus, pedestrian route choice can be considered as operating at the tactical level. Discrete choice models are widely used in transportation engineering in route choice (Hoogendoorn and Bovy 2004).

The route choice decision-making process can be categorized into two main sequential activities, route generation and route choice. Route generation refers to determining possible routes between preset origin and destination locations of the trip (i.e. a candidate set of routes of alternatives). Route choice is the mechanism of selecting one item from the candidate set. Previous studies suggested that trip-makers will select from among no more than six alternatives (Bovy and Stern 2012). The process of route choice is mainly explained through utility maximization framework, specifically through logit models (Ben-Akiva and Lerman 1985).

There are multiple factors affecting route choice, one of which can be categorized into five distinct categories. Network characteristics, route characteristics, personal characteristics, trip characteristics and environmental characteristic are the main set of factors used to model route choice (Daamen 2004). Some possible parameters associated with these factors are presented in Table 1. Note that in this thesis the words feature, factor, variable and attribute are used interchangeably.

Table 1 - Factors influencing route choice

Types of factors	Variables
Network characteristics	Number of available routes and overlapping routes
Route characteristics	Travel time and distance, scenery, directness, crowdedness, safety factors, weather protection, road type and gradient

Personal characteristics	Age and gender
Trip characteristics	Trip purpose, time budget, mode used and departure time
Circumstances	Weather conditions, road and traffic information and road works accidents on the route and day or night

While route choice factors are important for all modes, for each mode, they weigh differently in the route choice model. For example, route choice of drivers is mainly based on travel time while pedestrian route choice is mainly based on physical effort rather than travel time (Bovy and Stern 2012). There have been multiple studies investigating the factors for pedestrian route choice (Hintaran 2016; Hoogendoorn and Bovy 2004), which were evaluated in the following section.

2.3 Pedestrian route choice factors

Pedestrian route choice modeling in a precise way is a challenging process; due to the complexity and largely subconscious nature of the problem (Hill 1982). Several studies, using surveys, found that trip length is the dominant factor influencing the pedestrian route choice (Guo and Loo 2013; Seneviratne and Morrall 1985; Verlander and Heydecker 1997; Van der Waerden, Borgers, and Timmermans 2004; Weinstein Agrawal, Schlossberg, and Irvin 2008). However, studies have shown that the shortest path may only be chosen around 20 percent of the time (Borst et al. 2009). Additionally, the number of turns and gradient are also significant factors in route choices (Broach and Dill 2015). Transportation and land-use impacts on mode and route choice do not show consistent results (Badoe and Miller 2000), as some studies identify them as significant factors while other studies illustrate as marginal. Other significant factors that have been reported earlier in literature are related to the built environment and trip safety (Brown et al. 2007; Weinstein Agrawal et al. 2008).

The built environment can be defined as the human-made space in which people live, work, and recreate on a day-to-day basis (Roof 2008). It is a multidimensional concept that can be perceived in five dimensions, which are density and intensity, land use mix, street connectivity and aesthetic qualities. Aspects of the built environment can be measured in three categories, which are observed measures, geographic measures and perceived measures (Brownson et al. 2009). Observed measures mainly measure the physical features of the environment such as sidewalk width, street

slope/grade, land use frontage. The observed data is obtained by observers, like surveyors or data collectors, which may lack enough accuracy. The observational data is collected in lack of geographical measures. Geographic measures are mainly collected in zonal levels, and they include dimensions such as population density, land-use and street network. They are mainly set in geographic information system (GIS) programs. The main limitation of geographic measures is a lack of consistency between datasets needed for studies, because each dataset may be collected by a different agency, making it hard to be used together. Additionally, they may lack temporal consistency as well. In this study, the consistency of geographic measures has been considered by collecting data from the same source as much as possible. The perceived measures can be defined as people's perception of the built environment. Its attributes include aesthetics, sounds, and safety. The main challenge using these measures is that it needs surveys which are lengthy and hard to incorporate in transportation models. For example, scenery is a perceived built environmental attribute which is shown as an attractive street characteristic in stated preference studies. These studies are often qualitative in describing scenery, and are mainly based on the stated preferences of the respondents.

Researchers examined if features of the built environment in a micro level, such as width of the sidewalk, benches, trash bins, crossing aids: stoplights and crosswalks, had correlation with street segment pedestrian activity (Rodriguez, Brisson, and Estupinan 2009). They concluded greater pedestrian activity on segments are related with higher development intensity, mixed land uses, and more crossing aids. It was also noted that street connectivity and pedestrian friendly aids are related to higher pedestrian counts (Rodriguez et al. 2009).

Other studies have found that higher density of intersections, wider sidewalks, higher density of pedestrian friendly parcels are associated as attraction for a route while attributes such as large street crossings, poor lighting, litter, absence of people, or steep slopes act as deterrents (Ferreira et al. 2016). For example, in a study exploring the relationship between street characteristics and perceived attractiveness for elderly residents, concluded that attributes such as low slopes, zebra crossings, trees, gardens, bus stops, business buildings, catering establishments, city centre, and traffic volume are associated with attraction. While attributes such as litter, high-rise buildings, high neighborhood density were negatively related to perceived attractiveness (Borst et al. 2009).

In a study, built environment factors on short walking trips (less than 45 minutes) were investigated in Valencia, Spain. It was a stated preference study, consisting of three focus groups of non-shopping trips during the week. Some factors were unanimously considered positive, such as sidewalk width, the presence of trees, and low traffic volumes. Additionally, factors relating to safety, such as poor lighting or absence of people, were strong deterrents for walking for all groups. However, other factors such sidewalk cafes and bollards are considered as aesthetic improvements by some participants, while others found them as deterrents (Ferrer, Ruiz, and Mars 2015).

In a stated preference study, it was indicated that the primary factor is minimizing time and distance for pedestrian route choice. Additionally, safety, crossing delays, sidewalk conditions, a presence of other pedestrians were considered important. The participants of this study were morning commuters at five rail stations in San Francisco and Portland, Oregon (Weinstein Agrawal et al. 2008).

In another stated preference study, it was found that shops, good scenery, and crowdedness can also play a role in pedestrian decision making (Puay Ping Koh and Wong 2013). Another study by the same authors investigated influence of infrastructural compatibility on pedestrian route choice (P P Koh and Wong 2013). It was conducted to investigate which one of the following factors were considered important: distance, comfort, rain shelters, stairs/slopes, traffic accident risk, detour, crowded walkway, security, the number of road crossings/delay, shops along the route, good scenery, and directional signs. The pedestrians at transit stations in Singapore were interviewed regarding their preferences for the first/last mile of their walking trips. It was found that in different areas, the important factors could vary. For example, traffic accident risk, rain shelters and stairs/slopes are associated with greater importance in residential areas than in mixed land use areas. In residential areas, distance had the top priority, with availability of public transit and convenience of walking being influential. Mixed land use was found to be very similar to residential area. In industrial areas, traffic accident risk was also an important factor.

Pedestrian environment and its effect on the utility of walking is analyzed in a quantitatively as well, for example, the case study subway commuters' paths from the station to their workplace in downtown Boston (Guo 2009). The results showed more intersections, wider sidewalks and flat topography have a positive effect on utility (Guo 2009).

In another study, researchers have explored revealed preference in pedestrian route choice using GPS data (Broach and Dill 2015). They studied attributes such as distance, turns, steep upslope, substandard street, busy streets, commercial neighborhoods, unsignalized arterial crossings, and unmarked collector crossing. It was found that turns, upslopes, busy streets and unsignalized intersections were associated with negative utility. For example, an upward slope of 10 percent is regarded twice costly as less steep ground. The commercial nature of a neighborhood had positive impact on utility, as being considered comparable to 27% shorter trip.

The extent to which scenery in general is significant may depend on trip characteristics. For example, scenery is expected to be a significant factor for recreational trips, but it plays limited or no role for work-related walking trips (Bovy and Stern 2012). Some other studies found that the attractiveness of buildings (Guo and Loo 2013) can have positive impact as well. Therefore, understanding and clustering trip characteristics may lead to more accurate and reliable route choice models and related attributes (Bovy and Stern 2012; Hill 1982; Seneviratne and Morrall 1985).

2.4 Scenery in transportation context

The importance of scenery, which is one of the factors associated with pedestrian route choice, has been noted by several studies (Owen et al. 2004). The extent to which scenery plays a substantial role in route choice behavior on trip purpose. Scenery is very important for recreational trips, but it plays no role for work-related walking trips (Bovy and Stern 2012). However, its investigation has been challenging, because of the complexity of quantifying it and its subjective nature. The way scenery was mainly addressed is that it was incorporated in stated preference studies and participants were asked questions which indirectly assessed aesthetic characteristics and found positive impact of scenery on route choice (Ball et al. 2001; Puay Ping Koh and Wong 2013). However, the extent to which it can have a positive impact has not been estimated in a revealed preference context. Few studies have tried to evaluate scenery in a systematic way in transportation context, not necessarily pedestrian route choice (Alivand and Hochmair 2013; Chen et al. 2017; Quercia, Schifanella, and Aiello 2014). They are evaluated in the following section.

Researchers have tried to address scenic route planning for drivers using geo-tagged pictures obtained from services such as Panoramio and Flickr (Zheng et al. 2013). Their main assumption

is that a multitude of photos taken along a roadway imply that this roadway is probably appealing and catches the public's attention. In another study, researchers used Volunteered Geographic Information (VGI) data sources such as Panoramio and Flickr and websites where users uploaded tracks of traversed scenic routes (RouteYou, EveryTrail, and MyScenicDrive) (Alivand and Hochmair 2013). They assumed that users upload several scenic pictures during a day trip that are located along a route, it could be concluded that the traversed route is scenic. This assumption is derived from a work another work by Hochmair (Hochmair 2010). The roads that were already considered scenic by web services such as EveryTrail or GPSies, which were compared to the shortest path with respect to number of geo-tagged photos on Web 2.0 applications. The results show that the frequency was greater along scenic routes than along fastest routes. The study was based on the simple idea of null-hypothesis testing of whether number of geo-tagged photos found along scenic and fastest routes is equal, or that it is even higher for fastest route. The results showed photos obtained from obsolete Panoramio service show a higher spatial association with user posted routes.

Researchers attempted to integrate pedestrian perceptions of the urban context into their route choice model and specifically, route generation (Quercia et al. 2014). This approach is inspired from psycho-geography, which is defined as the study of the precise laws and specific effects of the geographical environment, consciously organized or not, on the emotions and behavior of individuals. Even though emotional responses are subjective and difficult to quantify, urban studies have repeatedly shown that specific visual cues in the urban contexts are consistently associated with attractiveness of city elements. For example, several studies (Peterson 1967; Salesses, Schechtner, and Hidalgo 2013) found that green spaces are mostly associated with attraction, while trash and broken windows with distaste.

Quercia et al. approached quantifying pleasantness of urban context through a crowd-sourcing platform that showed two street scenes in London (out of hundreds), and users voted on which one looks more beautiful, quiet, and happy (Quercia et al. 2014). Then they assigned scores to locations along each of the three dimensions. The next step was that to generate routes regarding these dimensions. Figure 1 depicts the authors' results of the proposed model given different route choice criteria. The results of their proposed model were validated through a stated preference study at two locations, in London, UK and in Boston, USA. To generalize the approach, they used

Flickr data to model beauty score obtained from the users. They suggested a regression model that could describe more than 30% of the variability of the beauty score by the presence of Flickr tags.

While the model may be used to explain some patterns in choice to travel around the city of London, nevertheless some limitations can be identified. For example, this route generation does not account for the trip purpose, therefore it cannot be used to capture the relationship between different types of trip purposes (e.g. work, school, etc.) and the expected route choice, as it is expected for a complex transportation problem. Additionally, the effect of working vs non-working hours were not investigate. Furthermore, their scenery prediction model shows a low R-square, which can degrade the generalization abilities of the model.

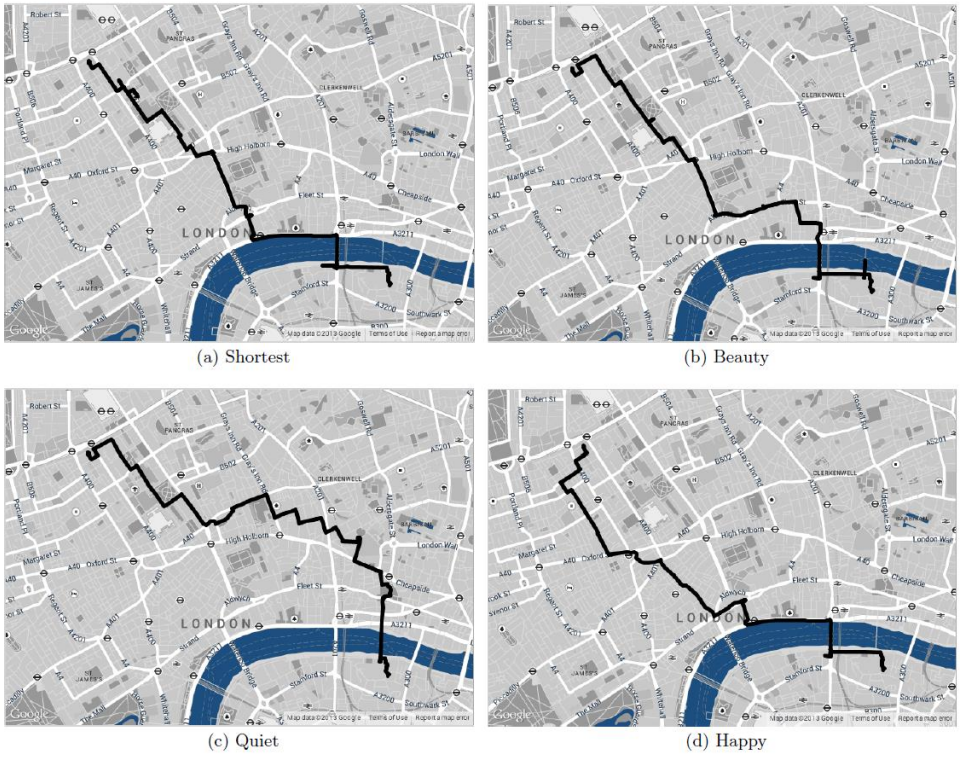

Figure 1- Alternative routes corresponding to different criteria between Euston Square and Tate Modern (Quercia et al. 2014)

Runge et al. tried to investigate scenicness for driving using Google Street View (GSV) images (Runge et al. 2016). They tried to use GSV images to take actual view on specific route segments rather than approximate pictures taken from volunteered geographic information (VGI). Then they used a pre-trained convolutional neural network (Places-CNN) created by (Zhou et al. 2014) to

categorize images into different tags (It will be discussed with further detail in this section). They grouped tags generated from the CNN and regrouped them in 6 categories which they assumed can be considered scenic. The main limitation of this research is that they did not provide a reason for selection of these categories are selected.

To address scenery in a systematic way, thanks to recent advancements in deep learning and data science, researchers were able to develop a scene classification model that significantly outperforms previous approaches (Zhou et al. 2017). They have used a repository of 10 million scene photographs, labeled with scene semantic categories, and applied a Convolutional Neural Networks (CNN). This scene classification model is called CNN Places365. It consists of 434 categories.

In a recent study (Seresinhe, Preis, and Moat 2017), researchers have investigated over 200,000 images through crowdsourcing from the existing online game Scenic-Or-Not. It was combined with the ability to extract hundreds of features from the images using the CNN Places365. The process of this study is that they have asked users to score images according to what they find scenic. This is called scenicness. Then they have acquired the tags using CNN Places365, and finally, they have modeled the effect of each tag on scenicness ratings using two modeling frameworks. The first is an elastic net model, whose sample of coefficients is depicted in Figure 2. The second model in a convolutional neural network.

Elastic net is a linear regression model trained with L1 and L2 prior as regularizer. Regularizations are terms added to the loss function of a problem to solve an ill-posed problem or to prevent overfitting (Bühlmann and Van De Geer 2011). The L1 regularization technique also called Lasso Regression adds a first order norm of the coefficients as a penalty term to the loss function. L2, also called as Ridge Regression, adds the second order norm. Elastic net uses both terms in its loss function. This combination allows for learning a sparse model where few of the weights are non-zero like Lasso, while still maintaining the regularization properties of Ridge. Elastic net is useful when there are multiple features that are correlated with one another (Bühlmann and Van De Geer 2011).

The second model is based on a convolutional neural network, applied via four different frameworks - AlexNet, VGG16, GoogleNet and ResNet152 (Seresinhe et al. 2017). The accuracy

rate of the predictions ranged between 0.445 to 0.654, which shows great improvement in comparison to previous work by Quercia et al. They have found that, for instance, as expected, natural features, such as 'Coast' and 'Mountain', are indeed associated with greater scenicness. Nevertheless, in urban built-up areas, the definition of scenicness is different. For example, man-made features can also be rated as scenic; such as 'Cottage' and 'Castle', as well as bridge-like structures, such as 'Viaduct' and 'Aqueduct'. Additionally, man-made features such as 'Construction Site' and parking Lots' are associated with lower scenicness in general as well as in urban built-up settings specifically. Figure 2 provides the elastic net coefficients, corresponding to scenicness proposed by Seresinhe et al.

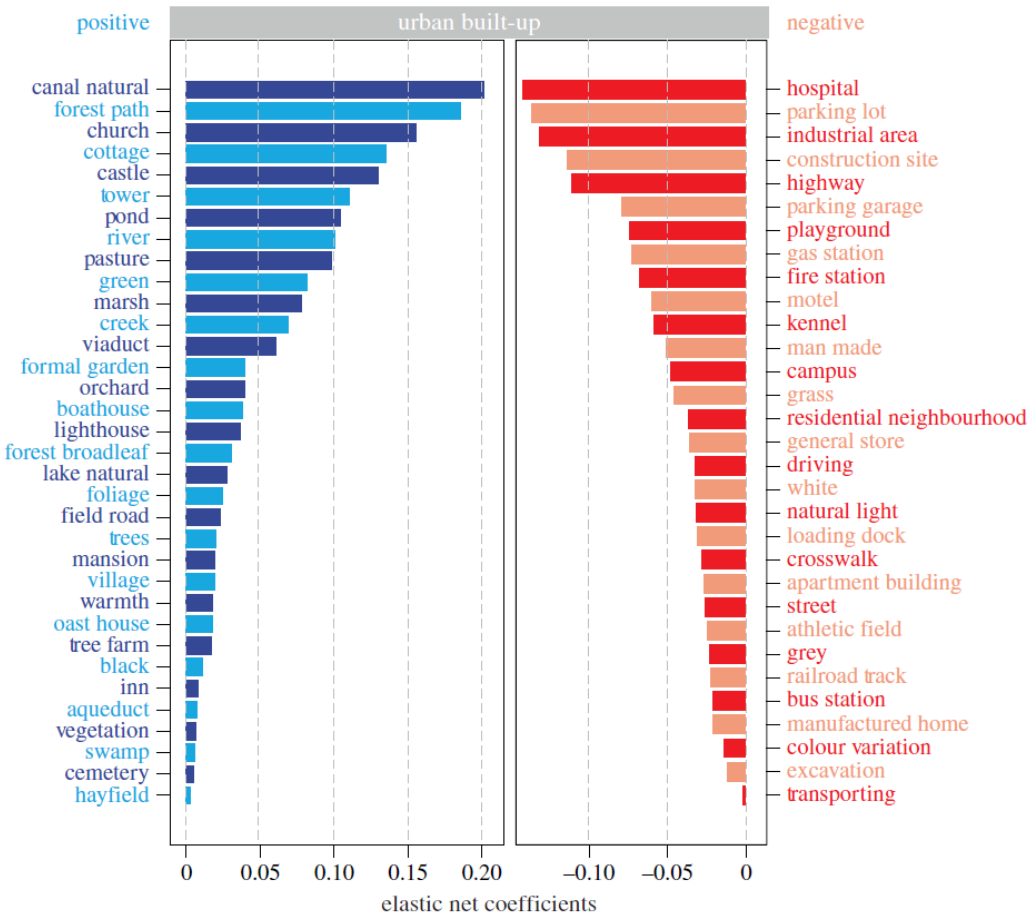

Figure 2 - Elastic net coefficients, corresponding to scenicness (Seresinhe et al. 2017)

2.5 Machine learning models

The classic definition is that machine learning is a field of study stemming from computer science that gives computers the ability to learn without being explicitly programmed (Samuel 1959). The learning process can be defined as acquiring new or modifying existing knowledge, behaviors, skills, values, or preferences (Gross 2015). In more practical terms, machine learning researchers study and create algorithms that can learn from patterns or make predictions with data. This field of science is useful when rule-based algorithms are not capable of solving the problems.

Machine learning addresses two main categories of tasks, making predictions and learning a pattern. Machine learning algorithms that tend to learn patterns are generally unsupervised. On the other hand, making predictions are categorized in supervised learning algorithms, which are the focus of this literature review due to their application in this thesis.

Supervised learning is defined as learning a function that maps an input to an output based on example input-output pairs (Russell and Norvig 2016). The two main problems that supervised algorithms solve are regression and classification. Regression problems map a feature space of inputs to single or multiple continuous outputs. On the other hand, the main question that classification algorithms tend to answer is that whether a set of data belongs to a certain category, that is they map the input into a discrete output space. For example, the well-known problem of identifying spam emails is considered a classification problem.

The first step in supervised learning is the observation of a phenomenon or random process which gives rise to an annotated training data set. The next step is to model this phenomenon by attempting to make a causal link between observation inputs and their corresponding observed observation outputs. This causal link is called the hypothesis/prediction function, where it is mainly referred to as a decision function in classifying tasks. What classifiers tend to do is maximize the conditional probability density function that governs the input-output space, which can then be used to define a suitable hypothesis.

The hypothesis is restricted to minimizing some measure of error over the observed training set while also maintaining a simple functional form. The first condition ensures that a causal link is in fact extracted from the observed data. The second condition avoids overfitting, that is producing a too closely or exactly to a particular set of data and may therefore fail to fit additional data or predict future observations reliably.

There are multiple algorithms that handle classification problems. In this thesis, the algorithms that were used are discussed in this section. That is, Decision Trees, Random Forest and Gradient Boosting Classifiers.

2.5.1 Decision Tree Learning

A decision tree is a mathematical representation of information, decisions by using graph theory. Its application is widespread in operations research and machine learning. It connects each decision with its possible consequence. Other information that can be represented include chance event outcomes, resource costs, and utility. *Decision tree learning* uses a decision tree to find the rules that lead observations about an item to conclusions about the item's target value. If the target value is a discrete set of values, the decision tree learning would be a classification tree, otherwise it is a regression tree. Classification trees are used to classify an object or an instance into a predefined set of classes based on their attribute values. Classification trees are frequently used in applied fields such as finance, marketing, engineering and medicine (Lior 2014).

A decision tree classifier is a classifier expressed as a recursive partition of the instance space. The decision tree consists of different nodes. The first node is called root that has no incoming edges. All other nodes have exactly one incoming edge. A node with outgoing edges is referred to as an “internal” node or a “test” node. All other nodes are called “leaves” (also known as “terminal” nodes or “decision” nodes). In a decision tree, each internal node splits the instance space into two or more sub-spaces according to a certain discrete function of the input attributes values. In most cases, each test considers a single attribute, such that the instance space is partitioned according to the attributes value. In the case of numeric attributes, the condition refers to a range (Lior 2014). Each leaf is assigned to one class representing the most appropriate target value. It is also possible that a probability vector is assigned to each leaf to indicate the probability of the target attribute having a certain value (Lior 2014). When the attributes are numeric, decision trees can be geometrically interpreted as a collection of hyperplanes, each orthogonal to one of the axes (Lior 2014).

There are multiple algorithms proposed to address decision tree learning, which include ID3, C4.5, and CART. These algorithms usually work top-down, that is they start by choosing a variable at each step that best splits the set of items and do this procedure recursively. To measure the ‘best’

split, various metrics have been introduced. These metrics are applied to each candidate subset, and the resulting values are aggregated, mainly averaged to provide a measure of the quality of the split. They include Gini impurity, information gain, variance reduction and gain ratio. To avoid splitting the instance space in a non-optimal way, these algorithms use a technique called pruning, which reduces the size of decision trees by removing sections of the tree that provide little power to classify instances. Pruning helps the decision tree classifier to avoid overfitting. In the following paragraphs, the advantages and disadvantages of these algorithms are introduced.

The ID3 algorithm is a very simple decision tree algorithm (Quinlan 1986). Using information gain as a splitting criterion, the ID3 algorithm ceases to grow when all instances belong to a single value of a target feature or when best information gain is not greater than zero (Lior 2014). ID3 does not apply any pruning procedure nor does it handle numeric attributes or missing values (Lior 2014). ID3 can be considered as the simplest decision tree algorithm. On the other hand, it does not guarantee an optimal solution. Since it does not have a pruning technique, it can overfit the training data. Furthermore, since it was designed for nominal attributes, continuous data needs to be converted to nominal bins.

C4.5, an evolution of ID3, uses gain ratio as splitting criteria. The splitting ceases when the number of instances to be split is below a certain threshold. C4.5 can handle numeric attributes. C4.5 uses a pruning procedure which removes branches that do not contribute to the accuracy and replace them with leaf nodes. C4.5 handles continuous attributes by splitting the attribute's value range into two subsets (binary split). Specifically, it searches for the best threshold that maximizes the gain ratio criterion. All values above the threshold constitute the first subset and all other values constitute the second subset (Lior 2014).

CART stands for Classification and Regression Trees (Breiman et al. 2005). Its structure is very similar to C4.5. An important advantage of CART over C4.5 is its ability to generate regression trees. Its distinguishable feature is that it constructs binary trees, namely each internal node has exactly two outgoing edges (Lior 2014).

2.5.2 Ensemble Methods

Ensemble methods use multiple learning algorithms to obtain better predictive performance that could be obtained from any of the constituent learning algorithms alone (Kowsari et al. 2018). If an ensemble is properly constructed; it can outperform single classifier-based approaches (Dietterich 2000). This is also shown in empirical studies (Brown et al. 2005; Freund and Schapire 1997).

There are multiple ways to construct an ensemble classifier, which include dividing a training set, manipulating data distribution, manipulating input features and manipulating learning algorithms.

Dividing a training set requires generating multiple data subsets from the base training dataset. For each of these subsets, multiple classifiers are constructed (Farrash 2016). Data subsets are generated by sampling and partitioning techniques. One of the most prevalent ensembles of classifiers that adopt dividing a training set is Bagging (Farrash 2016; Liang, Zhu, and Zhang 2011).

Breiman proposed bagging, which stands for bootstrap aggregating (Breiman 1996). It is an algorithm based on the idea of generating multiple subsets by repeatedly extracting samples with replacement (bootstrap) from the original dataset. Because of the boosting, that is sampling with replacement, any training instance may appear in a bootstrap more than once, while some training instances may not appear at all. It has been reported that on average 37% of training set instances do not appear in a bootstrap, particularly with large datasets (Skurichina and Duin 2002). After generating bootstraps, a base classifier model is built for each bootstrap by using a decision tree learning algorithm. The final ensemble decision is obtained by majority voting, that is selecting the best model with highest accuracy.

Manipulating data distribution is typically done by boosting. Freund and Schapiro introduced the prominent boosting algorithm called AdaBoost(Freund and Schapire 1997). In this algorithm, multiple classifiers are iteratively constructed from the entire dataset rather than a sample of the training data. At each step, the new base classifier improves classification on training instances that are incorrectly classified in the previous iteration (Farrash 2016). The final ensemble prediction is created from weighted voting, wherein each classifier's prediction is weighted according to its accuracy on the training datasets (Farrash 2016).

In manipulating input features, the training dataset is the same for all iterations of constructing multiple classifiers. The difference with other methods is that each classifier is built using different parts of a feature space, that is they sample from features rather than data points. Random forest is one of the most prevalent algorithms in this category, which was introduced by Breiman (Breiman 2001). Random forest uses decision tree as the base classifier and the word 'forest' refers to aggregation of many tree models that are constructed with pruning of fully grown trees (Zhang and Wang 2009). Each tree is constructed from all instances, which are drawn from the original training dataset by sampling with replacement from their features (attributes). In each tree node, a splitting attribute is selected from a randomly chosen sample of the training dataset's attributes (Farrash 2016). Ensemble prediction is made by majority voting. It is shown that the accuracy of random forest is most of the time greater or equal to AdaBoost (Breiman 2001). It also benefits from superiority in speed and robustness to outliers and noise (Breiman 2001). The structure of the random forest method is illustrated in Figure 3.

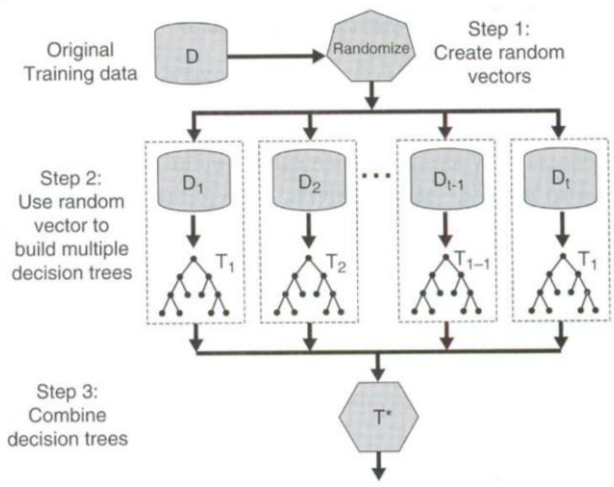

Figure 3- Random forest (Tan, Steinbach, and Kumar 2005, Page 279)

Manipulating learning algorithms is created by an ensemble through two possible approaches. The first approach is manipulating a base learning algorithm to create different models: changing the hyperparameters for the same classifier and combining them via an ensemble on the same training data set. The other approach is using multiple different learning algorithms that are each adopted to create a model from the same training dataset (Farrash 2016). These ensembles are called

heterogeneous ensembles of classifiers (Farrash 2016). The example of base classifiers for these models is artificial neural networks.

2.5.3 Gradient Boosting Classifiers

The Gradient Boosting (GB) classifier is a hybrid of the boosting and bagging approaches (Friedman 2001). In this algorithm, first, a random sample of the data is selected at each step of the boosting process. Second, boosting is based on a steepest gradient algorithm, with the gradient defined by deviance (twice the binomial negative log-likelihood) as a surrogate for misclassification rates. That is, in this step, the difference of the prediction and real value is generated (gradient) and then it is fed to the new classifier to be predicted. The gradient boosting is a general algorithm, that is it can use any base algorithm. The typical base algorithm is decision trees (Dietterich, Hao, and Ashenfelter 2008). Gradient tree boosting (GTB) consists mainly of fixed sized base learners, relatively small trees, with 6 terminal nodes being common size (Lawrence et al. 2004).

As with the other ensemble methods, larger trees are not formed, rather each tree developed during the process (often 100– 200 trees) is summed, and each observation is classified according to the most common classification among the trees. The combined effect of these differences from other boosting methods reduces GB sensitivity to inaccurate training data, outliers, and unbalanced data sets since, among other things, the steepest gradient algorithm places emphasis on misclassified training data that are close to their correct classification, rather than the worst classified data. GB has been shown in most cases to produce substantially higher accuracies with independent data (data that were not used to develop the trees) than either Classification tree analysis (CTA) of other boosting methods (Breiman et al. 2005). Finally, unlike CTA, which is highly prone to overfitting to training data, GB is highly resistant to overfitting since very small classification trees are used at each step of the boosting process.

With the gradient boosting algorithm, like other supervised classification models, the goal is to find the function f(x) using training sets where the misclassification error associated with the testing set will be as small as possible. To build f(x) in this setting, a probabilistic framework is applied. That is, first, a sigmoid function is applied to denote the probability of each point in a

class ([8]) , then the likelihood is calculated ([9]). Finally, values that correspond to the maximum of the log likelihood are chosen as the model parameters ([9]).

$$P(y = 1|x) = \frac{1}{1 + exp(-\sum_{m=1}^{M} h_m(x))} \quad \textbf{[8]}$$

$$L(y_i, f(x_i)) = \log(P(y_i|x_i)) \quad \textbf{[9]}$$

$$Q[f] = \sum_{i=1}^{n} L(y_i, f(x_i)) \quad \textbf{[10]}$$

The algorithm is introduced below (Friedman 2002):

Input: training set $Z = (x_1, y_1), \ldots, (x_n, y_n)$

M – number of iteration

1. $f_0(x) = log \frac{p_1}{1-p_1}$
2. For $m = 1 \ldots M$:
   2.1. $g_i = \frac{dL(y_i, f(x_i))}{df_m(x_i)}$ (gradient)
   2.2. Fit a decision tree $h_m(x)$ to the target g_i
   2.3. $\rho_m = \underset{\rho}{\operatorname{argmax}} Q[f_{m-1}(x) + \rho\, h_m(x)]$
   2.4. $f_m(x) = f_{m-1}(x) + \nu\rho_m\, h_m(x)$
3. Return: $f_m(x)$

Where:

each h_m is a decision tree

g_i is the gradient of likelihood

ν is regularization (learning rate) which is recommended to be less than 0.1

2.6 Machine learning models in transportation and choice modeling

Machine learning algorithms have become useful in transportation planning problems in recent years. Dougherty reviewed these algorithms including artificial neural networks, decision trees

and other classification algorithms (Dougherty 1995). For example, in a study by (Yamamoto, Kitamura, and Fujii 2002), decision trees and production rules algorithms were used to investigate driver route choice. They used two surveys to collect data on driver route choice between two alternative routes on expressway networks, which sets the route choice problem to be consistent with a binary classification framework.

In a recent study, researchers investigated using artificial neural network (NN) and Support Vector Machine (SVM) classifiers in route choice (Sun and Park 2017). They used a stated preference survey with 18 participants. With respect to attributes of three route alternatives including travel time, travel time fluctuations and fuel cost. The results show that the SVM has similar prediction accuracy as the NN, but it has a significantly higher computation efficiency (Sun and Park 2017).

The other choice problem that is addressed by machine learning techniques is mode choice. For example, there have been studies modeling mode choice as a pattern recognition problem in which multiple human behavior patterns reflected from explanatory variables determine the choices between alternatives or classes (Xie, Lu, and Parkany 2003). In a study, the capability and performance on work travel mode choice decision trees and neural networks is compared to multinomial logit model (MNL). The researcher used diary datasets from the San Francisco Bay Area Travel Survey (BATS) 2000 for model estimation and evaluation (Xie et al. 2003). The prediction results showed that the two data mining models offer comparable but slightly better performance than the MNL model in terms of the modeling results, while the decision tree model yielded highest estimation efficiency and most explicit interpretability and the neural network model gave a superior prediction performance in most cases (Xie et al. 2003). For their specific problem, the NN model (88.0%) shows a best overall performance over the other two models (86.0% and 86.7% for the DT and MNL model)

In another recent study, both Machine Learning and Discrete Choice modeling frameworks were used to predict the car ownership using transportation household survey data from Singapore (Paredes et al. 2017). The researchers compared a multinomial logit model against various machine learning models (e.g. Random Forest, Support Vector Machines) by using two datasets, one of them 2008 data to estimate models and 2012 ownership to predict the accuracy using the models already derived (Paredes et al. 2017). Their study found that machine learning models are inferior to the discrete choice model when using discrete choice features. However, after data engineering,

these models showed better performance in addressing choice. The prediction accuracy of machine learning models before data processing ranged from 0.642 to 0.742 and the multinomial logit model performance was 0.743. After feature engineering, the machine learning models ranged from 0.749 to 0.799 in accuracy. The feature engineering set dummy variables for discrete choice modeling dataset, while they have also incorporated other variables which cannot be discretized in the machine learning dataset (Paredes et al. 2017).

Using a Dutch travel diary data from the years 2010 to 2012 with variables on the built and natural environment as well as on weather conditions, researchers investigated multinomial logit models and compared the predictive performance of seven selected machine learning classifiers for travel mode choice analysis (Hagenauer and Helbich 2017). The results showed that machine learning models, specifically random forest, showed slightly better performance. Additionally, they investigated the importance of each factor using both Machine learning models and multinomial logit models. The results suggested that the analysis of variable importance with respect to the different classifiers and travel modes can be helpful for improved model analysis (Hagenauer and Helbich 2017).

2.7 Conclusion of literature review and contributions

In this thesis, the objective was to use smart-phone based raw GPS data to analyze revealed pedestrian trajectories to model route choice. As identified in the Literature Review, pedestrian route choice models use various input variables (factors), some of which are shown in Table 1.

The main contributions of this research are in two levels, which are introducing new variables quantitatively that may be explanatory and using machine learning algorithms that were not tested in route choice and compare their performance with discrete choice models.

To reduce the cost of acquiring perceived built environment measures, which is costly due to large surveys and low levels of participation, it is valuable to have a model to estimate scenery rather than doing the surveys. Additionally, understanding whether scenery would affect pedestrian route choice in a revealed preference setting rather than stated preference can be considered a contribution, because it has been done so rarely. Furthermore, the extent of different trip purposes on route choice when accounting for scenery is investigated.

The other novel approach in this study is the use of micro-level scale land use tags as a route attribute for each alternative. As discussed in the Literature Review, there are inconsistent finding regarding the interaction of land use and extent of its effect on route choice (Badoe and Miller 2000). This study investigated over 110 different places type in modeling route choice to seek whether each one influenced pedestrian route choice behavior or not.

There are multiple features which may affect route choice, however they may not be found significant due to assumptions of discrete choice models, especially, linearity of feature utility function. To address this matter in this study, with help of supervised learning algorithms, all the variables are used to build SVM, Random Forest and Gradient tree bosting models and their accuracy is compared with the traditional discrete choice framework.

3 Data Collection and Processing

To capture pedestrian behavior in a cohesive way, multiple data sets have been used. Because of the variety of data sources, each data set is described, and the processing techniques is presented in detail. The data consist of GPS data, image data and text labels. Figure 4 shows the data processing process in chronological order. The data used in this thesis is composed of the following five datasets:

1. MTL Trajet Database
2. Google Maps API data (Mainly Google Directions API)
3. Google Street View Images dataset
4. Places365 CNNs image tags
5. The level of scenery as defined by the coefficients shown in (Seresinhe et al. 2017)

3.1 MTL Trajet Database

This research has been conducted based on data collected through the App MTL Trajet. MTL Trajet was an instance of the smartphone travel survey app, DataMobile (Patterson 2017; Patterson et al. 2018; Patterson and Fitzsimmons 2016). The data is acquired by TRIP lab, Concordia University, Montreal, Canada, through an ongoing project which is now referred to as the Itinerum™ platform. MTL Trajet was released as part of a large-scale pilot study in a study that lasted 30 days. It is one of 70 projects in the 2015-2017 Montréal, Smart and Digital City Action Plan. The original purpose of the application was to collect travel behaviour information. The mobile application records the location of the respondent smartphones as they travel. The application captures movement, and if the user does not move for more than two minutes, a prompt is sent to ask whether he or she has ended the trip, the trip mode, and trip purpose. These prompts are used to validate which transportation mode has been chosen. Additionally, the socio-demographic information was asked in the initiation of the survey. These user descriptions include age bracket, sex, occupation, licence or transit pass, number of people within household, number of cars within household, latitude and longitude of participant's home, work and study location; primary and secondary travel mode(s) to work and/or to school.

The MTL Trajet dataset contains over 33 million location (primarily GPS) points. To detect trips and segments we used the rule-based trip-breaking algorithm developed in (Patterson and Fitzsimmons 2016). The algorithm detects segments based on 3-min gap in data while controlling for velocity and parameters relating to the public transit network (i.e. transit junctions and metro station location). Applying the trip-breaking algorithm on MTL Trajet dataset resulted in 623,718 trips, among which 102,904 trips were validated by respondents (Yazdizadeh, Patterson, and Farooq 2018).

Validated mode data was derived from the survey questions presented to respondents upon 35 installations. Respondents were asked the location of home, work and school, as well as the mode(s) of transport used for trips to these locations. Only validated trips from users who had declared they used only one mode option to travel between home and work or home and school were used. This procedure provided us with 10,518 validated trips (Yazdizadeh et al. 2018). With respect to trip purpose detection, six activity categories were used to predict trip purpose: "education," 40 "health," "leisure," "shopping/errands," "return home" and 'work." This dataset contained 4,996,501 rows (coordinates), collected from 2,414 distinct users, resulting to 10,800 trips. The Walking trips were only 1531.

After considering three filters, that is the speed should be consistent and less than 3 m/s three trip alternative and trip length more than 250 meters, 240 trips were selected. This data set can be considered a reasonable representative of the pedestrian route choice. This data spanned from 2016-10-17 to 2016-11-21. As can be seen in the Table 2, there is a good spread of people over age, however, there is no pedestrian recorded who is over 65. This can be associated with low penetration level of cell phones among elder people. As pedestrian have stated through multiple choice questions in the app, there has been a majority of educational and 'other' trips despite 'other' tag being not informative. Table 3 shows the travel mode preference by trip type. As can be seen, there are reasonable amount of people (212 out of 240), who prefer to walk to their study destination as primary mode choice. Figure 5 depicts the frequency of trips recorded per unique user.

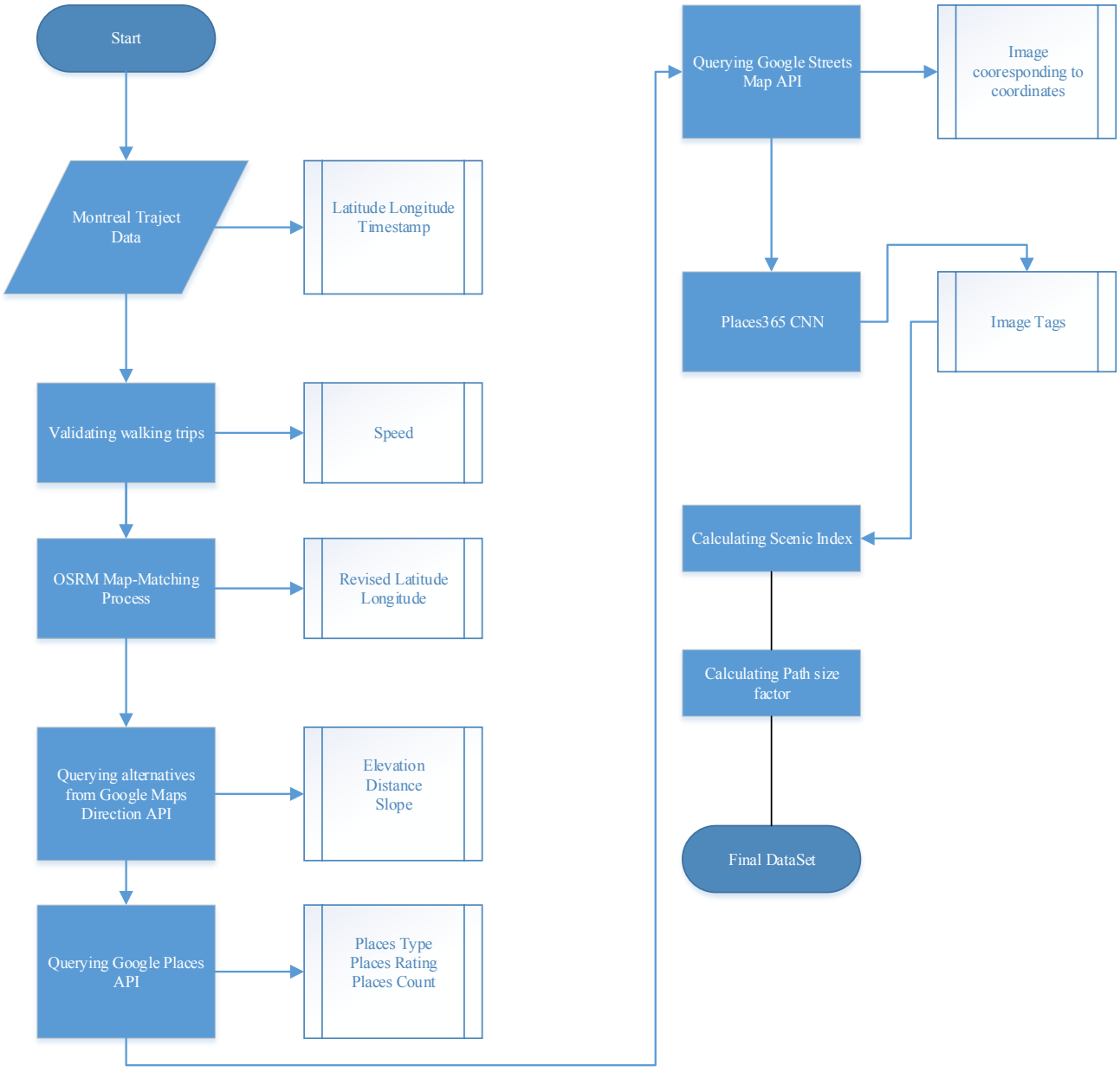

Figure 4- Data processing procedures and their outcome at each level

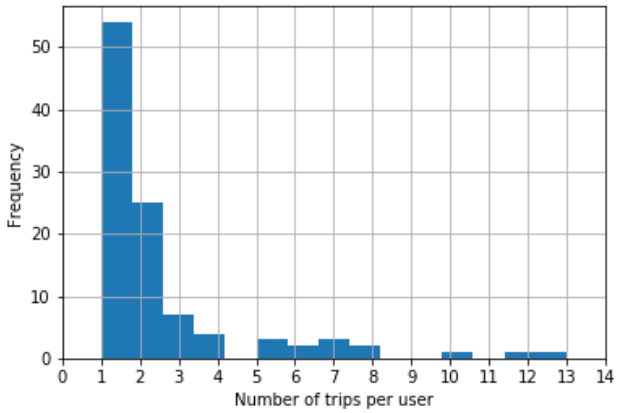

Figure 5-Frequency of trips recorded per unique user

Table 2-Sociodemographic characteristics and stated purpose of the pedestrian

Age Interval	**Frequency**	**Purpose**	**Frequency**
16-24	47	*Education*	87
25-34	74	*Health*	6
35-44	53	*Meal/snack/coffee*	13
45-54	53	*Leisure*	21
55-64	13	*Pickup*	3
65 or more	0	*Other*	78
Sum	**240**	***Sum***	**240**

Table 3 - Travel mode preference by trip type

Mode	**Main travel mode to work**	**Alternative travel mode to work**	**Main travel mode to study**	**Alternative travel mode to study**
On foot	140	64	212	4
Bicycle	64	14	9	13
Transit	27	1	16	5
Car	2	4	2	0

3.2 Map Matching Process

The location data is obtained by sampling the positions typically using GPS to produce data that in database terms is commonly referred to as trajectories. Unfortunately, this data is not precise due to the measurement error caused by the limited GPS accuracy, and the sampling error caused by the sampling rate (Brakatsoulas et al. 2005). A pre-processing step that matches the trajectories to the road network is needed. This technique is commonly referred to as map matching. Map-matching algorithms integrate positioning data with spatial road network data (roadway centerlines) to identify the correct link on which an observed point is located (Quddus, Ochieng, and Noland 2007). Researchers have reviewed multiple map-matching algorithms and their corresponding navigation sensors, test environment and accuracy (Quddus et al. 2007). There are available software packages that performs map matching as Open Source Routing Machine (OSRM). The Open Source Routing Machine or OSRM is a C++ implementation of a high-performance routing engine for shortest paths in road networks, it also supports map matching. Literature indicates its usage by as a map matching tool in academic research context (Yang and Meng 2015). Due to its reasonable results and ease of use, it was used to perform map matching on the GPS points. A sample of its performance is depicted in Figure 6.

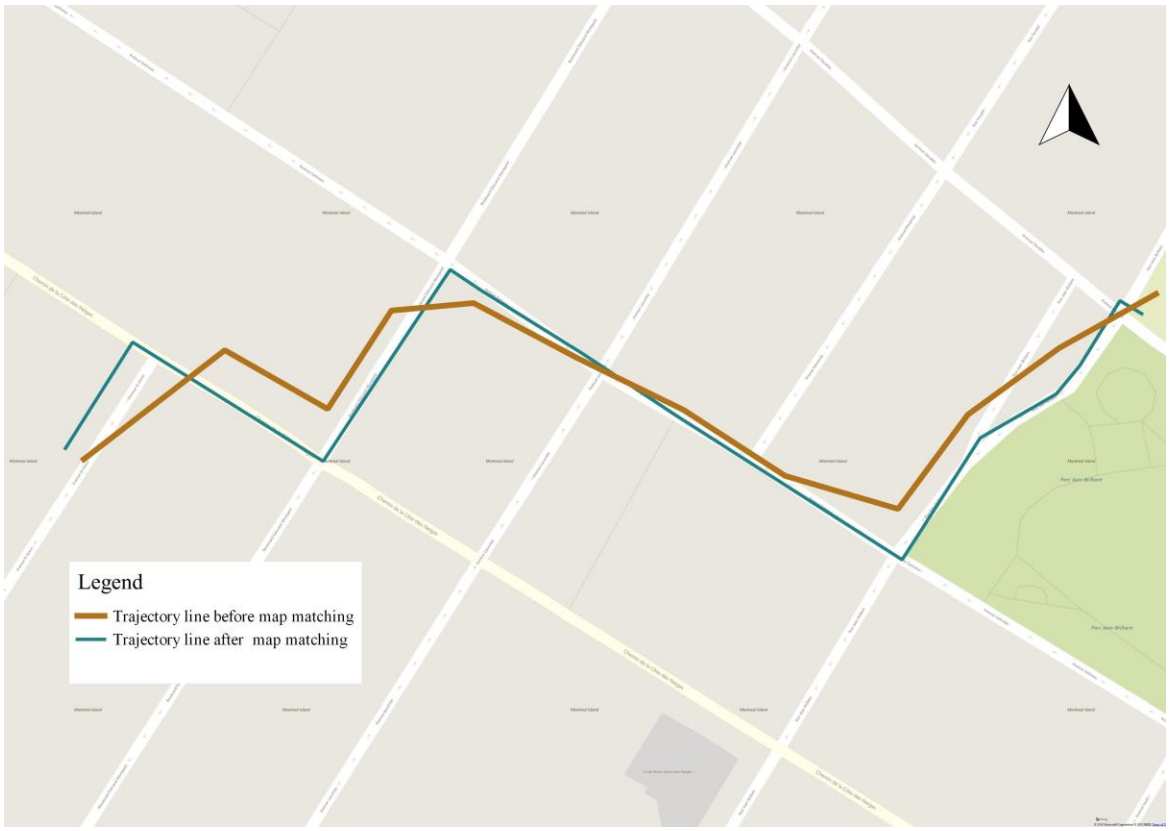

Figure 6-Map-matching comparison for a sample trip

3.3 Google Maps Directions API

The Google Maps Directions API is a service that calculates possible routes between given locations. It is possible to search for routes using different modes of transportation, including transit, driving, walking or cycling. There is evidence in the literature that shows usage of this API as a choice set generator in transportation context (Eluru, Chakour, and El-Geneidy 2012). In this thesis, this API was used to identify the alternative paths for a walking trip. The results were obtained in Encoded Polyline format, which were decoded and transformed to Spatial Reference System Identifier (SRID) of Montreal (4326), via PostGIS functions. Figure 7 shows all lines derived from Google Maps Direction API.

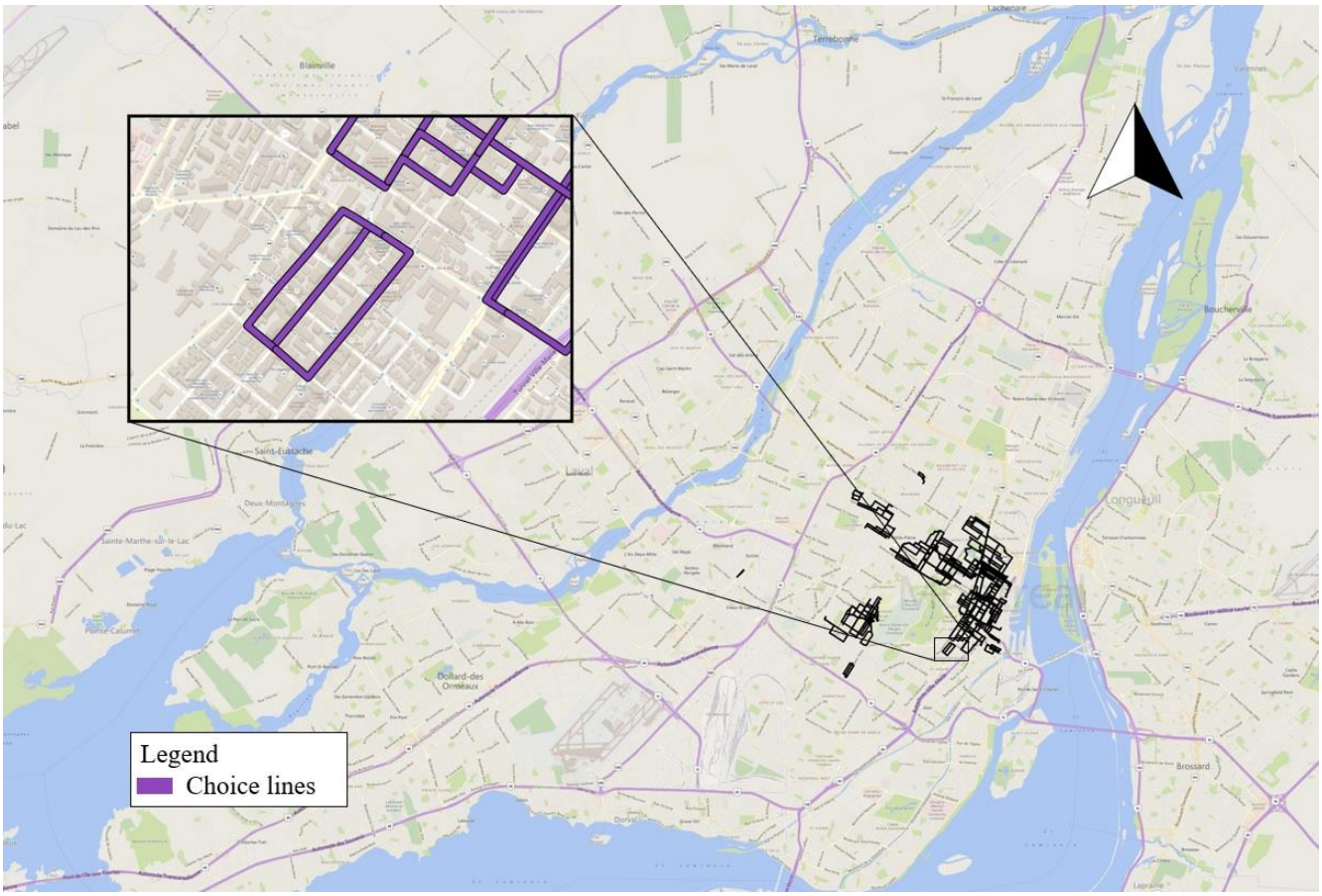

Figure 7 - Choice Set Lines

The queries of Google Maps Direction API resulted in two or three alternatives for each trip. Those trips with three alternatives were considered in the final set. For each alternative, distance, number of turns and elevation along sample points were taken. This led to calculating slope on GPS point interval. The description of the variables according to alternatives are presented in Table 4. As can be seen in this table, the traveled distances range from 284 meters to over 2505 meters, which shows a wide variety of pedestrian trips. Additionally, the average slope percentage is almost zero, which means that this feature cannot be accounted for in a route choice model. The alternatives that have the highest overlap percentage with the revealed path of each choice set were considered the chosen alternative.

Table 4 - Description of attributes acquired by Google Directions API

Attribute	Mean	Standard Deviation	Minimum	25%	50%	75%	Maximum
Distance (meters)	1243.14	468.09	284.0	905.75	1155.00	1586.75	2510
Number of turns	3.99	2.129	0	2	4	5	11
Average Slope Percentage	0.156	2.85	-15.63	-0.930	0.037	1.10	23.548

3.4 Google Street View

The Google Street View Image API module was used to collect the images along the routes of the analyzed trips. Through this module it is possible to query the street view by coordinates, with optional parameters of heading and field of view (*fov*). The heading indicates the compass heading of the camera and fov (*default is* 90) determines the horizontal field of view of the image. In this study, 12 images were queried for each coordinate. This number was chosen to change heading twelve times with thirty-degree intervals. This was chosen to ensure validity of tags obtained by the CNN Places365 model, which will be further discussed in the following subsection. A sample of twelve images acquired by for a coordinate is portrayed in Figure 8.

Figure 8-Sample of images acquired for a coordinate by Google Street View API

3.5 CNN Places365 Description

To calculate scenicness, the images were first transformed to their corresponding tags and later, the tags were used to calculate scenic index for each coordinate using scenery. The CNNPlaces365 model introduced by Zhou et al. was used (Zhou et al. 2017). Images corresponding to coordinates obtained by Google Street View were the inputs of the model and the outputs were the tags with their probability. An illustration of output of the model for a single image is depicted in Figure 11. As seen in this illustration, there is a distinct probability assigned to each scene category. According to Zhou et al., this model has an accuracy of 87 percent (Zhou et al. 2017). That is, at least one of the scene categories in top five predictions associated with an image is truly the tag that a human can identify, so it can be used as a rather reliable source of data. coordinates show frequency of unique scene tag over 50 for all coordinates of the study. As seen in this figure, the residential neighborhood and street tags are the two most frequent. This was expected because images from Google Street View were mainly street image in residential neighborhoods. However, there are irrelevant tags associated with each picture as well, which can be eliminated by probability ratios. In this thesis, all the probabilities for twelve images with the same tag has been summed for each of the coordinates and the top five has been selected as the meaningful tags. As shown in Figure 10, the probabilities will change the values and order of tags. For example, the second most frequent tag in Figure 10 is *parking_garage/outdoor*, instead of *street*. This procedure ensured the validity of the tags.

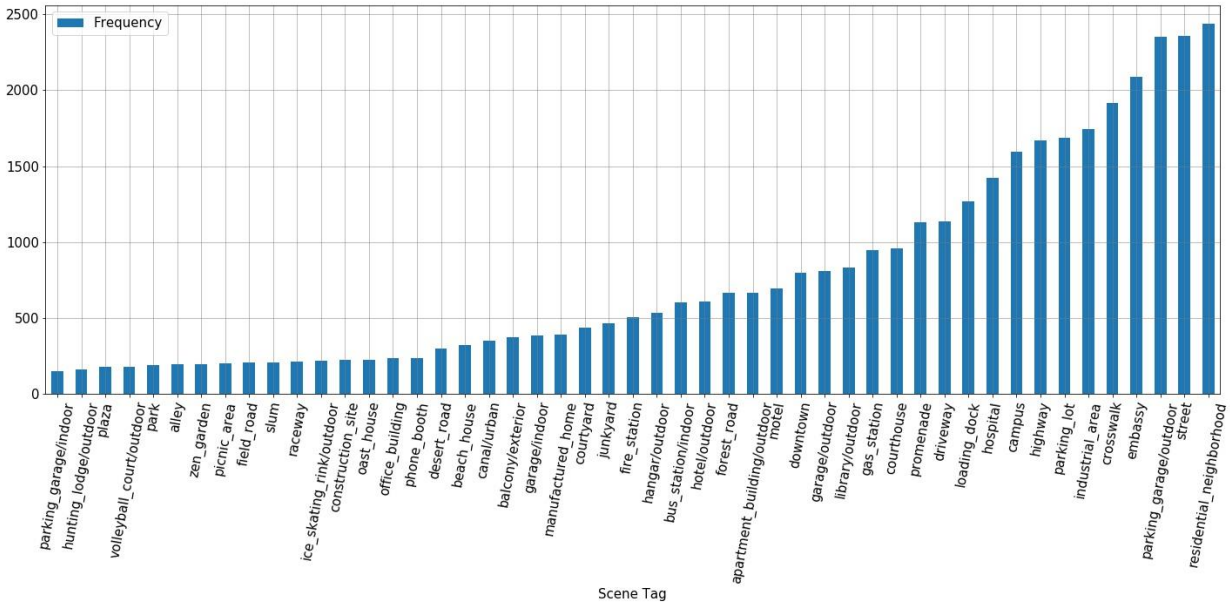

Figure 9 -Frequency of unique scene tag for all coordinates

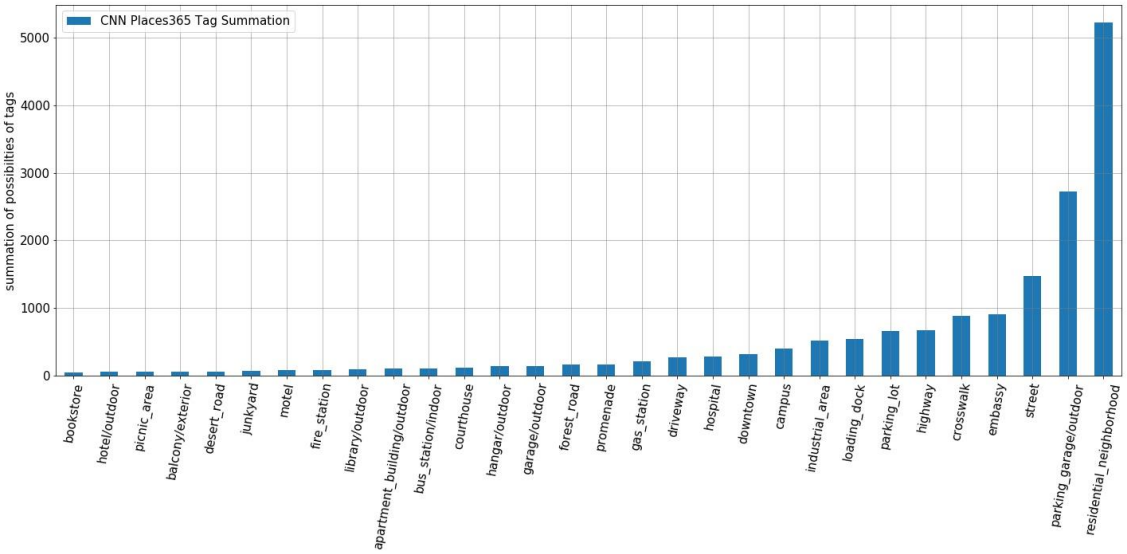

Figure 10 - CNN Places probability of summation

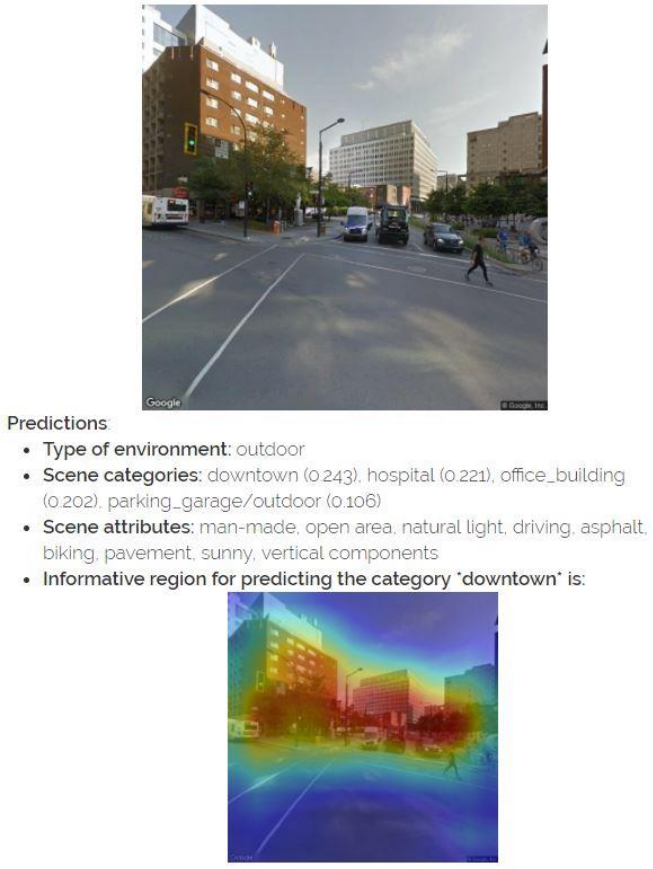

Figure 11 - Demo of CNNPlaces 365 Predictions, retrieved 4/17/2018

3.6 Scenic index calculation

The final step to calculate scenicness for each coordinate was to employ the Elastic Net model proposed by Seresinhe et al. The image tags for each coordinate were the inputs of the model and the outputs were the scenic score. However, since there were twelve pictures for each coordinate, it was necessary to filter weight the tags that were the most relevant in each coordinate. To this end, the summation of probabilities (SP) of each tag for a coordinate were calculated. The top five SP values were kept for further calculation. These SP values were used as weights to scenicness coefficients reported by Seresinhe et al. [11] illustrates this formulation. The scenic coefficients for each tag is shown in Figure 2. The results of this calculation were then averaged for all coordinates of each alternative, and the results are illustrated in Table 5.

$$SIC = \sum_{i \in SC} ENC_i * sp_i \quad [11]$$

Where:

SIC: scenic index of a coordinate

SC: set of top five tags associated with a coordinate

ENC: elastic net coefficient of tag i

sp: summation of probability of tag i for twelve images

Table 5- First order statistics of the scenic index

	Mean	**Standard Deviation**	**Minimum**	**25%**	**50%**	**75%**	**Max.**
Average Scenicness Index	-0.134	0.0636	-0.390	-0.169	-0.127	-0.091	0.114
Variance of scenic index	0.014	0.024	0.000	0.002	0.006	0.016	0.196

3.7 Google Places API

Google Places API is part of Google Maps service, which provides access to information about more than 100 million places around the World. These places are, usually, public places like touristic attractions, hospitals, and stores, malls, companies etc. There are over a hundred distinct places types that was tagged to locations. This API is used for academic purposes as well (e.g. See (Ermagun et al. 2017)).

In this thesis, this service was used to investigate effects of land use on pedestrian route choice. It was employed to see whether each place type would show any effect or not. For each alternative, sample points were extracted from alternative lines. For each sample point, a radius of 25 meters was queried by this API. That is, the place tags in proximity of 25 meters of each point were collected. These tags corresponded to macro and micro level land uses. Macro level land uses corresponded to a set of micro level land uses. In this thesis, food, residential, health, financial are considered macro level tags. For example, the food tag is a macro tag, which consists of restaurant and café and food related micro tags. The macro tags are health, food and finance. Figure 12 shows

a wide variety of the place tags that are associated to all the path alternatives. The top three are food, restaurant and health tags. In this thesis, the question of whether all the land uses have the same effect on pedestrian route choice is investigated by considering each of these place tags. As shown in this table, most tags can be considered commercial land uses, however; other types of place tags, such as public transportation infrastructure like transit_station also had high frequency. The tags of the corresponding coordinates of each alternative were used for further analysis.

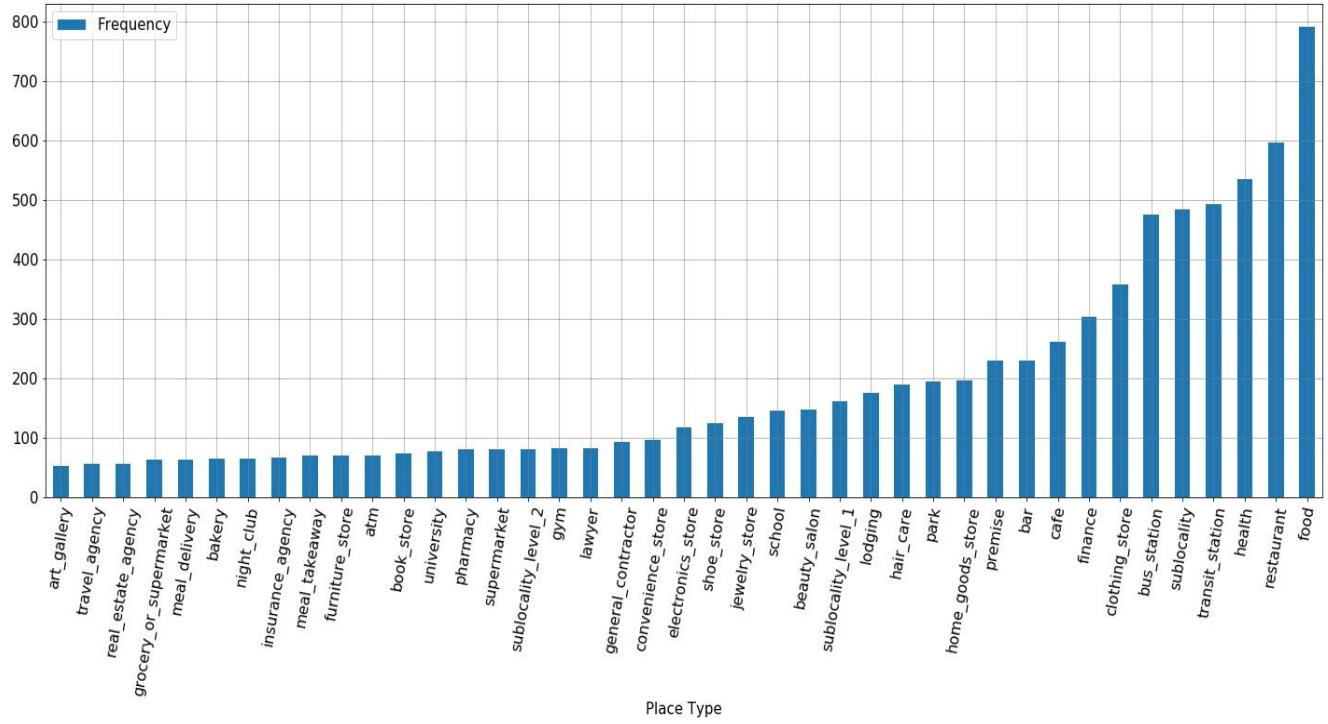

Figure 12-Frequency of places types over 50 of all alternatives

3.8 Path Size Factor

To calculate the path-size factor, the alternatives were split into their constructing points. Then, these points were used to create all the links related to alternatives. This procedure was scripted with help of PostGIS and Python. Finally, using the formulation presented in Table 7, the path size factor for each of the alternatives was calculated. denotes path size (PS) factor formulation characteristics of this study. The formulations of the PS factors guarantee a maximum value of zero. When ln(PS) has a value of zero, it means that there is no overlap within the choice set. As it is shown in Table 6- Path size factor formulation characteristics , the variables are close in the values of mean and standard distribution and the maximum is always negative, which means that all paths have at least one

overlapped linked. This makes usage of PS factor valuable. It is noticeable that for ln(PS3), φ was set to 14 based on finding of value proposed by Hoogendoorn et al. (Hoogendoorn-Lanser 2005).

Table 6- Path size factor formulation characteristics

Variable	Mean	Standard Deviation	min	25%	50%	75%	Max
ln(PS1)	-0.506	0.219	-1.098	-0.658	-0.490	-0.338	-0.044
ln(PS2)	-0.455	0.245	-1.095	-0.621	-0.453	-0.281	0.087
ln(PS3)	-0.498	0.336	-4.021	-0.694	-0.489	-0.278	0.466
PS4	-0.072	0.077	-0.484	-0.099	-0.044	-0.018	0

Table 7 - Path size factor formulation summary

Variable Name	Path size Equation	Study
ln(PS1)	$PS_{in} = \sum_{a \in \Gamma_i} \frac{L_a}{L_i} \frac{1}{\sum_{j \in C_n} \delta_{aj}}$	(Ben-Akiva and Bierlaire 1999)
ln(PS2)	$PS_{in} = \sum_{a \in \Gamma_i} \frac{L_a}{L_i} \frac{1}{\sum_{j \in C_n} \frac{L^*_{C_n}}{L_j} \delta_{aj}}$	(Ben-Akiva and Bierlaire 1999)
ln(PS3)	$PS_{in} = \sum_{a \in \Gamma_i} \frac{L_a}{L_i} \frac{1}{\sum_{j \in C_n} (\frac{L_i}{L_j})^{\varphi} \delta_{aj}}$	(Ramming 2002)
PS4	$PS_{in} = \sum_{a \in \Gamma_i} \frac{L_a}{L_i} ln(\frac{1}{\sum_{j \in C_n} \delta_{aj}})$	(Bovy et al. 2008)

3.9 Other data sources

Additionally, the number of signalized intersections for each alternative was calculated. This was made possible by using the Traffic lights - all intersections data set of the Montreal data. This dataset contains the location of all the traffic lights managed by the City of Montreal. It consists of the reference number of the intersection where the light is located, the names of the two streets that form the intersection, and the geographic coordinates of the center point of the intersection.

This data was imported in imported in QGIS as a layer and number of intersections with each alternative was counted. This number was reported as number of signalized intersection for each alternative.

All collected features are shown in Table 14

Appendix

Table 14-variables of supervised learning models

. This collected data set consists of geometrical features of alternatives such as distance and number of turns, built environment factors such as scenery and place tags and socio-demographic features of the trip makers. The details of the modeling procedure are described in Methodology section.

4 Methodology

In this thesis, the two main frameworks are discrete choice modeling and classification models. In this chapter, the data processing for each of the models, the frameworks main components and determination of hyperparameters of supervised learning algorithms are presented.

The data was divided in test and training sets. The training data is 85 percent of the data. This data composition is used for both modeling frameworks; however, their formats differ. Three model groups were built in this study. The first is discrete choice model (Model 1). Then, variables that were found important in this setting were used to train supervised classifiers (Model 2). Finally, Model 3 was built with all the variables using supervised classification. In Model 3, all micro level land uses were used. Table 8 illustrates the data and the variables for all the models.

Table 8 - Overall introduction to Models and their features

Model	Features
Discrete choice model (Model 1)	Limited
Supervised learning (Model 2)	Limited
Supervised Learning (Model 3)	All

For all three models, the prediction accuracy was used as measure of performance. Accuracy ratio is number of correct predictions to total number of predictions. To ensure validity of the models, the accuracy ratio was reported for the test set (see Comparison of models section). In the following sections, the details of parameters of each model is described.

4.1 Discrete choice model

As discussed in the Literature Review, the path-size logit model is an accepted method for analyzing route choice behavior, and therefore, is used in this study as well. The models were estimated using Python Biogeme 2.6.a. It is designed to estimate discrete choice models (Bierlaire 2016). This software package is commonly used in route choice modeling (Grond 2016; Lue and Miller 2018).

To create a discrete choice model, the evaluated variables included:

- Geometrical factors: distance, number of signalized intersection, number of turns and slope
- Scenery variables (maximum, minimum, average and variance scenic index)
- Macro level land uses: health, financial, food, residential land uses, points of interest
- User specific variables: age, gender, primary/secondary preferred mode and trip purposes
- Path size factors

The values that did not show significance were: number of signalized intersection and slope, all though were right-sided. The micro level features were not investigated due to lack of interpretability. For example, considering that people have chosen a route just because it has more bars may be irrelevant, although a model may indicate as such. Additionally, the interaction of user specific variables with other variables were investigated, yet a meaningful relationship which also had right sidedness were not found. The variables that were right sided but suffered from statistical insignificance were number of signalized intersection and slope. The only path size factor that showed importance was ln(PS3). The utility function of the best path size logit model is presented in equation [12].

$$U_i = \beta_D * D_i + \beta_{NOT} * NOT_i + \beta_{VSI} * VSI + \beta_{FLU} * FLU + \beta_{PS} * \ln(PS3)_i \quad \textbf{[12]}$$

Where:

D: distance

NOT: number of turns

VSI: variance scenic index

FLU: Food land uses

ln(PS3): path size factor introduced by Ramming.

To describe the variables in Table 9, the coefficients and their interpretation are discussed in results section. Additionally, introduction of Rho-square and Rho-square-bar is necessary. In discrete choice model, the log likelihood at equal shares (null log likelihood) is the same in all the estimated models. It describes the value of the log-likelihood function when all parameters are zero, i.e., when the alternatives are assumed to have equal probability to be chosen. It is computed as in [13].

Table 9 - Discrete choice model (Model 1) results

Variable	Coefficient	t-test
Variance scenic index	15.8	3.61
Length (per 100 m)	-0.637	-3.47
Food land-use	0.448	3.49
Number of turns	-0.242	-3.69
ln(PS)	1.26	3.18
$\overline{\rho}^2$ of model	0.231	
$\overline{\rho}^2$ of model	0.212	

$$\text{Log Likelihood (equal shares)} = \text{Number of observations} * \ln(0.5) \quad \textbf{[13]}$$

In logistic regression analysis, there is no agreed upon analogous measure, but there are several competing measures each with limitations to describe overall goodness of fit of the model. The McFadden rho (ρ^2) parameter is one of them and has a value between 0 and 1. Rho-squared is computed as equation [14].

$$\overline{\rho}^2 = 1 - \frac{\text{Log Likelihood (estimated model)}}{\text{Log Likelihood (equal shares)}} \quad \textbf{[14]}$$

The adjusted ρ^2, or $\overline{\rho}^2$, parameter considers the number of parameters included in a model and is computed as equation [15].

$$\overline{\rho}^2 = 1 - \frac{\text{Log Liklihood (estimated model)}}{\text{Number of parameters Log Likelihood (equal shares)}} \quad \textbf{[15]}$$

This value is reported by Python Biogeme, however, because this metric is not applicable to supervised classification techniques, the prediction accuracy is used in this thesis. To find the prediction accuracy, the models were simulated for the test set. To elaborate on this, the discrete choice model was created with the training set, and the coefficients were obtained. Using these coefficients and the test set, the simulations were made by Python Biogeme. The output was a file containing probability for each alternative and the choice. The alternative with the highest probability is the chosen predicted alternative. If it is the same as the revealed alternative, then it is considered as true prediction for a trip. This prediction accuracy is then converted to alternative based prediction accuracy to be comparable with the supervised learning models.

4.2 Supervised learning models

In this setting, the data format was changed from wide to long. That is each alternative was an instance. There were two sets of features to build machine learning models. The first set was the variables that showed significance in discreet choice model (Model 2). The second set consists of all variables that were available (Model 3). The details of the second set of features are demonstrated in Table 14

Appendix

Table 14-variables of supervised learning models

. These variables correspond to geographical aspects (number of intersection, distance, etc.), scenery, sociodemographic, purpose and land uses, both macro and micro levels. These variables are the inputs of the prediction (X matrix). The labels corresponding to each instance were a dummy variable, 0 or 1, which showed the instance was a chosen route or not (Y vector). As previously stated, the training and testing data were the same for all models. In this thesis, three supervised learning algorithms were employed, including Decision Tree (DT), Random Forest (RF) and Gradient Tree Boosting (GTB). The models were trained using Scikit-learn package in Python programming language, which is one of the most widely used packages in data science (Pedregosa et al. 2011).

For each classifier, there are multiple hyperparameters that can affect the performance of the model. Hyperparameters are parameters that are not directly learnt within estimators. Number of leaves in a decision tree is an example of hyper-parameters. To find the optimal value for these values, which would both increase the performance measure of the model and affect the generality of the model, using validation set seems reasonable. Figure 13 shows this data splitting visually. However, it would decrease the number of sample size which can be used for training the data. To address this problem, cross-validation technique is used. A simple algorithm for this general procedure is called k-fold cross validation. In k-fold, the training set is split into k smaller set, then a model is trained using $k - 1$ of the folds as training data, and the resulting model is validated on the remaining part of the data (i.e., it is used as a test set to compute a performance measure such as accuracy).

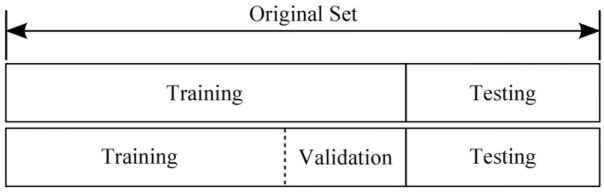

Figure 13 - Training, validation and test set

The performance measure reported by k-fold cross-validation is then the average of the performance measure of all folds of the data. This approach can be computationally expensive but

can be data efficient, which is a major advantage in problem such as inverse inference where the number of samples is very small.

4.2.1 Decision Tree

The decision tree, like other machine learning algorithms, consist of hyperparameters. To fit a decision tree, the hyper parameters, their description and the initial values are described in Table 11. To find the best parameters, the grid search technique was used. Grid search technique is exhaustive search over specified parameter values for an estimator. In another words, all the values were tested for each fold of validation set. The best parameters were selected based on the average accuracy over cross validation.

Table 10- Hyperparameters of Decision Tree

Hyper parameter	Description	Values	Best parameters Model 2	Best parameters Model 3
max_depth	The maximum depth of the tree.	Range of 10 to 500 with 20 as steps	370	110
min_samples_split	The minimum number of samples required to split an internal node	Range of 1to 20 with 2 as steps	3	7

4.2.2 Random Forest

As already discussed, random forest is an ensemble of decision trees. This model also has hyper-parameters similar to decision trees. the hyper parameters, their description and the initial values are described in Table 11. The grid search technique was applied here to find the best hyperparameter values. Additionally, best values are reported in Table 12.

Table 11 - Hyperparameters of Random Forest

Hyper parameter	Description	Values	Best parameters Model 2	Best parameters Model 3
N_estimator	The number of trees in the forest	10,20,50,100,200, 400,600,800,1000	400	400
max_depth	The maximum depth of the tree. If None, then nodes are expanded until all leaves are pure or until all leaves contain less than min_samples_split samples	3,5,10, 20, 30, 40, 50, 60, 70, 80, 90, 100, None	3	100
min_samples_split	The minimum number of samples required to split an internal node	2, 5, 10, 90	2	2
min_samples_leaf	The minimum number of samples required to be at a leaf node	1, 2, 4	2	1

4.2.3 Gradient Boosting Tree (GBT)

Gradient boosting is another ensemble model employed in this study. This model also has hyper-parameters similar to Random Forest. Table 12 shows the description of the hyperparameters and their corresponding value. The grid search technique was used to find the best hyperparameter values. Additionally, best values are reported in Table 12.

Table 12 - Gradient Tree Boosting Hyper-parameters

Hyper parameter	Description	Values	Model 2 best parameters	Model 3 best parameters

N_estimator	The number of boosting stages to perform	200, 400, 600, 800, 1000, 1200, 1400, 1600, 1800, 2000	600	1400
max_depth	maximum depth of the individual regression estimators. The maximum depth limits the number of nodes in the tree.	10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 110	90	40
min_samples_leaf	The minimum number of samples required to be at a leaf node	1, 2, 4	1	2
min_sample_split	The minimum number of samples required to split an internal node	2, 5, 10	5	10

5 Results

The discrete choice model coefficients and their interpretation are described in this chapter. Furthermore, the accuracy of these three models are compared and the effects of different factors are assessed as follows.

5.1 Length

As expected, the length had a negative coefficient on utility. This result is supported by both intuition and literature (Broach and Dill 2015; Guo and Loo 2013). In this study, this parameter was not sensitive to neither gender nor purpose of the trip.

5.2 Number of Turns

Number of turns is negatively associated with utility, as expected from the literature (Broach and Dill 2015). An excessive number of turns can cause more decision processing burden for the pedestrian, therefore associated with less utility, in contrast, straight lines need fewer decision process, therefore more attractive.

5.3 Scenery

In this thesis, the scenery was measured in four ways, which were average, maximum and minimum and variance of scenic index. The only variable that showed importance were variance of scenic index, with a coefficient of 15.4. Additionally, considering user characteristics (trip purpose, gender and primary and secondary intern action with scenery did not provide any meaningful coefficient, either due to statistical significance or right-sidedness.

This value indicates that people prefer places with more variation in scenery, which also reflects land use mix. More diverse land uses would result in more variance in scenic index. This confirms previous findings. Through the studies on walking from the fields of transportation, urban design and planning, and public health, it has been suggested that neighborhoods with higher residential and employment densities, more connected street patterns, and a variety of destinations show higher rate of walking (Cervero et al. 2009). This finding provides a quantitative approach for measuring land use mix in a standard way, which can be applied to other places as well.

5.4 Land Uses

The only macro level land use that showed importance were food related land ones. Other land uses such as financial, health and commercial did not show any significance. As previously stated, it has been shown in the literature that network density has strong correlation with pedestrian activity. Food land uses may indicate network density, which can be the reason why it is positively contributing to the utility.

5.5 Comparison of models

In this section, the comparison between supervised classifiers and discrete choice model (DCM) was done. Machine learning models are derived to reduce the prediction error instead of the estimation error, machine learning models outperform discrete choice models on this class prediction task. However, they lack model and parameter interpretability, desirable parameter properties, and behavioral theory soundness. There were many features that showed insignificance due to structure of discrete choice model, but they may be explanatory in other frameworks. In this part, model accuracies are presented in Table 13. It is possible to compare model 1 and model 2, because they have the same dataset. As the results show, the DCM model had lower prediction accuracy in comparison to other algorithms in model 2. It indicates that supervised learning models with same variables can show better accuracy, however they lose explanatory power.

In model 2, as expected, ensemble methods showed better predictive performance. In other words, the DT model showed slightly lower prediction accuracy (lower by 2 percent). To compare model 2 and 3, as it is seen in Table 13, using more data improved each prediction accuracy of modeling frameworks. For example, DT in model 2 has lower prediction accuracy than DT in Model 3. It indicates that although some variables showed lack of significance in DCM setting, they contain information that can lead to prediction accuracy improvement

As can be seen in this table, the GTB in model 3 showed the best result. It was expected due to advantages of GTB in handling outliers. By comparing this with DCM model, it indicated that by using built environment and user specific features with GTB model, it is possible to improve the results from 67 to 76 percent, which accounts for 9 percent improvement. This finding is also

supported by the literature, that is machine learning frameworks are capable of finding patterns that DCMs cannot find (Paredes et al. 2017).

Table 13- Comparison of accuracy of models

Model	Modeling framework	Accuracy
1	DCM	0.67
2	DT	0.72
	RF	0.74
	GTB	0.74
3	DT	0.73
	RF	0.75
	GTB	0.76

6 Summary and Conclusion

The pedestrian route choice is of interest for researchers to promote sustainability. Additionally, it is challenging due to subconscious nature of human choice. This research was conducted to model pedestrian route choice in a revealed setting using GPS data. The emphasis in this thesis was to better understand the built environment factors and personal characteristics that influence pedestrian route choice. Since stated preference studies on pedestrian route choice have shown their importance, however, they were not quantified due to complexity of measurement. The built environment is characterised by adjacent land use and scenery and their effects were investigated in this thesis.

This study consists of two main contributions. The first one is investigation of the effect of scenery and other built environment factors on pedestrian route choice. This was addressed using image recognition and deep learning techniques to estimate a measure for quantifying scenery. Additionally, incorporation of micro-level land uses was considered using Google Places API. The results showed that built environment factors could influence the pedestrian route choice, particularly food related land uses. The variance of scenic index was found to be significant, which combines both scenery and land use mix effects. The second contribution is usage of other frameworks rather than traditional discrete choice modeling framework to gain computational accuracy. This was possible using supervised machine learning techniques. Decision Tree, Random Forest and Gradient Tree Boosting were employed. The results showed that ensemble methods (Gradient Boosting Trees) showed improvement in prediction accuracy, at the expense of lack of interpretability. These models may be a reasonable substitute for the route choice step in traditional demand modeling.

6.1 Future Work

This thesis sought to measure built environment factors and especially scenery, on the pedestrian route choice. The methodology presented in our work is a novel approach to capture pedestrian route choice behavior, however additional improvements could be achieved by measuring how people interact with their environment. For example, there exists new data collection hardware that can track eye movements of people. These tools can be applied to measure how the person interacts with the environment visually. This way of measuring is out of context of transportation

due to lack of practicality, but it can be investigated if the walking behavior and the factors affecting it are part of human behavior studies.

Another improvement could focus on further enhancing some variables that are shown to be significant such as side walk width or weather. This was mainly due to lack of valid data sources for these variables. If the data were available, this can be addressed. Similarly, one might improve the proposed model by incorporating artificial intelligence frameworks such as reinforcement learning and contextual bandits, which are suitable for iterative choice analysis.

7 References

Aldrich, John H. and Forrest D. Nelson. 1984. *Linear Probability, Logit, and Probit Models*. Sage.

Alivand, Majid and Hartwig Hochmair. 2013. "Extracting Scenic Routes from VGI Data Sources." Pp. 23–30 in *Proceedings of the Second ACM SIGSPATIAL International Workshop on Crowdsourced and Volunteered Geographic Information*. ACM.

Antonini, Gianluca. 2005. "A Discrete Choice Modeling Framework for Pedestrian Walking Behavior with Application to Human Tracking in Video Sequences."

Badoe, Daniel A. and Eric J. Miller. 2000. "Transportation–land-Use Interaction: Empirical Findings in North America, and Their Implications for Modeling." *Transportation Research Part D: Transport and Environment* 5(4):235–63.

Ball, Kylie, Adrian Bauman, Eva Leslie, and Neville Owen. 2001. "Perceived Environmental Aesthetics and Convenience and Company Are Associated with Walking for Exercise among Australian Adults." *Preventive Medicine* 33(5):434–40.

Bekhor, Shlomo, Moshe E. Ben-Akiva, and M.Scott Ramming. 2006. "Evaluation of Choice Set Generation Algorithms for Route Choice Models." *Annals of Operations Research* 144(1):235–47.

Ben-Akiva, Moshe and Michel Bierlaire. 1999. "Discrete Choice Methods and Their Applications to Short Term Travel Decisions." Pp. 5–33 in *Handbook of transportation science*. Springer.

Ben-Akiva, Moshe E. and Steven R. Lerman. 1985. *Discrete Choice Analysis: Theory and Application to Travel Demand*. MIT press.

Bierlaire, Michel. 2016. *PythonBiogeme: A Short Introduction*.

Bierlaire, Michel and Emma Frejinger. 2008. "Route Choice Modeling with Network-Free Data." *Transportation Research Part C: Emerging Technologies* 16(2):187–98.

Bliemer, Michiel C. J. and John M. Rose. 2010. "Construction of Experimental Designs for Mixed Logit Models Allowing for Correlation across Choice Observations." *Transportation Research Part B: Methodological* 44(6):720–34.

Borst, Hieronymus C. et al. 2009. "Influence of Environmental Street Characteristics on Walking Route

Choice of Elderly People." *Journal of Environmental Psychology* 29(4):477–84.

Bouthelier, Fernando and Carlos F. Daganzo. 1979. "Aggregation with Multinomial Probit and Estimation of Disaggregate Models with Aggregate Data: A New Methodological Approach." *Transportation Research Part B: Methodological* 13(2):133–46.

Bovy, Piet, Shlomo Bekhor, and Carlo Prato. 2008. "The Factor of Revisited Path Size: Alternative Derivation." *Transportation Research Record: Journal of the Transportation Research Board* (2076):132–40.

Bovy, Piet H. and Eliahu Stern. 2012. *Route Choice: Wayfinding in Transport Networks: Wayfinding in Transport Networks*. Springer Science & Business Media.

Brakatsoulas, Sotiris, Dieter Pfoser, Randall Salas, and Carola Wenk. 2005. "On Map-Matching Vehicle Tracking Data." Pp. 853–64 in *Proceedings of the 31st international conference on Very large data bases*. VLDB Endowment.

Breiman, L., J. H. Friedman, R. A. Olshen, and C. J. Stone. 2005. "Classification and Regression Trees, Wadsworth International Group, Belmont, California, USA, 1984; BP Roe et Al., Boosted Decision Trees as an Alternative to Artificial Neural Networks for Particle Identification." *Nucl. Instrum. Meth. A* 543:577.

Breiman, Leo. 1996. "Bagging Predictors." *Machine Learning* 24(2):123–40.

Breiman, Leo. 2001. "Random Forests." *Machine Learning* 45(1):5–32.

Broach, Joseph and Jennifer Dill. 2015. "Pedestrian Route Choice Model Estimated from Revealed Preference GPS Data." in *Transportation Research Board 94th Annual Meeting*.

Brown, Barbara B., Carol M. Werner, Jonathan W. Amburgey, and Caitlin Szalay. 2007. "Walkable Route Perceptions and Physical Features: Converging Evidence for En Route Walking Experiences." *Environment and Behavior* 39(1):34–61.

Brown, Gavin, Jeremy Wyatt, Rachel Harris, and Xin Yao. 2005. "Diversity Creation Methods: A Survey and Categorisation." *Information Fusion* 6(1):5–20.

Brownson, Ross C., Christine M. Hoehner, Kristen Day, Ann Forsyth, and James F. Sallis. 2009. "Measuring the Built Environment for Physical Activity: State of the Science." *American Journal of Preventive Medicine* 36(4):S99–123.

Bühlmann, Peter and Sara Van De Geer. 2011. *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer Science & Business Media.

Cascetta, Ennio. 2001. "Transportation Systems." Pp. 1–22 in *Transportation Systems Engineering: Theory and Methods*. Springer.

Cascetta, Ennio, Agostino Nuzzolo, Francesco Russo, and Antonino Vitetta. 1996. "A Modified Logit Route Choice Model Overcoming Path Overlapping Problems. Specification and Some Calibration Results for Interurban Networks." in *TRANSPORTATION AND TRAFFIC THEORY. PROCEEDINGS OF THE 13TH INTERNATIONAL SYMPOSIUM ON TRANSPORTATION AND TRAFFIC THEORY, LYON, FRANCE, 24-26 JULY 1996*.

Cervero, Robert, Olga L. Sarmiento, Enrique Jacoby, Luis Fernando Gomez, and Andrea Neiman. 2009. "Influences of Built Environments on Walking and Cycling: Lessons from Bogotá." *International Journal of Sustainable Transportation* 3(4):203–26.

Chen, Chao, Xia Chen, Zhu Wang, Yasha Wang, and Daqing Zhang. 2017. "ScenicPlanner: Planning Scenic Travel Routes Leveraging Heterogeneous User-Generated Digital Footprints." *Frontiers of Computer Science* 11(1):61–74.

Daamen, Winnie. 2004. "Modelling Passenger Flows in Public Transport Facilities."

Dietterich, Thomas G. 2000. "Ensemble Methods in Machine Learning." Pp. 1–15 in *International workshop on multiple classifier systems*. Springer.

Dietterich, Thomas G., Guohua Hao, and Adam Ashenfelter. 2008. "Gradient Tree Boosting for Training Conditional Random Fields." *Journal of Machine Learning Research* 9(Oct):2113–39.

Dougherty, Mark. 1995. "A Review of Neural Networks Applied to Transport." *Transportation Research Part C: Emerging Technologies* 3(4):247–60.

Eluru, Naveen, Vincent Chakour, and Ahmed M. El-Geneidy. 2012. "Travel Mode Choice and Transit Route Choice Behavior in Montreal: Insights from McGill University Members Commute Patterns." *Public Transport* 4(2):129–49.

Ermagun, Alireza, Yingling Fan, Julian Wolfson, Gediminas Adomavicius, and Kirti Das. 2017. "Real-Time Trip Purpose Prediction Using Online Location-Based Search and Discovery Services." *Transportation Research Part C: Emerging Technologies* 77:96–112.

Farrash, Majed. 2016. “Machine Learning Ensemble Method for Discovering Knowledge from Big Data.”

Ferreira, Inês A., Maria Johansson, Catharina Sternudd, and Ferdinando Fornara. 2016. “Transport Walking in Urban neighbourhoods—Impact of Perceived Neighbourhood Qualities and Emotional Relationship.” *Landscape and Urban Planning* 150:60–69.

Ferrer, Sheila, Tomás Ruiz, and Lidón Mars. 2015. “A Qualitative Study on the Role of the Built Environment for Short Walking Trips.” *Transportation Research Part F: Traffic Psychology and Behaviour* 33:141–60.

Frejinger, Emma. 2008. “Route Choice Analysis: Data, Models, Algorithms and Applications.” *École Polytechnique Federale de Lausanne*.

Frejinger, Emma, Michel Bierlaire, and Moshe Ben-Akiva. 2009. “Expanded Path Size Attribute for Route Choice Models Including Sampling Correction.” in *International Choice Modelling Conference. Harrogate. 30 March 2009–1 April 2009*.

Freund, Yoav and Robert E. Schapire. 1997. “A Decision-Theoretic Generalization of on-Line Learning and an Application to Boosting.” *Journal of Computer and System Sciences* 55(1):119–39.

Friedman, Jerome H. 2001. “Greedy Function Approximation: A Gradient Boosting Machine.” *Annals of Statistics* 1189–1232.

Friedman, Jerome H. 2002. “Stochastic Gradient Boosting.” *Computational Statistics & Data Analysis* 38(4):367–78.

Grond, Kathryn. 2016. “Route Choice Modeling of Cyclists in Toronto.”

Gross, Richard. 2015. *Psychology: The Science of Mind and Behaviour 7th Edition*. Hodder Education.

Guo, Zhan. 2009. “Does the Pedestrian Environment Affect the Utility of Walking? A Case of Path Choice in Downtown Boston.” *Transportation Research Part D: Transport and Environment* 14(5):343–52.

Guo, Zhan and Becky P. Y. Loo. 2013. “Pedestrian Environment and Route Choice: Evidence from New York City and Hong Kong.” *Journal of Transport Geography* 28:124–36.

Hagenauer, Julian and Marco Helbich. 2017. “A Comparative Study of Machine Learning Classifiers for Modeling Travel Mode Choice.” *Expert Systems with Applications* 78:273–82.

Hill, Michael R. 1982. “Spatial Structure and Decision-Making Aspects of Pedestrian Route Selection

through an Urban Environment."

Hintaran, R. E. 2016. "Unravelling Urban Pedestrian Trips: Developing a New Pedestrian Route Choice Model Estimated from Revealed Preference GPS Data."

Hochmair, Hartwig H. 2010. "Spatial Association of Geotagged Photos with Scenic Locations."

Hoogendoorn-Lanser, Sascha. 2005. *Modelling Travel Behaviour in Multi-Modal Networks*.

Hoogendoorn, Serge P. and Piet H. L. Bovy. 2004. "Pedestrian Route-Choice and Activity Scheduling Theory and Models." *Transportation Research Part B: Methodological* 38(2):169–90.

Koh, P. P. and Y. D. Wong. 2013. "Influence of Infrastructural Compatibility Factors on Walking and Cycling Route Choices." *Journal of Environmental Psychology* 36:202–13.

Koh, Puay Ping and Yiik Diew Wong. 2013. "Comparing Pedestrians' Needs and Behaviours in Different Land Use Environments." *Journal of Transport Geography* 26:43–50.

Kowsari, Kamran, Mojtaba Heidarysafa, Donald E. Brown, Kiana Jafari Meimandi, and Laura E. Barnes. 2018. "RMDL: Random Multimodel Deep Learning for Classification." *arXiv Preprint arXiv:1805.01890*.

Lawrence, Rick, Andrew Bunn, Scott Powell, and Michael Zambon. 2004. "Classification of Remotely Sensed Imagery Using Stochastic Gradient Boosting as a Refinement of Classification Tree Analysis." *Remote Sensing of Environment* 90(3):331–36.

Leccese, Michael and Kathleen McCormick. 2000. *Charter of the New Urbanism*. McGraw-Hill Professional.

Liang, Guohua, Xingquan Zhu, and Chengqi Zhang. 2011. "An Empirical Study of Bagging Predictors for Different Learning Algorithms." in *AAAI*.

Lior, Rokach. 2014. *Data Mining with Decision Trees: Theory and Applications*. World scientific.

Lue, Gregory. 2017. "Estimating a Toronto Pedestrian Route Choice Model Using Smartphone GPS Data: It's Not the Destination, but the Journey, That Matters."

Lue, Gregory and Eric J. Miller. 2018. *Estimating a Toronto Pedestrian Route Choice Model Using Smartphone GPS Data*.

Manski, Charles F. 1977. "The Structure of Random Utility Models." *Theory and Decision* 8(3):229–54.

McFadden, Daniel. 1978. “Modeling the Choice of Residential Location.” *Transportation Research Record* (673).

McFadden, Daniel. 1986. “The Choice Theory Approach to Market Research.” *Marketing Science* 5(4):275–97.

Owen, Neville, Nancy Humpel, Eva Leslie, Adrian Bauman, and James F. Sallis. 2004. “Understanding Environmental Influences on Walking: Review and Research Agenda.” *American Journal of Preventive Medicine* 27(1):67–76.

Paredes, Miguel, Erik Hemberg, Una-May O’Reilly, and Chris Zegras. 2017. “Machine Learning or Discrete Choice Models for Car Ownership Demand Estimation and Prediction?” Pp. 780–85 in *Models and Technologies for Intelligent Transportation Systems (MT-ITS), 2017 5th IEEE International Conference on*. IEEE.

Patterson, Zachary. 2017. “MTL Trajet 2016.” *Presented at the 11th International Conference on Travel Survey Methods, Esterel, Quebec* 1–15.

Patterson, Zachary and Kyle Fitzsimmons. 2016. “DataMobile: Smartphone Travel Survey Experiment.” *Transportation Research Record: Journal of the Transportation Research Board* (2594):35–43.

Patterson, Zachary, Kyle Fitzsimmons, Michael Widener, Jessica Reid, and David Hammond. 2018. *Recruitment, Burden, Incentives and Participation in Smartphone Travel Surveys*.

Pedregosa, Fabian et al. 2011. “Scikit-Learn: Machine Learning in Python.” *Journal of Machine Learning Research* 12(Oct):2825–30.

Peterson, George L. 1967. “A Model of Preference: Quantitative Analysis of the Perception of the Visual Appearance of Residential Neighborhoods.” *Journal of Regional Science* 7(1):19–31.

Prato, Carlo Giacomo. 2009. “Route Choice Modeling: Past, Present and Future Research Directions.” *Journal of Choice Modelling* 2(1):65–100.

Quddus, Mohammed A., Washington Y. Ochieng, and Robert B. Noland. 2007. “Current Map-Matching Algorithms for Transport Applications: State-of-the Art and Future Research Directions.” *Transportation Research Part c: Emerging Technologies* 15(5):312–28.

Quercia, Daniele, Rossano Schifanella, and Luca Maria Aiello. 2014. “The Shortest Path to Happiness: Recommending Beautiful, Quiet, and Happy Routes in the City.” Pp. 116–25 in *Proceedings of the*

25th ACM conference on Hypertext and social media. ACM.

Quinlan, J.Ross. 1986. "Induction of Decision Trees." *Machine Learning* 1(1):81–106.

Ramming, M. S. 2002. "Network Knowledge and Route Choice. PhD."

Rodriguez, Daniel A., Elizabeth Brisson, and Nicolas Estupinan. 2009. *Relationship Between Segment-Level Built Environment Attributes and Pedestrian Activity Around Bogota's Bus Rapid Transit Stations*.

Roof, Karen. 2008. "Public Health: Seattle and King County's Push for the Built Environment." *Journal of Environmental Health* 71(1):24.

Runge, Nina, Pavel Samsonov, Donald Degraen, and Johannes Schöning. 2016. "No More Autobahn!: Scenic Route Generation Using Googles Street View." Pp. 147–51 in *Proceedings of the 21st International Conference on Intelligent User Interfaces*. ACM.

Russell, Stuart J. and Peter Norvig. 2016. *Artificial Intelligence: A Modern Approach*. Malaysia; Pearson Education Limited,.

Salesses, Philip, Katja Schechtner, and César A. Hidalgo. 2013. "The Collaborative Image of the City: Mapping the Inequality of Urban Perception." *PloS One* 8(7):e68400.

Samuel, Arthur L. 1959. "Some Studies in Machine Learning Using the Game of Checkers." *IBM Journal of Research and Development* 3(3):210–29.

Schüssler, Nadine. 2010. "Accounting for Similarities between Alternatives in Discrete Choice Models Based on High-Resolution Observations of Transport Behaviour."

Seneviratne, P. N. and J. F. Morrall. 1985. "Analysis of Factors Affecting the Choice of Route of Pedestrians." *Transportation Planning and Technology* 10(2):147–59.

Seresinhe, Chanuki Illushka, Tobias Preis, and Helen Susannah Moat. 2017. "Using Deep Learning to Quantify the Beauty of Outdoor Places." *Open Science* 4(7):170170.

Simon, Herbert A. 1959. "Theories of Decision-Making in Economics and Behavioral Science." *The American Economic Review* 49(3):253–83.

Skurichina, Marina and Robert P. W. Duin. 2002. "Bagging, Boosting and the Random Subspace Method for Linear Classifiers." *Pattern Analysis & Applications* 5(2):121–35.

Sun, Bingrong and Byungkyu Brian Park. 2017. “Route Choice Modeling with Support Vector Machine.” *Transportation Research Procedia* 25:1806–14.

Talen, Emily and Julia Koschinsky. 2013. “The Walkable Neighborhood: A Literature Review.” *International Journal of Sustainable Land Use and Urban Planning* 1(1).

Tan, Pang-Ning, Michael Steinbach, and Vipin Kumar. 2005. “Introduction to Data Mining. 1st.”

Train, Kenneth E. 2009. *Discrete Choice Methods with Simulation*. Cambridge university press.

Verlander, Neville Q. and Benjamin G. Heydecker. 1997. “Pedestrian Route Choice: An Empirical Study.” PTRC Education and Research Services Ltd.

Van der Waerden, PJHJ, A. W. J. Borgers, and H. J. P. Timmermans. 2004. “Choice Set Composition in the Context of Pedestrians’ Route Choice Modeling.”

Walker, Joan Leslie. 2001. “Extended Discrete Choice Models: Integrated Framework, Flexible Error Structures, and Latent Variables.”

Wardman, Mark. 1988. “A Comparison of Revealed Preference and Stated Preference Models of Travel Behaviour.” *Journal of Transport Economics and Policy* 71–91.

Weinstein Agrawal, Asha, Marc Schlossberg, and Katja Irvin. 2008. “How Far, by Which Route and Why? A Spatial Analysis of Pedestrian Preference.” *Journal of Urban Design* 13(1):81–98.

Xie, Chi, Jinyang Lu, and Emily Parkany. 2003. “Work Travel Mode Choice Modeling with Data Mining: Decision Trees and Neural Networks.” *Transportation Research Record: Journal of the Transportation Research Board* (1854):50–61.

Yamamoto, Toshiyuki, Ryuichi Kitamura, and Junichiro Fujii. 2002. “Drivers’ Route Choice Behavior: Analysis by Data Mining Algorithms.” *Transportation Research Record: Journal of the Transportation Research Board* (1807):59–66.

Yang, Jian and Liqiu Meng. 2015. “Feature Selection in Conditional Random Fields for Map Matching of GPS Trajectories.” Pp. 121–35 in *Progress in Location-Based Services 2014*. Springer.

Yazdizadeh, Ali, Zachary Patterson, and Bilal Farooq. 2018. “An Automated Approach from GPS Traces to Complete Trip Information.” *Transportation Research Record: Journal of the Transportation Research Board*.

Zhang, Heping and Minghui Wang. 2009. “Search for the Smallest Random Forest.” *Statistics and Its Interface* 2(3):381.

Zheng, Yan-Tao et al. 2013. “GPSView: A Scenic Driving Route Planner.” *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 9(1):3.

Zhou, Bolei, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2017. “Places: A 10 Million Image Database for Scene Recognition.” *IEEE Transactions on Pattern Analysis and Machine Intelligence* 8828(c):1–14.

Zhou, Bolei, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. 2014. “Learning Deep Features for Scene Recognition Using Places Database.” Pp. 487–95 in *Advances in neural information processing systems*.

Appendix

Table 14-variables of supervised learning models

Variable	Description	Values
travel_mode_study	A dummy variable indicating primary travel mode to work, Values correspond to the following respectively: On foot, Bicycle, Public transportation, Car, Car and public transportation, Other modes, Other combinations	0,1,2,3,4,5,6
travel_mode_alt_study	A dummy variable indicating secondary travel mode to study, Values correspond to the following respectively: not applicable, on foot, bicycle, public transportation, car, car and public transportation, other modes, other combinations	0,1,2,3,4,5,6,7
travel_mode_work	A dummy variable indicating primary travel mode to work, Values correspond to the following respectively: On foot, Bicycle, Public transportation, Car, Car and public transportation, Other modes, Other combinations	0,1,2,3,4,5,6
travel_mode_alt_work	A dummy variable indicating second travel mode to work, Values correspond to the following respectively: not applicable, on foot, bicycle, public transportation, car, car and public transportation, other modes, other combinations	0,1,2,3,4,5,6,7

Sex	Participant gender , Values correspond to the following respectively: Male, Female, Other/Neither	0, 1, 2
age_16_24	A dummy variable indicating whether the participant has age between 16 to 24	0,1
age_25_34	A dummy variable indicating whether the participant has age between 25 to 34	0,1
age_35_44	A dummy variable indicating whether the participant has age between 35 to 44	0,1
age_45_54	A dummy variable indicating whether the participant has age between 45 to 54	0,1
age_55_64	A dummy variable indicating whether the participant has age between 55 to 64	0,1
age_65	A dummy variable indicating whether the participant has more than 65 years old	0,1
purpose_education	A dummy variable indicating whether the trip purpose was educational	0,1

purpose_health	A dummy variable indicating whether the trip purpose was health	0,1
purpose_other	A dummy variable indicating whether the trip purpose was not specified	0,1
purpose_meal_snack_coffee	A dummy variable indicating whether the trip purpose was having meal snack coffee	0,1
purpose_leisure	A dummy variable indicating whether the trip purpose was leisure	0,1
purpose_pick_up	A dummy variable indicating whether the trip purpose was to pick up other family members	0,1
Distance	The distance of each alternative in meters	float
duration_seconds	The estimated duration of each alternative in seconds	float
number_of_turns	The number of turns in each alternative	integer
avg_num_places	Average number of place types in all coordinates that correspond to an alternative	float

var_num_places	Variance of number of place types in all coordinates that correspond to an alternative	float
avg_rating_avg	Average of rating averages, The ratings for places provided by Google Places API were averaged for each coordinate. Then for each alternative, these values were averaged	float
var_rating_average	Variance of rating averages, The ratings for places provided by Google Places API were averaged for each coordinate. Then for each alternative, these values variances were calculated	float
avg_scenic_index	Average scenic index, The average of scenic indices of coordinates of each alternative	float
var_scenic_index	Variance of scenic index, The variance of scenic indices of each alternative	float
max_scenic_index	Maximum scenic index, The maximum of scenic indices of each alternative	float

min_scenic_index	Minimum scenic index	float
sum_scenic_index	Summation of scenic index	float
Numpoints	Number of signalized intersection in each alternative	integer
sum_accout	Summation of all accounting land uses in each alternative	integer
sum_tags_art_gallery	Summation of all art gallery land uses in each alternative	integer
sum_tags_atm	Summation of all Automated Teller Machine land uses in each alternative	integer
sum_tags_bakery	Summation of all bakery land uses in each alternative	integer
sum_tags_bank	Summation of all bank land uses in each alternative	integer
sum_tags_bar	Summation of all bar land uses in each alternative	integer
sum_tags_beauty_salon	Summation of all beauty salon land uses in each alternative	integer
sum_tags_bicycle_store	Summation of all bicycle store land uses in each alternative	integer
sum_tags_book_store	Summation of all book store land uses in each alternative	integer
sum_tags_bowling_alley	Summation of all bowling alley land uses in each alternative	integer

sum_tags_bus_station	Summation of all bus station land uses in each alternative	integer
sum_tags_cafe	Summation of all cafe land uses in each alternative	integer
sum_tags_car_dealer	Summation of all car dealer land uses in each alternative	integer
sum_tags_car_rental	Summation of all car rental land uses in each alternative	integer
sum_tags_car_repair	Summation of all car repair land uses in each alternative	integer
sum_tags_car_wash	Summation of all car wash land uses in each alternative	integer
sum_tags_cemetery	Summation of all cemetery land uses in each alternative	integer
sum_tags_church	Summation of all church land uses in each alternative	integer
sum_tags_city_hall	Summation of all city hall land uses in each alternative	integer
sum_tags_clothing_store	Summation of all clothing store land uses in each alternative	integer
sum_tags_convenience_store	Summation of all convenience store land uses in each alternative	integer
sum_tags_courthouse	Summation of all courthouse land uses in each alternative	integer
sum_tags_dentist	Summation of all dentist land uses in each alternative	integer

sum_tags_department_store	Summation of all department store land uses in each alternative	integer
sum_tags_doctor	Summation of all doctor land uses in each alternative	integer
sum_tags_electrician	Summation of all electrician land uses in each alternative	integer
sum_tags_electronics_store	Summation of all electronics store land uses in each alternative	integer
sum_tags_embassy	Summation of all embassy land uses in each alternative	integer
sum_tags_establishment	Summation of all establishment land uses in each alternative	integer
sum_tags_finance	Summation of all finance land uses in each alternative	integer
sum_tags_fire_station	Summation of all fire station land uses in each alternative	integer
sum_tags_florist	Summation of all florist land uses in each alternative	integer
sum_tags_food	Summation of all food related land uses in each alternative	integer
sum_tags_funeral_home	Summation of all funeral home land uses in each alternative	integer
sum_tags_furniture_store	Summation of all furniture store land uses in each alternative	integer

sum_tags_gas_station	Summation of all gas station land uses in each alternative	integer
sum_tags_general_contractor	Summation of all general contractor land uses in each alternative	integer
sum_tags_grocery_or_supermarket	Summation of all grocery or supermarket land uses in each alternative	integer
sum_tags_gym	Summation of all gym land uses in each alternative	integer
sum_tags_hair_care	Summation of all hair careland uses in each alternative	integer
sum_tags_hardware_store	Summation of all hardware store land uses in each alternative	integer
sum_tags_health	Summation of all health land uses in each alternative	integer
sum_tags_home_goods_store	Summation of all home goods store land uses in each alternative	integer
sum_tags_hospital	Summation of all hospital land uses in each alternative	integer
sum_tags_insurance_agency	Summation of all insurance agency land uses in each alternative	integer
sum_tags_jewelry_store	Summation of all jewelry store land uses in each alternative	integer

sum_tags_laundry	Summation of all laundry land uses in each alternative	integer
sum_tags_lawyer	Summation of all lawyer land uses in each alternative	integer
sum_tags_library	Summation of all library land uses in each alternative	integer
sum_tags_liquor_store	Summation of all liquor store land uses in each alternative	integer
sum_tags_local_government_office	Summation of all local government office land uses in each alternative	integer
sum_tags_locality	Summation of all locality land uses in each alternative	integer
sum_tags_locksmith	Summation of all locksmith land uses in each alternative	integer
sum_tags_lodging	Summation of all lodging land uses in each alternative	integer
sum_tags_meal_delivery	Summation of all meal delivery land uses in each alternative	integer
sum_tags_meal_takeaway	Summation of all meal take away land uses in each alternative	integer
sum_tags_mosque	Summation of all mosque land uses in each alternative	integer
sum_tags_movie_rental	Summation of all movie rental land uses in each alternative	integer

sum_tags_movie_theater	Summation of all movie theater land uses in each alternative	integer
sum_tags_moving_company	Summation of all moving company land uses in each alternative	integer
sum_tags_museum	Summation of all museum land uses in each alternative	integer
sum_tags_neighborhood	Summation of all neighborhood land uses in each alternative	integer
sum_tags_night_club	Summation of all night club land uses in each alternative	integer
sum_tags_painter	Summation of all painter land uses in each alternative	integer
sum_tags_park	Summation of all park land uses in each alternative	integer
sum_tags_parking	Summation of all parking land uses in each alternative	integer
sum_tags_pet_store	Summation of all pet store land uses in each alternative	integer
sum_tags_pharmacy	Summation of all pharmacy land uses in each alternative	integer
sum_tags_physiotherapist	Summation of all physiotherapist land uses in each alternative	integer
sum_tags_place_of_worship	Summation of all worship related land uses in each alternative	integer

sum_tags_plumber	Summation of all plumber land uses in each alternative	integer
sum_tags_point_of_interest	Summation of all points of interest in each alternative	integer
sum_tags_police	Summation of all police land uses in each alternative	integer
sum_tags_political	Summation of all political land uses in each alternative	integer
sum_tags_post_office	Summation of all post office land uses in each alternative	integer
sum_tags_premise	Summation of all premises in each alternative	integer
sum_tags_real_estate_agency	Summation of all real estate agency land uses in each alternative	integer
sum_tags_restaurant	Summation of all restaurant land uses in each alternative	integer
sum_tags_roofing_contractor	Summation of all roofing contractor land uses in each alternative	integer
sum_tags_route	Summation of all route tags in each alternative	integer
sum_tags_school	Summation of all school land uses in each alternative	integer
sum_tags_shoe_store	Summation of all shoe store land uses in each alternative	integer
sum_tags_shopping_mall	Summation of all shopping mall land uses in each alternative	integer

sum_tags_spa	Summation of all spa land uses in each alternative	integer
sum_tags_storage	Summation of all storage land uses in each alternative	integer
sum_tags_store	Summation of all store land uses in each alternative	integer
sum_tags_sublocality	Summation of all sub locality in each alternative	integer
sum_tags_sublocality_level_1	Summation of all land uses in each alternative	integer
sum_tags_subway_station	Summation of all sublocality_level_1 in each alternative	integer
sum_tags_supermarket	Summation of all supermarket land uses in each alternative	integer
sum_tags_synagogue	Summation of all synagogue land uses in each alternative	integer
sum_tags_train_station	Summation of all train station in each alternative	integer
sum_tags_transit_station	Summation of all transit stations in each alternative	integer
sum_tags_travel_agency	Summation of all travel agency land uses in each alternative	integer
sum_tags_university	Summation of all university land uses in each alternative	integer
sum_tags_veterinary_care	Summation of all veterinary care land uses in each alternative	integer