An Ensemble Approach to Route Choice

HAOTIAN WANG

Supervisor: David Levinson
Associate Supervisor: Emily Moylan

A thesis submitted in fulfilment of
the requirements for the degree of
Doctor of Philosophy

School of Civil Engineering
Faculty of Engineering
The University of Sydney
Australia

12 June 2025

CHAPTER 1

Statement of Originality

This thesis has not been submitted for any degree or other purposes. I certify that the intellectual content of this thesis is the product of my own work and that all the assistance received in preparing this thesis and sources have been acknowledged. I have read and understood the University of Sydney Student Plagiarism: Coursework Policy and Procedure. I understand that failure to comply with the University of Sydney Student Plagiarism: Coursework Policy and Procedure can lead to the University commencing proceedings against me for potential student misconduct under chapter 8 of the University of Sydney By-Law 1999 (as amended).

Author: Haotian WANG

Signature:

Date: 21/08/2024

CHAPTER 2

Author Attribution Statement

The work contained in the body of this dissertation, except otherwise acknowledged, is the result of my own research and investigations. The work related to choice set generation was published in the Transportation Research Record (Wang et al. 2024). The work related to the ensemble approach has been accepted for publication in Transportation Research Part C. Appendix C was published in Findings (Wang et al. 2022). All manuscripts were written by Haotian Wang and supervised by David Levinson and Emily Moylan.

In addition to the statements above, in cases where I am not the corresponding author of a published item, permission to include the published material has been granted by the corresponding author.

-Haotian Wang

Signature:

21 August 2024

As the supervisor for the candidature upon which this thesis is based, I can confirm that the authorship attribution statements above are correct.

-David Levinson

Signature:

21 August 2024

-Emily Moylan

Signature:

21 August 2024

CHAPTER 3

Abstract

Understanding automobile drivers' route preferences allows for the calculation of traffic flow on network segments and helps in assessing facility requirements, costs, and the impact of network modifications. This can be seen as a two-stage problem. For a large urban network, available routes from an origin to a destination, which are difficult to itemize, could be numerous. Therefore, as the first step, a clear choice set of available routes for a trip should be extracted from the network before attempting to predict choice of route. The second step is predicting individuals' route choices, which starts at the individual level, whereas traffic assignment begins at the network level and aims to determine how many travelers use each link in the network. Much previous research employs logit-based choice methods to model the route choices of individuals, but machine learning models are gaining increasing interest. However, all of these methods typically rely on a single 'best' model for predictions, which may be sensitive to measurement errors in the training data. Moreover, predictions from discarded models might still provide insights into route choices. The ensemble approach combines outcomes from multiple individual models using various pattern recognition methods, assumptions, and/or data sets to deliver improved predictions. When configured correctly, ensembles offer greater prediction accuracy and better account for uncertainties.

To examine the advantages of ensemble techniques, a hybrid method, which combines labelling and link-penalty approaches, is applied to a high resolution road network to prepare the choice set for route choice modelling. A data set from the I-35W Bridge Collapse study in 2008, and another from the 2011 Travel Behavior Inventory (TBI), both in Minneapolis - St. Paul (The Twin Cities) are each used to train a set of route choice models. These models are combined with ensemble techniques. The analyses consider travellers' socio-demographics and trip attributes.

The trained models are applied to two datasets, the Longitudinal Employer-Household Dynamics (LEHD) commute trips and TBI morning peak trips, for validation. Predictions are also compared with the loop detector records on freeway links. Traditional Multinomial Logit and Path-size Logit models, along with machine learning methods such as Decision Tree, Random Forest, Extra Tree, AdaBoost, Support Vector Machine, and Neural Network, serve as the foundation for this study. Ensemble rules are tested in both case studies, including hard voting, soft voting, ranked choice voting, and stacking.

Based on the results, using the 10 best labels identified in the study, the choice set captures most observed trajectories without needing to remove any links from the road network. A new similarity measure, which considers the influence of overlap, attribute similarity, and spatial similarity between routes, is proposed and applied to evaluate route choice models. We conclude that ensembles, when properly applied, perform better than base models in route choice prediction in the datasets in the study. Additionally, heterogeneous ensembles using soft voting outperform both the base models and other ensemble rules on testing sets, with unscaled weights in soft voting proving to be more robust based on external validation.

Acknowledgements

I would like to express my deepest gratitude to those who have supported and guided me throughout my doctoral journey.

First and foremost, I would like to express my sincere thanks to my supervisor, Professor David Levinson. I will never forget the first day we met — he gave me a tour of the office, which ended in his own office, where we came up with many incredible ideas, shared interesting everyday experiences, and discussed how advanced technologies might influence our lives. His invaluable advice, encouragement, and expertise have been essential to my research. I am profoundly grateful for his great mentoring and insightful guidance throughout the development of my Ph.D. journey. I would also like to express my deepest gratitude to my co-supervisor, Dr. Emily Moylan. She has always been there to help me navigate both academic and life challenges, offering solutions and encouragement whenever I found myself in difficult situations. Without their help, I would not be where I am today.

Second, I'm grateful to my external reviewers, Professor Carlo Prato and Professor Shlomo Bekhor, for providing invaluable feedback on the early drafts. I also wish to thank my committee members, Associate Professor Mohsen Ramezani and Dr. Benjy Marks, for their insightful feedback on my annual progress.

Furthermore, I would like to thank my friends, Bahman Lahoorpoor, Changle Song, Hema Rayaprolu, Laya Hossein Rashidi, Ruihao Zeng, Tingsen Xian, Yang Gao, Yadi Wang, Yue Yang, Zhuopeng Xie and Zhaohan Wang. Their friendship has been a vital source of strength and comfort during this demanding period. The moments we shared have enriched my life and provided much-needed balance. I am also grateful to the members of TransportLab at the University of Sydney. Their collaboration, feedback, and camaraderie have significantly contributed to the success of this research. Working alongside such dedicated and talented individuals has been a truly rewarding experience.

In particular, I would like to thank my girlfriend, Jingwei Hu, for her steadfast love and support, helping me to rise again whenever I felt defeated.

Lastly, I deeply appreciate my mother, Lihong Kou, and my father, Jun Wang, whose financial support and constant encouragement have made this journey possible. Their belief in my abilities has been a continual source of motivation.

Nomenclature

α	is the weight of attribute similarity in measuring overall similarity
$\boldsymbol{\beta}$	is the coefficient vector
$\delta_{a,j}$	is 1 if alternative route j includes link a, and 0 for otherwise
$\epsilon_{n,i}$	is independent and identically distributed noise following an extreme value distribution
Λ	is the area of the region which is formed by predicted route and chosen route
Φ	is the total number of loop detectors (indexed by ϕ) included in this study
ψ_x	is the Shapley value for player (feature) x
Ω	is the overlap rate between two routes
a	is the link a in route i
c	percentage of the set of observed trajectories that are captured by the generated routes under a threshold
d	is the average deviation between predicted route and chosen route
D	is the deviation between predicted route and chosen route
$1-D$	is the spatial similarity between predicted route and chosen route
E_h	is the employment opportunities at destination h
E_{net}	is the network percentage error between prediction and loop detector observation
$f(j)$	is the model's prediction for instance j
F_i	is the freeway coverage of route i
JSD	is the Jensen-Shannon Divergence
L_k	is the total length of the non-overlapped links in the predicted route
L_s	is the total length of links that are shared between predicted route and chosen route
L_t	is the total length of the chosen route
l_a	is the length of link a in alternative route i
l_i	is the total length of the alternative route i
$N_{g,h}$	is the number of trips from origin g to destination h

R_h	is the resident population at origin h
S	is the similarity between two routes
s	is the 'attribute similarity' between predicted and chosen route
$\mathbf{X}$	is a vector which includes all input variables
$\mathbf{M}$	is a subset of $\mathbf{X}$ that does not include feature x
$U_{n,i}$	is the utility of alternative route i for traveller n
$V_{n,i}$	is the deterministic part in the utility function
w_m	is the weight of base model m
w_x	is the weight of variable x in similarity measure
$\hat{x_b}$	is attribute x of non-overlapped parts of predicted route
x_b	is variable x of non-overlapped parts in chosen route

Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

1.1 Background

Route choice modelling predicts the routes that individuals or vehicles will take when traveling from one location to another. It is the core of traffic assignment, which is the last step in the strategic transport modelling process. By knowing how many trips are generated between an origin and a destination, traffic assignment allocates these trips to routes that connect the origin and destination. Route choice models include characteristics of alternative routes and travelers to predict which route(s) among the alternatives will be chosen. Since a route consists of a series of links, aggregating trips on links from multiple routes allows forecasting link-level traffic.

Traditionally, the routes with the shortest distance or minimum travel time for travelers are taken as the chosen route (Golledge 1995), which assumes people have perfect knowledge about path costs. However, it is unrealistic for travelers to possess perfect information, and different people perceive travel costs in different ways. For instance, Prato and Bekhor (2006), Zhu (2010) and Zhang et al. (2009) show that the shortest routes by distance or travel time are not always, or even generally, chosen by individuals. For a large urban network, potential routes, which are difficult to itemize, could number in the thousands (Prato 2009a), and thus a clear choice set of available routes for a trip should be explicit from the network before route choice modelling.

With a finite choice set, and based on random utility theory and discrete choice modelling framework, a series of logit models has been applied to model people's route choice. These

models include a deterministic and a random component in their utility function. The deterministic part measures the utility based on observed attributes of the route, and the random component captures the the variability and randomness in preferences. The main advantages of traditional statistical approaches like logit are their easy interpretability and implementation. With a well-fit logit model, the estimated parameters of tested features can directly reflect how they influence choice behaviour. In addition, the probability of choosing a specific route can be calculated.

Machine learning approaches, which can more easily capture the non-linear relationships that logit models cannot, and generally provide more accurate predictions than logit models (Lee et al. 2018; Darwish et al. 2024; Zhao et al. 2020), have attracted attention (Van Cranenburgh et al. 2022). However, most machine learning models are highly sensitive to data, running the risk of over-fitting. There is no guarantee that machine learning models will always outperform statistical models, especially when applied to a new dataset. For both statistical and machine learning models (Zhu and Levinson 2015; Zhang and Levinson 2008; Zhang 2011; Xiong et al. 2018; Zhu and Levinson 2018), the aim has traditionally been to find the 'best' model based on the given information. However, the 'best' model can shift based on the level of noise or different data sets used for training or testing. Moreover, the 'non-best' models still have a chance to make a correct prediction when the 'best' model makes mistakes. Therefore, we hypothesise that including all the results from each individual model, rather than abandoning them, can provide an even more accurate result than the single 'best' model.

Ensemble approaches combine various models or data from different sources to produce an 'ensemble solution.' These have been shown to provide a more accurate result than individual models in weather forecasting (Palmer 2019), economic prediction (Armstrong 2001; McNees 1990), and disease forecasting (Ray et al. 2020; Sharma et al. 2021), and the transport field shares similar uncertainty with those fields (Wu and Levinson 2021). Techniques such as bagging (bootstrap aggregating) primarily reduce variance by combining predictions from diverse models trained on different data subsets (Breiman 1996a), while boosting sequentially reduces bias by focusing on errors made by previous models (Breiman 1996b). Heterogeneous ensembles combine different types of base models to leverage the strengths of each model

type and capture a wider range of patterns in the data, effectively reducing the overall bias. By combining the predictions of multiple models, the ensemble reduces the impact of the variance from any single model. The combining process for different ensembles varies based on the problem's structure, where the errors made by individual models in ensembles tend to cancel out, leading to more accurate and reliable route choice predictions compared to single models.

Unlike discrete choice problems such as mode choice, where alternatives are fairly distinct, the routes in the choice set for the route choice problem typically share some links. For example, the shortest distance route may share links with the minimum travel time route. This overlap between alternative routes makes route choice less discrete than other choice problems and increases the difficulty in modelling and prediction. This is why Path-size Logit was proposed over a simpler logit form, but most studies include the overlap between routes as a variable or component in utility function. The overlap also can appear between the predicted route and the chosen route.

1.2 Problem Statement

Despite numerous advances in route choice modelling, existing models frequently struggle with predicting route choices accurately on high-resolution urban road networks. Individual preferences and behaviors can vary widely and are often unpredictable, making modelling people's route choices difficult. Existing applications often rely on one 'best' model and simplistic assumptions, leading to non-robust and sub-optimal performance. There is a pressing need for more powerful approaches that can integrate various data and capture the true pattern in a more accurate way.

1.3 Research Objectives and Goals

The primary objective of this research is to develop the ensemble approach for route choice that leverages multiple prediction techniques to improve accuracy. The specific goals of this study are to:

(1) Review and analyze the limitations of current route choice models.
(2) Demonstrate that the 'best' route choice model varies with the training and testing data.
(3) Design, test, and implement ensemble approaches that combine different predictive algorithms.
(4) Develop new methods to evaluate route choice prediction.
(5) Test whether improvement in model performance results from applying ensemble techniques to multiple models.
(6) Validate the network-wide performance of the ensembles, and other tested models, to predict travellers' route choice by comparing the aggregate of predicted choices to observed values from loop detector records.
(7) Evaluate a set of ensemble techniques and provide recommendations for implementation and future research.

1.4 Significance and Contributions

The proposed ensemble approach aims to address the shortcomings of existing route choice models by offering a more robust and accurate solution. By integrating diverse data sources and leveraging the strengths of multiple predictive algorithms, this study contributes to the development of more accurate and adaptive transport planning tools.

This research attempts makes several contributions to route choice, including:

(1) Developing a new choice set generation technique without removing any non-cyclic links in a large, high-resolution network;

(2) Developing a new similarity measure between routes by including the effect of overlap, attribute similarity, and spatial similarity; and
(3) Developing a framework for applying ensemble techniques to route choice modelling

Each of these contributions is described below.

New Choice Set Generation Approach for High-Resolution Network. Since the numerous links and nodes in the high-resolution road network significantly increase the difficulty of forming a satisfactory choice set, many studies reduce links from the original network to simplify the problem. However, the high similarity of the high-resolution graph with the real-life network suggests that the analysis results are more useful. Therefore, this study finds a strategy to capture the most observed trajectories with a minimum number of algorithms. The proposed hybrid method uses free-flow and congested speed data, which include a distribution of vehicles' travel speeds for road links between 5:00 am and 9:00 am. Compared to networks of similar size, the proposed method requires a smaller choice set but covers more trips.

New Similarity Metric between Routes. Alternative routes in route choice modelling share some links, resulting in a choice set which differs from many other discrete choice problems and also makes route choice modelling more difficult. An incorrect prediction might share most of the links with the selected route, or a wrong prediction might have the same route attributes but just deviate a few meters from the chosen route. These mis-predicted results are much better than a route that is totally distinct from the chosen route, especially for link-level forecasting. However, a measure to describe the similarity between routes by considering not only the overlap between links from two routes but also how close the routes are in terms of characteristics and spatial proximity is missing. Therefore, this study proposes a new similarity metric that includes all the aspects discussed above, and demonstrates its effectiveness on a simple grid network and a real urban network.

Ensemble Techniques for Route Choice Modelling. The advantages in prediction accuracy and robustness of ensembles have been demonstrated in many fields but have not yet been demonstrated for the route choice problem. This study includes both traditional logit

models and popular machine learning models as base models and tests different strategies for ensemble learning to combine results from individual models. By applying models in multiple random tests on two different data sets, ensembles using soft voting techniques show better performance in prediction accuracy and robustness. Using ensemble techniques like bagging and feature randomness improves the model's sensitivity and log-likelihood. The SHAP values help to explain ensembles, but remain harder to interpret than logit models.

1.5 Structure of the Dissertation

This dissertation is organized as follows: Chapter 2 reviews the relevant literature on choice set generation, route choice models, and ensemble learning techniques. Chapter 3 introduces the datasets used in this research, including their sources and characteristics. Chapter 4 details the research methodology, encompassing choice set generation, model development, model evaluation, and validation processes. Chapter 5 presents the results from the case studies and the validation process, providing a thorough analysis of the ensemble's performance. Chapter 6 discusses the findings of this study and concludes, including limitations and recommendations for future research, while highlighting the contributions and potential applications of the study.

CHAPTER 2

Literature Review

Route choice modelling can be seen as a two-step problem: generating a feasible choice set, and predicting individual route choices.

The term 'route choice' differs from 'route assignment': the first refers to the decision-making process of individual travelers, the second refers to the broader process of allocating traffic to different routes within a network. Ideally, the second is just the aggregation of the first, however, when assigning many travelers the characteristics of the road network change, in particular travel times, which affects individual choices, and so equilibrium processes are often used (Wardrop 1952). Route choice applications typically assume travel times are exogenous based on actual traffic conditions, and don't solve for the equilibrium.

This chapter reviews various existing choice set generation methods and understand their strengths and limitations. Since the size of the road network directly influences the difficulty of creating a satisfactory choice set, the performance of the reviewed methods in different studies with various networks are included. This review will help us to develop the new choice set generation approach and provides benchmarks for assessing proposed approaches. We review both various logit models and machine learning models in route choice modelling.

2.1 Choice Set Generation

Many studies focus on methods for creating choice sets (Azevedo et al. 1993; Ben-Akiva et al. 1984; Bovy and Hoogendoorn-Lanser 2005; De La Barra et al. 1993; Frejinger et al. 2009), and choice set generation approaches can be categorized based on the way they produce

potential routes (Prato 2009a). The measures all build on variations of the idea of the best path through the network, and are summarised below:

(1) K-shortest path: Calculate the K best paths based on a generalised cost of links. Some algorithms allow cycles in the path (Bellman and Kalaba 1960; Eppstein 1998), and some others focus on cyclic paths (Yen 1971; Hadjiconstantinou and Christofides 1999).

(2) Link elimination (Azevedo et al. 1993): The algorithm extends the K-shortest path algorithm. When searching the $K + 1$ shortest path (normally measured in distance or time), some links or all links in previous K shortest paths will be eliminated from the network.

(3) Link penalty (De La Barra et al. 1993): Similar to link elimination, but instead of removing links from the original network, this method adds a fixed penalty factor to links which are included in the previous shortest paths before re-estimating the shortest path.

(4) Labelling (Ben-Akiva et al. 1984): Labelling differs from the previous methods. It defines a target label such as 'Minimise traffic lights' before calculating a path, and then searches for routes that optimise the target. For example, 'Minimise traffic lights' means finding the path with the fewest traffic lights.

(5) Constrained enumeration: Unlike the methods listed above which assume travelers deciding based on minimum generalised costs, constrained enumeration assumes people selecting routes based on behavioural rules. For instance, a branch and bound algorithm as introduced by Prato and Bekhor (2006), uses a branching rule that reflects behavioral assumptions through the definition of thresholds. A directional threshold and a loop threshold removes routes with either high overlap with existing alternative routes or very long travel time.

(6) Probabilistic methods: Probabilistic methods calculate the probability of links based on the distance of them to the shortest path (Frejinger and Bierlaire 2007). For an OD pair, at each way point, a repeated random walk process adds the probable next link based on similarity to the shortest path. The route probability then equals the product of the probability of each link comprising the route.

(7) Doubly stochastic generation function (Bovy and Fiorenzo-Catalano 2007): This assumes that travelers have a perceived cost with error for paths, and the generation function includes a stochastic terms for cost and to account for the heterogeneity of travelers. These random terms are assumed to follow a probability distribution.

TABLE 2.1: Summary of the choice set generation techniques observed in the literature.

Methodology	Study	# of Nodes	# of Links	travelers	Generated routes	Capture rate (Overlap = 80%)
Link elimination	Bekhor et al. (2006)	13,000	34,000	188	30	71%
	Frejinger and Bierlaire (2007)	3,077	7,459	2,978	15	80%
	Prato and Bekhor (2007)	419	1,427	236	10	70%
	Pillat et al. (2011)	7,703	22,620	1,089	1-13	60%
	Rieser-Schüssler et al. (2013)	408,636	882,120	500	100	75%
	Ding et al. (2014)	7,808	11,106	997	30	79%
	Zhu and Levinson (2015)	8,618	22,477	143	29-58	25%
	Yao and Bekhor (2020)	8,583	21,151	6000	24	83%
Link penalty	Bekhor et al. (2006)	13,000	34,000	188	40	80%
	Prato and Bekhor (2007)	419	1,427	236	15	62%
	Zhu and Levinson (2015)	8,618	22,477	143	15	55%
	Yao and Bekhor (2020)	8,583	21,151	6000	27	96%
Labelling	Bekhor et al. (2006)	13,000	34,000	188	1	46%
	Prato and Bekhor (2007)	419	1,427	236	1	31%
	Spissu et al. (2011)	no infomation	18,000	393	1	47%
	Quattrone and Vitetta (2011)	4,480	16,029	332	5	75%
	Zhu and Levinson (2015)	8,618	22,477	143	1	23%
	Tang and Levinson (2018)	8,618	22,477	124	1	28%
	Yao and Bekhor (2020)	8,583	21,151	6000	1	59%

Directly using the K-shortest path approach for the route choice problem is uncommon in recent research. The high similarity (overlapping) of generated routes means the alternatives cannot be easily distinguished as a different routes by travelers. Instead, link elimination, link penalty, and labelling approaches are commonly applied in literature. As shown in Table 2.1, these methods generally provide high capture rates for a fixed overlap threshold (80%), which indicates that a large number of trajectories at least share 80% of links with the generated route in the choice set.

Prato (2009a) noted that some links which are included in both actual trajectories and alternative routes might be removed before searching new shortest paths in link elimination, and thus, the real trajectories cannot be captured. Link penalty keeps those 'used' links in the network, and the new shortest path still has the chance to use them. Therefore, link penalty approaches are more likely to generate real paths. For labelling, choosing good labels can make the generation process efficient, but the label selection process relies on the modelers' experience, and not all real trips have clear labels. For constrained enumeration, the difficulty is setting the thresholds for behavioural constraints.

Bekhor and Prato (2009) argued that the empirical results in a small network can differ greatly from a large urban network. One shortcoming of probabilistic methods is the unrealistic loops created by repeating the random walk process.

All the methods above generate alternative routes based on predefined rules and are further categorized to be explicit methods by Yao and Bekhor (2020). Compared with explicit methods, implicit methods do not need alternative routes to be defined before model estimation (Fosgerau et al. 2013). However, implicit methods have high computational costs and are unsolvable when there are cycles in the paths. More detailed description about the implicit methods are introduced in section 2.2 as link-based models.

Overall, using a K-shortest path algorithm to generate alternative routes is easy to implement but is unlikely to capture all observed trajectories. Similar shortcomings can be found in constrained enumeration methods. For stochastic-shortest path based methods, the high time cost and high reliance on the implementation of suitable probability distributions mean it is unsuitable for many data sets. For instance, Prato (2009a) argued that the doubly stochastic generation function method, which assumes that not only travellers perceive path costs with error, but also different travellers have different perceptions, is computationally prohibitive for larger road networks. Based on previous studies, the majority of research focuses on improving link elimination, link penalty, and labelling approaches, and both the number of observed trips and the size of the network are small. In high-resolution networks, Rieser-Schüssler et al. (2013) covered approximately 75% observed trips with 80% overlap threshold but generated 100 alternative routes per OD pair. This can be improved.

2.2 Route Choice Modelling

The second step is a discrete choice problem. Both statistical logit models and data-driven machine learning approaches are reviewed. Moverover, even though this study focuses on the route-based route choice models, we still include the link-based route choice models to provide a more comprehensive review.

2.2.1 Logit Models

Unlike other discrete choice problems, in route choice links are often shared between alternatives, creating additional complexity. Therefore, the traditional Multinomial Logit (MNL) model is inappropriate for route choice problems, and various modified MNL models have been developed (Bliemer and Bovy 2008; Zhao and Liang 2023).

2.2.1.1 Route-Based Models

Cascetta et al. (1996) proposed the C-Logit model, which includes a *commonality* factor to account for the similarity between alternatives and is robust to the choice set size. However, the commonality factors only captures part of the similarity, and there are no standard rules for defining commonality factors.

The Path-size Logit model (PSL), proposed by Ben-Akiva and Bierlaire (1999), has a similar formulation to C-logit, introducing a path-size term to measure the overlap between alternative routes. It has been shown that the Path-size Logit model outperforms C-Logit in general (Ramming 2001; Prato and Bekhor 2006), but the path-size term only captures part of the correlation between alternative routes. Bovy et al. (2008) adopting an idea from random utility introduced the Path Size Correction Logit model, which replaced the original path-size term with a correction factor that expands the value range from (0, 1) to $(-\infty, 0)$, but attains similar results to the original PSL model on a grid network (Prato 2009b). Duncan et al. (2020) proposed a internally consistent Adaptive Path Size Logit (APSL) model to overcome the challenges of traditional PSL by making routes contribute to path size terms according to

the ratio of route choice probabilities, and found that APSL shows better log-likelihood than PSL and MNL.

Instead of adding a correction term in MNL to include the effect of similarity between alternative routes, Generalized Extreme Value (GEV) models (Prato 2009b) allow for a flexible specification of the error term distribution to incorporate the similarity effect from alternative routes in the choice set.

Prashker and Bekhor (2000) proposed a Paired Combinatorial Logit (PCL) formula for the Stochastic User Equilibrium (SUE) problem, in which no travelers can unilaterally switch routes to improve their perceived travel times, but due to its computational cost, this method is rarely applied to real networks. The upper level of the nesting structure for PCL consists of alternative pair comparisons across all possible route pairs, so the number of nests grows rapidly with the network size.

Prashker and Bekhor (2000) also formulated a Cross-Nested Logit (CNL) for the SUE problem. All links are assumed to have the same coefficient in CNL, and when the coefficient equals one, CNL simplifies to MNL. According to Bovy and Hoogendoorn-Lanser (2005), CNL does not perform better than modified MNL models, and CNL collapses to MNL for analysing traveler behaviour (Ramming 2001; Bovy and Hoogendoorn-Lanser 2005).

A generalised CNL model, Generalized-Nested Logit (GNL), was introduced by Bekhor and Prashker (2001). It shares some properties with the CNL model but increases the complexity of the formulation of the nest coefficient by introducing an additional parameter γ. However, the nest coefficient tends to 1, which means it simplifies to an MNL model in real case studies (Ramming 2001; Bovy and Hoogendoorn-Lanser 2005).

While logit models cannot perfectly solve the problem, the Path-size Logit model shows better performance and is easier to implement on large networks than other variants. Therefore, PSL will be included in this study.

2.2.1.2 Link-Based Models

The recursive logit model, introduced by Fosgerau et al. (2013), does not require forming a choice set before modeling. Unlike the models above, the recursive logit model considers route choices in a dynamic way based on links rather than routes, where a route is a sequence of links that connect the origin to the destination. At each node in the dynamic case, travelers optimise the utility of the subsequent links before moving, and the route choice is a sequence of link choices (Zimmermann and Frejinger 2020). Based on the recursive logit model, Mai et al. (2015) proposed a nested recursive logit model (NRL), which avoids the independence of irrelevant alternatives (IIA) assumption by enabling scale parameters to be link-specific, which results in improved model fit and prediction. These link-based models eliminate the requirement of a finite choice set in route choice modelling.

As argued by Lai et al. (2019), Sun and Park (2017) and Sobhani et al. (2018), traditional logit models do not perform well in predicting individuals' route choices on high-resolution networks. Therefore, approaches that advance in prediction accuracy and handle big data are needed.

2.2.2 Machine Learning Approaches

Unlike statistical models, which are at least partially theory-driven, machine learning models are largely data-driven. This gives them more flexibility and greater predictive accuracy, but lower interpretability. Based on the way route-based and link-based models solve route choice, the machine learning techniques they incorporate differ. Since route-based models consider complete routes as the decision units and capture the holistic decision-making process, supervised machine learning approaches attract more attention than the other machine learning methods. For link-based models, emphasizing local decisions at each network node makes reinforcement learning techniques more popular.

2.2.2.1 Route-Based Models

Machine learning approaches have attracted attention in discrete choice modelling (Van Cranenburgh et al. 2022), especially in transport mode choice (Zhang and Xie 2008; Wang et al. 2020; Cantarella and Luca 2005; Hillel et al. 2021). Improved accuracy is demonstrated by many studies (Lee et al. 2018; Darwish et al. 2024; Zhao et al. 2020). For route choice modelling, Yamamoto et al. (2002) applied a Decision Tree to predict driving route choice from two alternative routes on expressway networks, and the Decision Tree yielded more correct predictions than the logit model. Considering the small size of choice set and simplicity of the test network, this early approach is criticised for not being robust (Lai et al. 2019).

Random forest, as an 'upgraded' version of Decision Tree, has captured more attention. Tribby et al. (2017) used a Random Forest to model pedestrian route choice and showed that using a Random Forest to select variables significantly improves the model's goodness of fit. For modeling train route schedulers' choice behavior, Schmid et al. (2022) tested Random Forest, Multinomial logit model, Mixed logit model, Support Vector Machine, and Artificial Neural Network, and found that Random Forest yields the best accuracy.

Apart from tree-based models, other machine learning has been used to model transport choices. Support Vector Machine (SVM) has been applied to drivers' route choice in a lab-based simulator (Sun and Park 2017). The results showed that SVM has accuracy similar to Neural Network (NN), another powerful machine learning algorithm, but with better time efficiency. Artificial Neural Network (ANN) is a class of machine learning algorithms that consist of an input layer, an output layer, and one or more hidden layers. Politis et al. (2023) showed Artificial Neural Networks have better performance than logit in predicting drivers' choices. Lai et al. (2019) tested both statistical and machine learning models with different choice set sizes for modeling taxi drivers' route choice and also found better performance with RF models. In the broader machine learning space, ANNs demonstrate high accuracy when dealing with big data and have therefore attracted the attention of route choice modellers (Lai et al. 2019).

All the machine learning models above will be included in this study.

2.2.2.2 Link-Based Models

For link-based route choice modelling, reinforcement learning approaches with complex structures and various architectural designs are also attracting increasing attention. Zhao and Liang (2023) proposed a deep inverse reinforcement learning (IRL) framework to solve the route choice problem in a general way. Similar to the recursive logit and nested recursive logit models, the IRL approach is also based on link choice rather than route choice. The goal of reinforcement learning is to learn a process to obtain the behaviour that optimises the predefined reward function, and inverse reinforcement learning aims to recover the rewarded function from the data to explain the choice behaviour of travelers. The reward function includes some features and parameters associated with the features and is analogous to the utility function in logit models. A proposed context-dependent adversarial IRL (AIRL) approach shows better performance than conventional PSL and recursive logit models (Zhao and Liang 2023). Liu et al. (2023) combined ensemble techniques and inverse reinforcement learning to form an AdaBoost-Bagging deep inverse reinforcement learning for autonomous taxi cruising route choice, and the proposed model also shows better performance.

Even though link-based and route-based modelling approaches solve the problem from different view, they both show that using machine learning techniques can outperform traditional methods.

2.2.3 Bounded Rational Route Choice Models

The route-based and link-based logit models above are based on perfect rationality, which assumes the path with the highest utility or the least generalized cost will be selected by travelers. As opposed to perfect rationality, Simon (1957) proposed that people are bounded rational when making route choice decisions, which means ‘a good enough path’ is accepted by the traveler. According to Di and Liu (2016), studies on bounded rational route choice model can be categorized as based on substantive bounded rationality and procedural bounded rationality. The former uses game-theory approaches and focuses on equilibrium link flows in a road network for planning purposes (Lou et al. 2010; Sheffi 1985), while the later aims

to predict individuals' decision-making results and estimate bounded rationality parameters (Gao et al. 2011; Di et al. 2017). A comprehensive review of both substantive and procedural bounded rational route choice models can be found in the work of Di and Liu (2016). Machine learning approaches are data driven rather than are grounded in economic theory or behavior theory.

2.3 Ensembles

In general, ensemble learning includes generation and combination of multiple base models to solve the learning task. The ensemble approach is well-known in machine learning, but it does not have a fixed form or algorithm. As described in chapter 1 section, including all the results from each individual model, rather than abandoning them, can provide an even more accurate result than the single 'best' model. Using ensemble techniques not only involves individual models but also leverages the benefits of GPS trajectory data to provide a more accurate prediction. As shown in Figure 2.1, ensembles could be seen as a two-stage process, which first trains each base model to gain a set of results and then aggregates those results based on ensemble strategies to provide a final prediction. Based on the type of base models (Zhou 2021), ensembles could be categorised into:

(1) *Homogeneous ensembles* comprising base models of the same type. For example, all base models in a Decision Tree ensemble will be Decision Trees
(2) *Heterogeneous ensembles* containing base models and algorithms of different types. For example, a combination of the predictions from a Decision Tree, a Support Vector Machine and a Neural Network.

The following describes the common ways ensembles are used in machine learning, and reviews the application of ensemble techniques in transport.

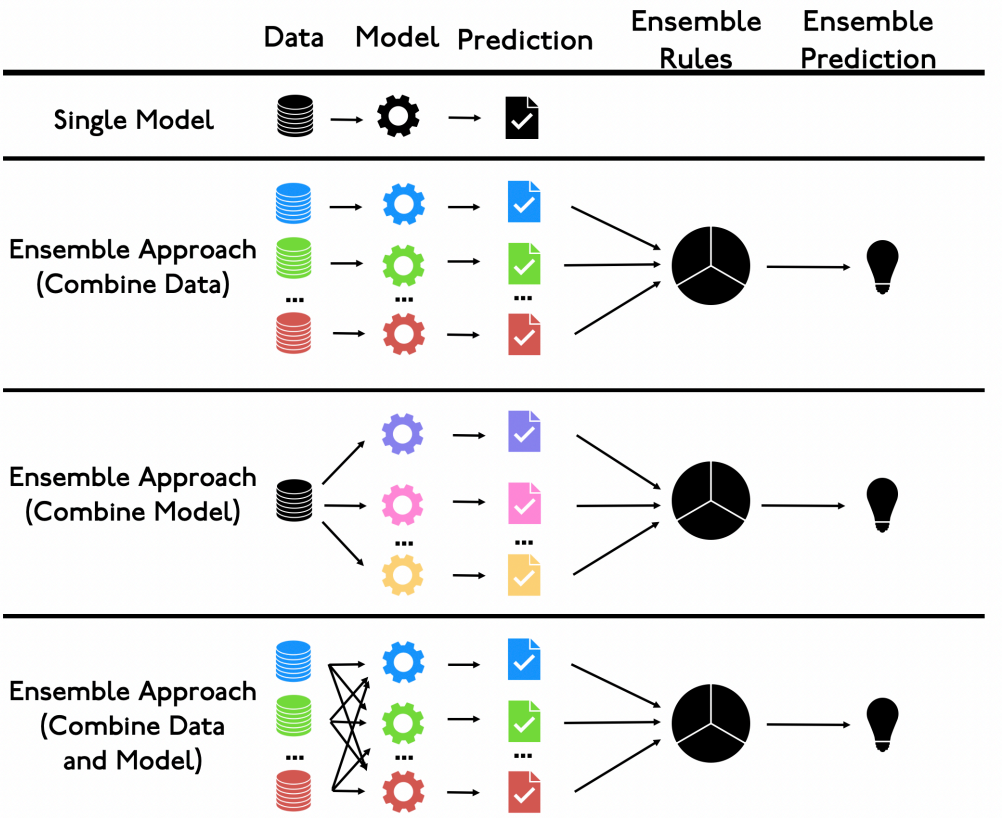


FIGURE 2.1: Ensemble Types

2.3.1 Bagging

Bagging is a popular technique in homogeneous ensembles. As shown in Figure 2.2, for bagging (bootstrap aggregation), each base model is trained with a sample that is randomly taken from the original data set with replacement (Breiman 1996a). Those samples are constituted differently for learners but have the same size. Since the sample is taken with replacement, some data in the original data set may be included in different samples. The main advantage of bagging is decreasing variance without rising bias of the predictions and at achieving more accurate point predictions the same time (Breiman 1996a; Petropoulos et al. 2018). As each base model is trained independently, the parallel structure reduces the computational cost. For classification tasks, bagging adopts hard voting to combine the (discrete) results from each learner.

Random Forest (RF) is an extension of bagging. The base models in RF are Decision Trees, and similar to bagging, each individual tree is trained based on a stochastic sample from an original training set, but each tree incorporates a random subset of all features (explanatory variables). Results from each tree are then combined based on a hard voting rule. In our study, RF is treated as both a homogeneous ensemble and a base model.

2.3.2 Boosting

Boosting, which is another well-known technique in homogeneous ensembles, is a group of algorithms that convert base models into strong learners. As shown in Figure 2.2, the idea behind boosting is that one base model is trained first, and then the distribution of the sample is adjusted based on the results of the first base model before training the next learner (Freund and Schapire 1997). This adjustment allows subsequent learners to pay more attention to previously misclassified samples. Boosting improves prediction accuracy for weak base models (Freund and Schapire 1997; Ferreira and Figueiredo 2012). There are various approaches to performing boosting, such as AdaBoost and XGboost. The techniques differ, but their goal is the same. Each base model's training depends on the previous learner, and thus, the model estimation time is normally higher than bagging.

2.3.3 Stacking

Stacking, as shown in Figure 2.2, is one of the various ways to form heterogeneous ensembles and is characterized by its two-level learner structure. This differs from bagging and boosting, in which results from base models are combined based on ensemble rules such as hard voting to make an ensemble prediction. Stacking takes outputs from base models (which may be the same or different types of models) as inputs of the meta-learner to provide final results. In machine learning approaches, data are normally split into training and testing sets. In stacking, the training set is further split into two sub-training sets. The first sub-training set is used to train base models (first-level learners). Results gained from first-level learners are then taken as input to the second-level learner. In order to reduce the risk of over-fitting,

the second-level learner is trained based on the second sub-training set, and the two-level learner ensemble is tested on the test set. The base models (first-level learners) in stacking can be either homogeneous or heterogeneous, while the second-level learner could be one of the models in first-layer learners or other kinds of models. Graczyk et al. (2010) shows that using stacking technique can obtain more accurate results, but the performance can drop significantly on different cases.

Ensemble techniques like bagging, boosting, and stacking are all included in this study. For a more detailed review of ensembles, see Wu and Levinson (2021)

2.3.4 Ensembles in Transport

Ensembles have gained more and more attention in transport field (Wu and Levinson 2021; Wu and Levinson 2022; Ji and Levinson 2020; Wu et al. 2023; Cui et al. 2022). Rasouli and Timmermans (2014) and Cheng et al. (2019) used tree-based ensembles for transport mode choice prediction, Cheng et al. (2020) introduced an ensemble based on Multinomial logit model and boosting and random feature selection technique for modelling travelers' choice behaviour in travel demand, and Sahoo et al. (2022) tested heterogeneous ensembles using bagging and stacking with four different models to predict the delay in air cargo transport. All of them show ensembles have advantages in prediction accuracy measure. However, ensembles that combine different types of models (heterogeneous ensembles) have not previously been applied to predict route choice, and different ways to combine results from individual models have not been evaluated to model commuters' route choice.

Therefore, this study aims to take advantage of the ensemble approach to predict travelers' route choices. This study tests only path-based route choice models, including Multinomial Logit, Path-size Logit, and supervised machine learning approaches, such as Decision Trees, Random Forests, AdaBoost, Support Vector Machines, and artificial Neural Networks.

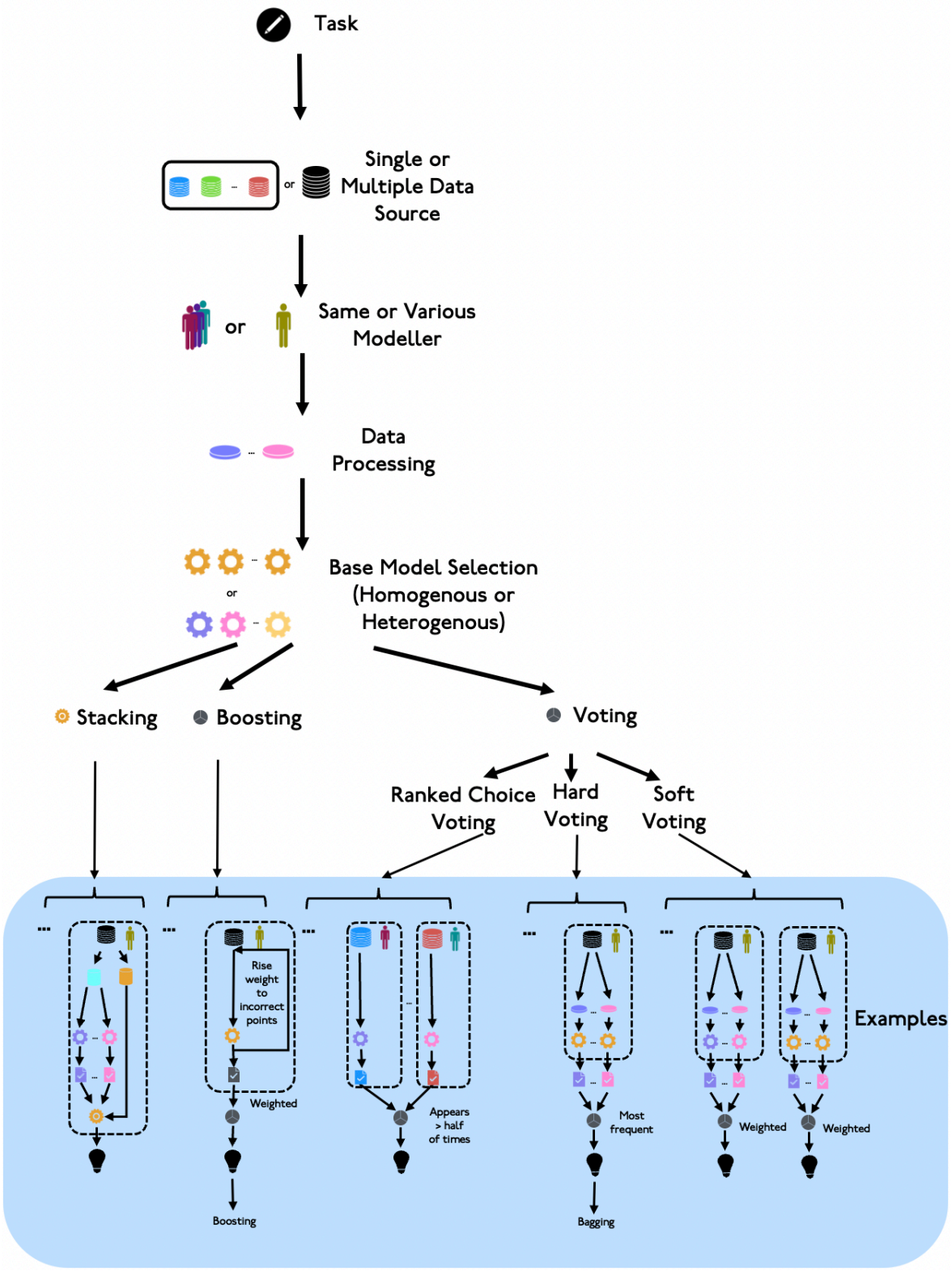


FIGURE 2.2: Schematic of Ensemble Techniques

CHAPTER 3

Data

This chapter introduces the five data sets used in this study. In section 3.2, we include descriptions of various travel speeds from different sources and two extreme scenarios regarding travel speed. These speed data are used to calculate travel times on each link. section 3.1 shows the road network of the Twin Cities. In section 3.3 and section 3.4, we present two data sets that are used for training and testing route choice models. section 3.5 and section 3.6 describe two data sets that are used for validating the route choice models trained based on section 3.3 and section 3.4.

3.1 Network

As shown in Figure 3.1, The Lawrence Group (TLG) road network (Craig 2005) for the Twin Cities is used in this study, and it includes 108,561 nodes and 277,747 links.

3.2 Travel Speed Data

Three sources of real-time speed data are applied in this study.

The first is TomTom speed data which was gained by aggregating thousands of GPS logging and navigation devices from the Twin Cities metro region, (Cohn 2009; Cui and Levinson 2018; Tang, Levinson et al. 2015). No travellers in the I-35W study used TomTom as their guidance, and the TomTom speed data was collected in a different year (2011). Additionally, this study focuses on morning trips, so morning peak hour (7-9 am) data on the TomTom network is matched to and overlaid on the TLG network. For a few roads in the TLG networks,

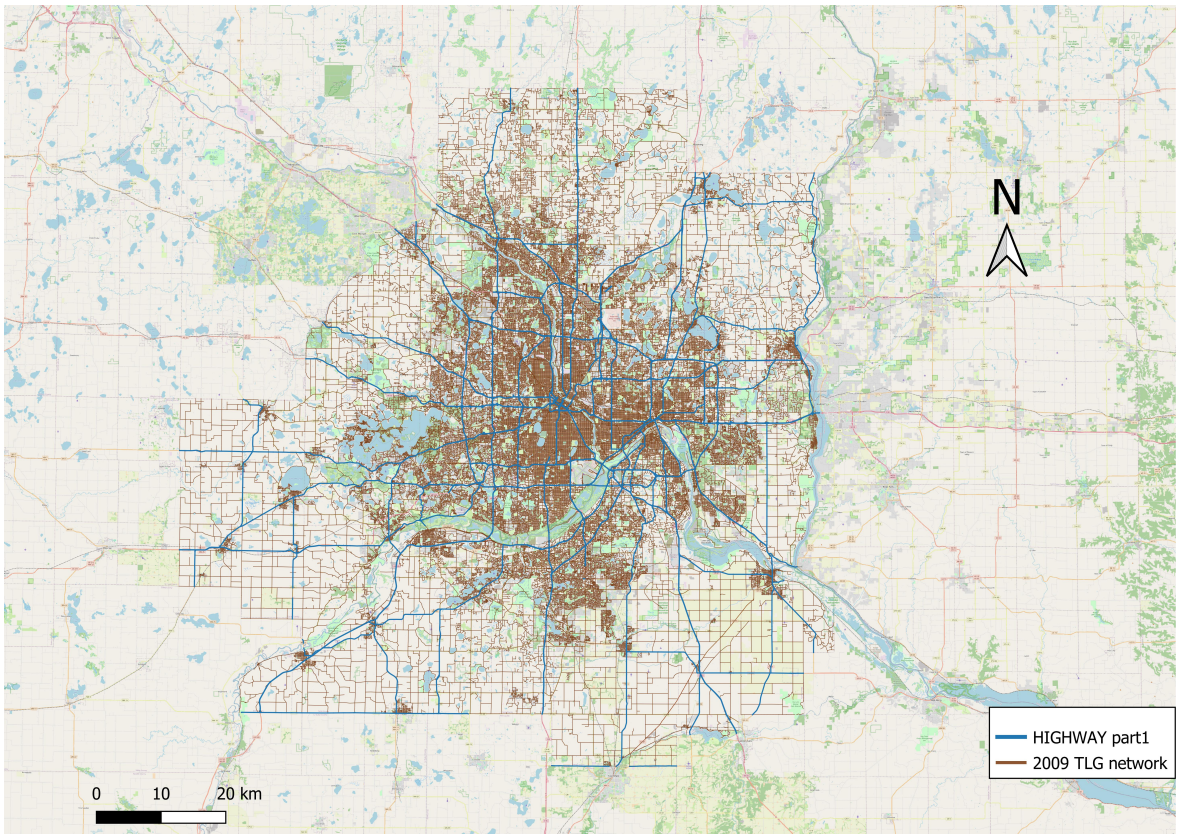


FIGURE 3.1: The Twin Cities TLG Road Network

the TomTom network splits those roads into several links and records speeds for those links separately. In this case, the speed on the longest link is used to represent the travel speed for the road. The TomTom data records the travel speed distribution on each link. For example, the 5th percentile travel speed represents the lowest 5% recorded speed on each link, and for every 5% from 5th to 95th percentile, TomTom data provide an aggregated travel speed. Two book-end scenarios are used in this study:

(1) Perfectly correlated scenario ($M1$): which assumes the travel speed on all links is drawn from the same percentile.
(2) Perfectly independent scenario ($M2$): which assumes travel speed on all links are independent. The percentile level for each link is randomly selected.

Additionally, travel speeds from loop detectors and the GPS speeds from all 153 participants in the I-35W Bridge Collapse study are also included in this study. Both have smaller sample

sizes compared with the TomTom speed data. We could have used the GPS data itself to generate a speed map and used the average speed of all links by link classification to represent the travel speed for unused links (Zhu 2010), but as the 2011 year TomTom data has recorded travel speed on most of the links and is available for us, we used it to represent the link travel speed in this study.

For all three sources of real-time travel speed data, not all links in the TLG network have speed records. The 2008 to 2009 street images in Google Maps are used to check those links, and we find most of the links are local streets. Based on that, we assume these are low traffic roads, and therefore, speeds are assumed to be 15 miles per hour (25 $km \cdot h^{-1}$).

3.3 2010 Travel Behavior Inventory GPS Trajectory Data (CS1)

The trip data in case study 1 (CS1) are obtained from the 2010 Travel Behavior Inventory (TBI) (Huang and Levinson 2015; Transport Publications 2013). The GPS trajectories were collected from travellers in Minneapolis - St. Paul (The Twin Cities) between 2010 and 2012. Each household was issued GPS units to carry for a 7-day period, but the recorded data does not contain information about transport modes. There were 3116 driving trips from 155 households are included in this study (Tang and Levinson 2018). The 2011 road network from the OpenStreetMap (OSM) includes 144,717 nodes and 374,358 links, is used.

3.4 I-35W Bridge Collapse Study (CS2)

The GPS information in case study 2 (CS2) were collected in the Minneapolis - St. Paul region (The Twin Cities) as part of the I-35W Bridge Collapse study in 2008 (Zhu and Levinson 2015; Zhu 2010). Within a 13-week period, 43,117 trips were recorded from 153 participants using either a logging GPS device (QSTARZ BT-Q1000p GPS Travel Recorder powered by DC output from in-vehicle cigarette lighter) or a real-time communicating GPS device (adapted from the system deployed in the Commute Atlanta study (Rates 2007)) installed in

TABLE 3.1: Data processing for CS2

Cleaning process	Travelers removed	Unique travelers	Unique Trips
Raw data	0	153	43117
Trips completed between 5am to 9 am	21	132	5990
Trip duration longer than 1 minute	1	131	4965
Trips forming a complete route	0	131	4839
Trips shorter than 150% shortest distance path	0	131	4538

the travellers' vehicles. In this study, an observed trip is defined as an observed journey from a single origin to a single destination for one traveller at a specific time. For example, for the same origin and destination, even if a traveller uses the same route on two different days, these two journeys are defined as two separate observed trips. In this study, only morning trips are considered. Trips lasting under 1 minute are assumed to be short journeys for parking and are removed. In a few cases where travellers detour 1.5 to 3 times the shortest distance to their workplace, these trips are assumed to pick up someone or have other purposes and are excluded. The filtering process and results are presented in Table 3.1.

3.5 Twin Cities 2010 LEHD Data (Validation Data 1)

The first is the 2010 Origin-Destination Employment Statistics from the Longitudinal Employer-Household Dynamics (LEHD), which includes home and workplace location at the census block level for 1,121,821 people in the Twin Cities. For the LEHD data set, the origins and destinations are assumed to be allocated at the centroid of the census blocks, and loaded on the network at the nearest node.

3.6 2010 TBI data Survey Data (Validation Data 2)

The second is the 2010 Travel Behavior Inventory (TBI) data in the Twin Cities, which includes travellers' origin and destination at the transport analysis zone (TAZ) level, transport modes, trip purpose, departure times, a weight of each OD pair and other information for approximately 1% of the population. The higher value of the weight means the more people that record represents. While the GPS data from the TBI is used for model estimation for Case Study 1, this is the non-GPS based survey data from a different set of subjects

CHAPTER 4

Methods

The proposed method, including choice set generation, route choice modeling, evaluation and validation, is shown in Figure 4.1.

The data sets used in this study were shown in chapter 3. As the data sets include numerous GPS trajectory data, section 4.1 details how GPS points are matched to road network links. A hybrid method combining the link penalty and labelling approaches is proposed in section 4.2.

After obtaining the choice set, section 4.3 introduces the key concepts of ensembles and how they are constructed in this study. Within that, statistical methods and machine learning approaches are used to model route choice, and these are included as base models for the ensembles in subsection 4.3.3. Ensemble rules are strategies to combine results from base models to form ensemble results, and detailed description of them is included in subsection 4.3.4. subsection 4.3.5 and subsection 4.3.6 discuss how homogeneous ensemble and heterogeneous ensemble are formed using the ensemble rules and base models.

After that, in order to evaluate the performance of the proposed models in predicting route choice behaviors, section 4.5 includes three criteria for assessing model performance on testing sets. section 4.7 uses the trained model to forecast Vehicle Kilometers Traveled (VKT) on freeways by aggregating all travelers' route choice results, and the results are compared with the observations from loop detectors.

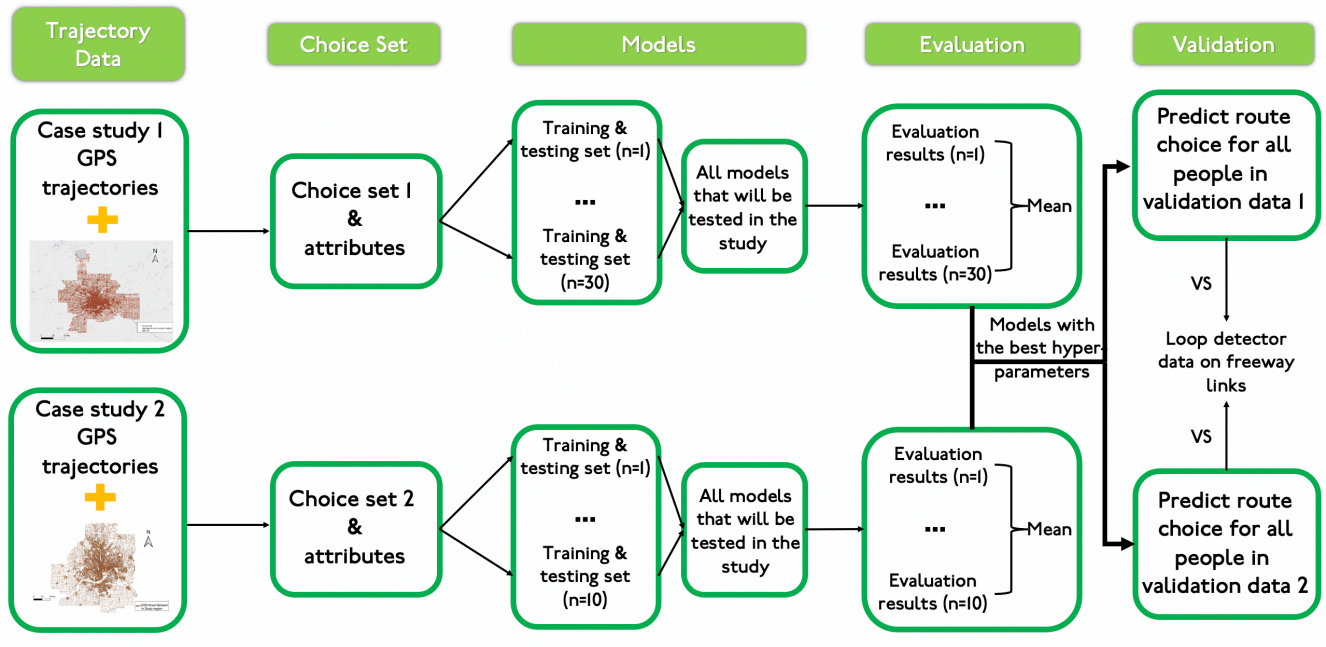


FIGURE 4.1: Flow Chart of Ensemble

4.1 Map Matching

The accuracy of recorded GPS trajectories can be affected by various aspects, such as the time gap between two records, the quality of the receivers, and the influence of dense, tall buildings and weather conditions. Case study 1 has dense and consistent GPS trajectories, and using the k-nearest neighbour approach to match all trajectory points to the road links in OpenStreet Map shows great results. However, the TLG road network in case study 2 does not match well with the trajectories when using the k-nearest neighbour approach. Instead, an algorithm was developed and applied to ensure that all matched links form a complete route, as shown in Figure 4.2.

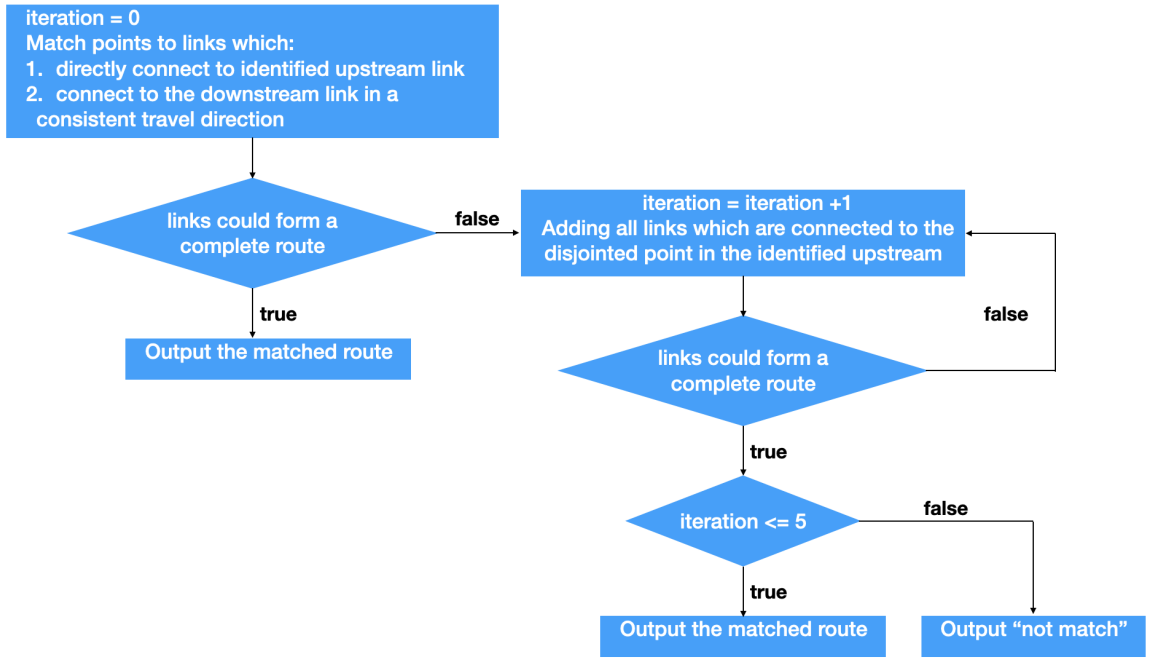


FIGURE 4.2: Map Matching Algorithm

In rare cases, multiple complete routes are formed by adding potential links connecting disjointed points upstream. In this case, the shortest one is selected as assuming people avoid local detours. For privacy, start and end recorded points are not real origins and destinations. In this case, a 100-meter tolerance zone is set at their reported origin and destination. Finally, all 4538 trips are successfully matched, and 1940 OD pairs are included.

4.2 Choice Set Generation

Choice set generation is an important step before modelling individual route choices, and quality and size significantly affect the modelling results. The high-resolution TLG road network provides an advantage in accuracy of the paths, but it also results in higher computational complexity in the route search process. As TomTom speed data records 19 travel speed percentage bins (every 5% from 5% to 95%), a simulation method is used to find alternative routes.

Overall, 19 draws based on 19 travel speed percentage bins defined in TomTom speed data are performed for the perfectly correlated scenario $M1$, and thus 19 travel times are obtained for each route alternative for a given trip's origin-destination pair.

For the perfectly independent scenario $M2$, since travel speeds of all links are independent, for each simulation, travel speeds of the links are randomly drawn from 5% speed to 95% speed. We performed 20 draws for the $M2$ scenario because of computational cost. However, as computational costs drop, more draws would be preferred, noting that we observed additional draws are subject to diminishing returns as alternative routes tend to repeat, and speed variance on many links is fairly small. Finally, 20 simulated travel times are gained for each link for scenario $M2$. Several packages can generate a network and find the shortest path, like NetworkX, SNAP and LightGraphs, and the each package has advantages in different aspects. Since evaluating those packages is not the main aim for this study, the most familiar package, NetworkX is used. In both cases, the A-star algorithm is applied in Python's NetworkX package to find the shortest path.

The labelling and link penalty approaches are combined to generate alternative routes with TomTom and free-flow speed data. The general steps are:

(1) Define a label.
(2) Determine a weight (penalty or bonus) factor and multiply the factor on all related links in the road network.
(3) Search the path which satisfies the predefined label in the first step for all OD pairs, such as 'minimise travel distance'.
(4) For some labels, we want multiple paths rather than a single path. Before finding the $k + 1^{th}$ path for these labels, we multiply a penalty factor to all links in the k^{th} path and update those links in the road network.
(5) Once all labels are completed, form a choice set with the generated alternative routes.

Labels used in this study are categorized into time-based paths and distance-based paths.

4.2.1 Time-based Labels

Time-based paths find alternative routes with the least travel time under different conditions. The measurable link attributes for these kind of labels are travel time. Freeways, which normally have faster travel speed, good road surface condition, and no traffic lights, are attractive for many travelers. However, as they have fewer intersections, travelling on freeways often results in a somewhat longer travel distance. To investigate the trade-off for the benefits and costs of freeway travel, a link penalty method is applied. A set of weighting factors (0.33, 0.6, 0.7, 0.8, 0.9, 0.95, 1.0, 1.05, 1.11, 1.25, 1.43, 1.67, 3.33) are multiplied by the link travel time on freeway links to identify the weighted fastest path. The same process is implemented for free-flow travel time, travel time under the perfectly correlated scenario ($M1$), and travel time under the perfectly independent scenario ($M2$). In addition, for every 5% from 5% to 95% travel speed in the perfectly correlated scenario, alternatives are generated based on all weighting factors in that list. Therefore, one route is found for each weighting factor (13 factors in total) for free-flow travel time, 19 routes are found for each weighting factor (13 factors in total) for $M1$, and 20 routes are found for each weighting factor (13 factors in total) for $M2$.

As the TLG network includes many links and nodes, the computational costs for generating k-paths are high. Therefore, 20 shortest time paths are found for each three factors (0.3, 0.8, and 1.0) with free-flow travel time. For scenario $M1$, 20 shortest time paths are searched for weighting factor 1 with 50th, 70th, and 90th percentile travel time. Since computational time for scenario $M2$ is high, and the travel time simulation process itself is like searching k-shortest paths, no additional k-shortest paths are generated for $M2$. The link weighting method is applied to searching k alternatives. For travel time based on detector speed and travel time based on GPS speed, only a factor equal to 1 (no preference for freeway) is applied.

4.2.2 Distance-based Labels

Distance-based paths find alternative routes with the least travel distance under different conditions. The measurable link attributes for these kind of labels are link length, and based

on the shortest distance path(s), some further searching process are made to optimize target labels.

(1) 'Shortest distance': One shortest distance path is found for each OD-pair.

(2) 'Minimum left turns': Since turns are actions of vehicles rather than properties of links, it hard to directly find the minimum left-turn path for a given OD pair by using NetworkX package in Python (Hagberg et al. 2008). Instead of finding the minimum left-turns path in the network, this study identifies the minimum left-turns path of the 20 shortest distance paths.

The link weighting approach is used to generate 20 shortest path in this study. For example, before finding the $k + 1^{th}$ shortest path, a weighting factors is multiplied to each link in the k^{th} shortest path, and the the $k + 1^{th}$ is found based on the updated network with those weighted links. Penalty factors of 1.05, 1.2 and 99 are used in this study.The reason for using 99 is that we find that even if we use the 1.2 weighting factor to exclude the link in the last shortest path, we still get the same path or a path with just small differences to the last shortest path. However, if we eliminate the link in the last shortest path, we might lose the links in the observed trips. Therefore, we used a very high weighting factor for the links in the last shortest path to obtain a very different path. Overall, as shown in Table 4.1, 47 time-based labels and 5 distance-based labels are included.

Labels Class	Weighting Factor	Number of Labels
Time-based labels		
Free-flow travel time scenario		
Freeway preferred	0.33, 0.6, 0.7, 0.8, 0.9 and 0.95	6
Freeway avoided	1.05, 1.11, 1.25, 1.43, 1.67 and 3.33	6
No preferences (Shortest time)	1.0	1
K-shortest path	0.3, 0.8 and 1.0	3
Perfectly correlated scenario (M1)		
Freeway preferred	0.33, 0.6, 0.7, 0.8, 0.9 and 0.95	6
Freeway avoided	1.05, 1.11, 1.25, 1.43, 1.67 and 3.33	6
No preferences (Shortest time)	1.0	1
K-shortest path with 50% travel time	1.0	1
K-shortest path with 70% travel time	1.0	1
K-shortest path with 90% travel time	1.0	1
Perfectly independent scenario (M2)		
Freeway preferred	0.33, 0.6, 0.7, 0.8, 0.9 and 0.95	6
Freeway avoided	1.05, 1.11, 1.25, 1.43, 1.67 and 3.33	6
No preferences (Shortest time)	1.0	1
Other time-based labels		
Shortest time with loop detector speed data	1.0	1
Shortest time with GPS speed data	1.0	1
Distance-based labels		
Shortest distance	1.0	1
Minimum left turns	1.0	1
K-shortest path	1.05, 1.2 and 99	3
Total number of labels		52

TABLE 4.1: Summary of all tested labels in choice set generation

4.2.3 Choice Set Evaluation

4.2.3.1 Overlap

The common choice set performance indicators are overlap rate, and capture rate. Overlap rate (Ω) measures the percentage by which generated routes overlap the observed trajectory, as shown in Equation 4.1.

$$\Omega_{i,n} = L^c_{i,n}/L_n \tag{4.1}$$

Where:

- $\Omega_{i,n}$ is the share of generated routes i overlapping observed trip n
- L^c_i is the total length of all common links in generated route i and observed trip n.
- L_n is the total length of the observed trajectory.

For an observed trip, the overlap rate between each generated alternative route and the actual route is calculated. The same process is repeated for all observed trips

Capture rate (c) describes the percentage of the set of observed trajectories that are captured by the generated routes under a threshold, as shown in Equation 4.2. For an observed trip, if any alternative route in the choice set has an overlap rate greater than the threshold, then this observed trip is defined as 'captured'. For example, capture rate is 43% for 80% overlap threshold means that more than 80% of the length of trajectories spatially coincide with the selected algorithm for 43% of all observed trajectories in the sample set. Both overlap rate and capture rate are used to show the performance of the choice set generation algorithms. Overlap thresholds of 50%, 60%, 70%, 80%, 90%, and 100% are used to measure capture rates.

$$c = \frac{N_c}{N} \times 100 \tag{4.2}$$

Where:

- c percentage of the set of observed trajectories that are captured by the generated routes under a threshold.
- N_c is the number of observed trajectories that satisfy the threshold.
- N is the total number of observed trajectory.

4.2.3.2 Deviation

Since on a high-resolution road network, there exist many alternative routes which slightly deviate from target routes (observed trip) (Rieser-Schüssler et al. 2013), some generated routes which have 0 overlap with observed trips, might only be one block away from the used route. The rationale for those small deviations varies with factors such as traffic light phase, influence of temporary road conditions, a gap in the traffic, and preferences which could not be identified from the collected information. Compared to routes which are far from observed routes, these 'closer alternative routes' are also important (Wang et al. 2022), so only using overlap rate to evaluate choice sets might be insufficient. Therefore, a term called 'average deviation' ($d_{i,n}$) is calculated based on Equation 4.3, and applied in this study.

$$d_{i,n} = \Lambda_{i,n}/L_n \tag{4.3}$$

Where:

- $d_{i,n}$ is the average deviation between observed trip n and the generated route i. Unit is meter.
- Λ is the area of the region between the generated route i and observed trip n. Unit is square meters.
- L_n is the total length of the observed trajectory. Unit is meter.

4.2.4 Explanatory Variables for Linear Models

After obtaining a choice set with good performance, in order to understand and predict route choices, this study evaluate which variables explain route choice. Both linear models and logit

Variable	Description
Trip Length[1,2]	The length of the alternative route
Traffic Lights[1,2]	The number of traffic lights passed along the trip
Traffic lights coverage[1]	Length of link with traffic lights divide by trip length
Bus Stops[1,2]	The number of bus stops passed along the trip
Bus stops coverage[1]	Length of link with bus stops divide by trip length
Minimum Legal Travel Time[1,2]	Travel time based on speed limit
Maximum Legal Speed[1,2]	Trip length divided by free-flow time
Freeway Percentage[1,2]	Equals total freeway length divided by trip length, as shown in Equation 4.4.
Left Turns[1,2]	The number of left turns along the trip
Right Turns[1,2]	The number of right turns along the trip
Path Size[1,2]	Measures the effect of overlap between alternative routes. Path size Z is between 0 and 1, where 1 indicates that the two routes are the same, as shown in Equation 4.5.
Age[2]	The age of the traveler
Gender[2]	The gender of the traveler {male, female}
Income[2]	The income of the traveler in US dollars

TABLE 4.2: Factors affecting trip characteristics and route choice

models are used in this study. We model the overlap and route choices as being determined by a set of independent variables ([1]) in Table 4.2.

[1]: explanatory variables for estimation

[2]: input variables for prediction

The freeway percentage is calculated based on Equation 4.4.

$$F_i = \frac{\sum_a^A f_a l_a}{\sum_a^A l_a} \tag{4.4}$$

Where:

- F_i is the reeway percentage for alternative route i.
- f_a equals to 1 if link a is a freeway link.

The Path-size term (Z) is calculated based on Equation 4.5.

$$Z_{i,C} = \sum_{a \in i} \frac{l_a}{L_i} \frac{1}{\sum_j \delta_{a,j}} \tag{4.5}$$

Where:

- $Z_{i,C}$ is the path-size term for alternative route i in choice set C.
- l_a is the length of link a in alternative route i.
- L_i is the total length of the alternative route i.
- $\delta_{a,j}$ is 1 if alternative route j includes link a, and 0 for otherwise.
- $\sum_j \delta_{a,j}$ is the number of alternative routes contain link a.

If the angle between the two connected road segments is between 30 and 150 degrees, the movement through these two segments is defined as a turn.

4.2.5 Linear Overlap and Deviation Models

In this study, a linear model which predicts the performance (overlap or deviation) of the alternative route is introduced. As GPS trajectories are collected from the same group of participants for a period of time, panel regression models are applied. All attributes shown in the Explanatory Variables section are taken as independent variables, and dependent variables including overlap rate $\Omega_{i,n}$ and average deviation $D_{i,n}$ are modelled.

For both $\Omega_{i,n}$ and $D_{i,n}$, the Breusch-Pagan test is adopted to test the heteroskedasticity, and the Durbin-Watson test is used for checking auto-correlation. The Hausman test is used to test for endogeneity in the panel data. Hypotheses in this study include:

(1) For overlap: alternative routes with shorter trip lengths, less travel time, fast travel speeds, higher percentage with freeway, and fewer turns are more likely to have higher overlap with the observed trip.
(2) For deviation: alternative routes with long trip lengths, longer travel time, lower travel speeds, lower percentage with freeway, and more turns are more likely to have higher deviation with the observed trip.

A similar study focusing on cyclists is included in the chapter C.

4.3 Prediction

4.3.1 Input Variables

Finally, for each origin-destination (OD) pair, on average, 40 unique alternative routes from the same best ten labels comprise the choice set. The route attributes and socio-demographic data are included as the input values for models, and detailed descriptions for them are presented in Table 4.2 with a superscript [2].

4.3.2 Standardisation

Unlike numerical features, some machine learning methods are not good at processing categorical attributes like gender, and thus, transforming the categorical attributes to numerical value is required. One-hot encoding converts the textual class to a numerical matrix.

The scale and spread of numerical variables will influence their importance in data-driven approaches. A standardisation technique that adjusts the attributes with the distribution value between 0 and 1 is applied before modelling in order to avoid inappropriate weighting of the variables.

After encoding and standardisation, the 4538 trips from 131 travellers are separated into training sets and testing sets. Since multiple trips are collected from one traveller, the day one trip can be exactly the same as the day n trip where n is any other day during the recording period. The testing set will include data from the same person as the training set if the sample is directly divided based on trips, and models will outperform in that situation. Therefore, this study uses the routes from 80% of travellers to form the training set and the routes from the remaining 20% as a testing set. Meanwhile, the total number of routes in the training set is checked to be roughly 80% of the total number of trips. Since the number of routes in the training set is not 100% identical because the number of recorded trips for each traveller are varied, the training and testing samples are randomly split 30 times for CS1 and 10 times for CS2.

4.3.3 Base Models

All base models share the same attributes of the alternative routes for each OD pair. These inputs are used according to each base model's approach and hyperparameter setting. The model output includes the chosen route as well as the probability of choosing each alternative route for a given OD pair. For all statistical and machine learning approaches described below, each model is trained with the same training set, and predictions for the route choice are assessed against the same testing set. The Python package Scikit-Learn is used for training and testing all base models.

The hyperparameters in machine learning models are the parameters that are set before the learning process, and they influence both the training process of the model and the model's performance in testing data. In this study, for all ML methods, the grid search technique with five cross-validations is applied to find the local best hyperparameters. The same hyperparameters are used when predicting with base models alone and ensembles. For each base model, the tested hyperparameters and the suggested value are recorded in Table 4.3.

4.3.3.1 Choice Models

Multinomial Logit (MNL) and Path-size Logit (PSL). As the currently dominant approaches, conventional MNL and PSL (Ben-Akiva and Bierlaire 1999) models are an important benchmark for route choice modelling. Both MNL and PSL models are based on the principles of random utility theory, and the utility function is shown in Equation 4.6. Travellers are assumed to make choices based on the utilities (U) they associate with each alternative route, and the one maximizing their overall utility is the chosen route.

$$U_{n,i} = V_{n,i} + \epsilon_{n,i} = \boldsymbol{\beta} X_{n,i} + \epsilon_{n,i} \tag{4.6}$$

Where:

- $U_{n,i}$ is the utility of alternative i for traveller n
- $V_{n,i}$ is the deterministic part in the utility function
- $\boldsymbol{\beta}$ is the coefficient vector, where $\beta \sim f(\beta|\theta)$ for any distribution of f
- $\mathbf{X}$ is a vector which includes all variables in the subsection 4.3.1
- $\epsilon_{n,i}$ is independent and identically distributed noise following an extreme value distribution

Equation 4.7 calculates the probability of choosing an alternative routes for MNL and PSL.

$$P(i|C) = \int \frac{e^{Un,i}}{\sum_j e^{Un,j}} f(\beta|\theta) d\beta \tag{4.7}$$

Where:

- $P(i|C)$ is the probability of choosing alternative i from the choice set C
- $U_{n,i}$ is the utility of alternative i for traveller n
- $\sum_j e^{Un,j}$ is the sum of exponentiated utilities across all alternatives in the choice set C

Generated alternative routes in the choice set might overlap partially or completely with each other. The Path-size logit model includes a Path-size (Z) term, as shown in Ben-Akiva and Bierlaire (1999), to correct for the correlation between the routes in the deterministic part of the utility function (V) Equation 4.8, and the formula for Z is presented in Equation 4.5.

$$V_{n,i} = \boldsymbol{\beta} X_{n,i} + \beta_Z \ln Z_{n,i} \tag{4.8}$$

Where:

- $V_{n,i}$ is the deterministic part in the utility of alternative i for traveller n
- $\boldsymbol{\beta}$ is the coefficient vector
- $\mathbf{X_{n,i}}$ is a vector which includes all variables in the subsection 4.3.1 section
- $Z_{n,i}$ is the path-size term

Since observed trips are recorded for multiple participants, heterogeneity might exist across people. However, MNL and PSL models do not consider variation among travellers. Therefore, a Mixed Logit (MXL) model, which allows the coefficient $\boldsymbol{\beta}$ to be random in utility function, is applied to take heterogeneity into account. Moreover, as shown in Equation 4.9 and Equation 4.10, the Path-size introduced in Equation 4.5 is added as a attributes to the MXL, and the result of MXL with and without the Z term are compared. A Python package called 'pylogit' is applied to perform mixed logit modelling (Brathwaite and Walker 2018).

$$U_{n,i} = \beta_n X_{n,i} + \beta_n \ln(Z) + \epsilon_{n,i} \tag{4.9}$$

$$P(i|C) = \int \frac{e^{\beta X_{n,i} + \beta_n \ln(Z)}}{\sum_j e^{\beta X_{n,j} + \beta_n \ln(Z)}} f(\beta|\theta) d\beta \tag{4.10}$$

- $P(i|C)$ is probability of choosing alternative route i in choice set C.
- $f(\beta|\theta)$ is the density function of the random coefficients β

As the Mixed Logit model typically requires simulation-based methods (e.g., Monte Carlo integration) for estimation, which can be computationally expensive, especially with large

datasets or when included in ensemble methods that require multiple model estimations, it is not included in any of the ensembles discussed below. For the later paragraphs when we say all models in subsection 4.3.3, we refer to all models in subsection 4.3.3 except MXL.

Random MNL and Random PSL. Random MNL (RMNL) was first introduced by Cheng et al. (2020). Similar to the structure of the Random Forest, RMNL involves MNL as the base model and randomly samples the original data (bootstrap) and sub-features from all features. Samples and sub-features are taken with replacement. We propose a Random Path-size Logit model (RPSL), which uses the same two-level randomness as RMNL but implements the PSL model as the base model, as shown in Figure 4.3.

4.3.3.2 Trees

Decision Tree (DT). Decision tree has a tree-like structure to identify breakpoints in each attribute that help to distinguish route alternatives. Each node in a DT represents a 'test' on an attribute, each branch shows the possible outcomes of the test, and each leaf node represents a class label (decision taken after computing all attributes). Hyperparameters defining maximum depth, maximum features, and criteria, including gini, entropy, and log-loss, are identified using grid search. The selected values are included in Table 4.3.

In the context of predicting route choices for travellers, the nodes can be seen as the process of assessing the attributes of alternative routes, while the leaf nodes represent people choosing a specific alternative route if its classification rules are certified. For example, consider a choice set with three alternatives: the shortest route (A), the fastest route (B) and the route with minimum left-turns (C). The length, the free-flow travel time, and the number of left turns of each alternative are included in the Decision Tree. By training a DT, it is possible to learn, for example, that if the length of route A is greater than 20 km, the minimum travel time of route B is less than 20 mins, and route C contains fewer than 4 left turns, then there is a high probability of choosing route B.

A Decision Tree cannot directly provide the probability of choosing an alternative route in the same way as the logit model. Instead, Decision Trees use the proportion of training samples

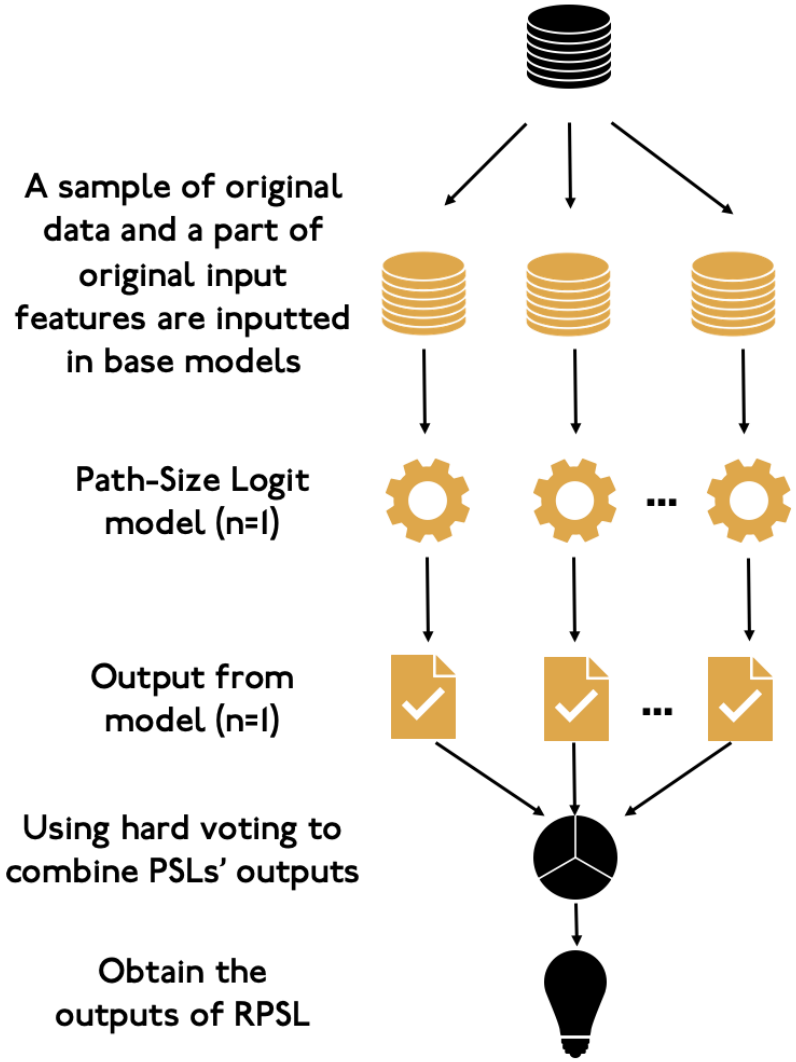


FIGURE 4.3: Structure of Random Path-size Logit Model

that belong to a particular class in each leaf node to represent the probability. This probability is later utilized in a heterogeneous ensemble with a soft voting technique.

Random Forest (RF). Random Forest (RF) is a homogeneous ensemble, but it is also treated as a base model in the heterogeneous ensemble in this study. The base model of RF is the Decision Tree model. Each tree in a Random Forest provides its own result, and RF combines all the results by hard voting. In a homogeneous ensemble, using the same

data to train each Decision Tree will yield the same result. Therefore, RF adopts two-level randomness to gain diversity:

(1) The training set of each Decision Tree has the same size and is randomly taken from origin data set with replacement. This is the bootstrapping approach.
(2) The set of explanatory variables which are used for each Decision Tree are randomly selected, but the number of variables of each tree remains the same.

Hyperparameters defining the numbers of estimators, maximum depth, maximum features, and criteria are found using grid search, and the selected values are listed in Table 4.3.

When using RF to predict the route choice for an OD pair, each DT in the RF will yield a route alternative, and the most frequently occurring route will be the final prediction of the RF model. For instance, if 8 out of 10 tree models in an RF model predict that a traveller chooses the shortest route (A), the final prediction of the RF model will be route A. For a given observation, the Random Forest aggregates the selected route probabilities from all the trees and takes the mean of them as the final probability for route predicted by the RF.

Extra Tree (ET). Extra Trees, short for Extremely Randomised Trees, shares common features with RF. They both have many DTs and use hard voting for classification. The main differences are: (1) RF uses bootstrap to take part of the sample with replacement, but ET use the whole sample; (2) When choosing the split point at each mode, RF uses the best cut point for the attribute while ET cuts it randomly. The output of ET is like Decision Tree models. Hyperparameters defining the numbers of estimators, maximum depth, maximum feature, and criteria are found using grid search. The used Extra Tree has a greater maximum depth than RF and DT, and includes fewer DTs than RF. The used values are shown in Table 4.3.

AdaBoost (AB). AdaBoost, short for Adaptive Boosting, uses boosting to assemble a group of base models, in this case decision tress, to produce a stronger model. The core principle of AdaBoost is to sequentially fit Decision Trees on repeatedly re-sampled versions of the data. Each sample carries a weight that is adjusted after each training step, such that misclassified samples will be assigned higher weights. The re-sampling process with replacement takes

into account the weights assigned to each sample. Samples with higher weights have a greater chance of being selected multiple times in the new data set, while samples with lower weights are less likely to be selected – this is boosting.

The model output includes predicted routes and their probabilities. Similar to Random Forest, the probability of the selected alternative is the weighted mean of the probabilities of the predicted route from all of the Decision Trees. Hyperparameters, defining the number of estimators and the learning rate, are found using grid search and illustrated in Table 4.3.

4.3.3.3 Support Vector Machine (SVM)

Support Vector Machines map the data into high dimensional space and find the hyperplane that most accurately divides the observations by observed route alternative. For multi-class classification tasks, Scikit-Learn provides a 'one-versus-one' approach, which makes $\frac{N_{\text{class}} \cdot (N_{\text{class}} - 1)}{2}$ binary classifiers. The input data and attributes are identical to the other models for fair comparison. The output from SVM is the predicted route alternative and the probability of choosing each alternative route in the choice set.

Hyperparameters defining regularisation and kernel size are found using grid search. When the kernel functions are 'RBF', 'Poly', and 'Sigmoid', the gamma value, which implies the influence of a single training example, is also tested. However, since 'Linear' is ultimately adopted as the kernel function, the gamma value is not relevant. Table 4.3 shows the selected value of these hyperparameters.

4.3.3.4 Neural Networks (NN)

Neural networks, inspired by the structure and function of the human brain, consist of interconnected nodes called neurons, organized into layers. Each neuron receives input, processes it through an activation function, and then passes the output to the next layer. In this study, initially, one-hidden-layer NN, three-hidden-layer NNs and ten-hidden-layer NNs are tested. The difference in performance is not very significant. Therefore, to save computational cost, only one hidden layer NN, which is also called Multi-Layer Perceptron (MLP), is

included in the remaining tests. Similar to SVM, MLP outputs the predicted alternative route ID and the probability of selecting each alternative route. An activation function determines the output of each neuron in the Neural Network. The solver is an optimisation algorithm used to update the weights of the Neural Network during training. Batch size refers to the number of training examples used in each iteration of training, and it determines how many samples are processed before updating the weights of the Neural Network. The learning rate

controls the step size taken during optimisation and influences the speed of convergence. All of these four hyperparameters are involved in the grid search.

Base Model	Hyperparameter	Tested settings	Selected Value
Multinomial Logit Model	Not applicable	Path Size Z not Included	
Path Size Logit Model	Not applicable	Path Size Z Included	
Random Multinomial Logit Model	The number of Estimators	Every 50 from 100 to 500	200
	The Number of Samples to Draw	Every 50 from 100 to 500	100
	The Number of Features to Draw	Every 50 from 100 to 800	140
	Samples Are Drawn with Replacement	True	True
	Features Are Drawn with Replacement	True	True
Random Path Size Logit Model	The Number of Estimators	Every 50 from 100 to 500	200
	The Number of Samples to Draw	Every 50 from 100 to 500	100
	The Number of Features to Draw	Every 50 from 100 to 800	140
	Samples Are Drawn with Replacement	True	True
	Features Are Drawn with Replacement	True	True
Decision Tree	Maximum Depth of The Tree	from 5 to 50	11
	Criterion	Gini, Entropy, and Log Loss	Gini
	Maximum Feature	“Log2” and “Sqrt”	“Sqrt”
Extra Tree	The Number of Estimators	Every 50 from 100 to 500	100
	Maximum Depth of The Tree	from 5 to 50	25
	Criterion	Gini, Entropy, and Log Loss	Gini
	Maximum Feature	"Log2" and "Sqrt"	“Sqrt”
Random Forest	The Number of Estimators	Every 100 from 1000 to 3000	1500
	Maximum Depth of Trees	from 5 to 50	20
	Criterion	Gini, Entropy, and Log Loss	Gini
	Maximum Feature	“Log2” and “Sqrt”	“Log2”
AdaBoost	Base Estimator	The DT with the best setting of hyperparameter	Maximum Depth is 20
	The Number of Estimators	Every 5 from 20 to 70	55
	Algorithm	"SAMME", "SAMME.R"	"SAMME"
	Learning Rate	Every 0.05 from 0.05 to 0.5	0.15
Support Vector Machine	Regularisation Parameter (C)	Every 0.05 from 0.05 to 0.9	0.2
	Kernel	Linear, Poly, RBF, and Sigmoid	Linear
	Probability	True. Probability Required for Soft Voting	True
	Decision Function Shape	“OvO”. Since Route Choice Is Multi-Class Classification	“OvO”
Neural Networks	Activation	“Identity”, “Logistic”, “Tanh”, and “Relu”	“Relu”
	Solver	“LBFGS”, “SGD” and “ADAM”	“Adam”
	Learning Rate	Use a Constant Learning Rate	Constant
	Initial Learning Rate	0.001, 0.01, and Every 0.05 from 0.05 to 0.9	0.02

TABLE 4.3: Base models and the tested setting of their hyperparameters in Scikit Learn

4.3.4 Ensemble Rules

Ensemble rules are strategies to combine results from base models. Here, we first introduce how this can be accomplished with base models that predict classes or discrete outcomes.

4.3.4.1 Voting

Voting is commonly applied to classification tasks, normally including hard and soft voting. The process of hard voting is shown in Figure 4.4, and the result that appears most frequently is the ensemble result. For soft voting (weighted voting), as shown in Figure 4.5, it considers the probability of the result given by base models. For example, the probability of being purple $p = (35\% + 20\% + 40\% + 30\% + 8\% + 35\% + 50\% + 2\%)/8 = 27.5\%$. In that case, base models with a higher probability of their result have a more significant influence on the ensemble result.

In addition, inspired by the Australian election system, we propose a ranked choice voting strategy to aggregate results from base models as shown in Figure 4.6. For multi-class classification problems, each base model includes the preference order of alternatives, and the first choice from each base model consists of the initial result set. If none of the alternatives in the initial set obtains more than 50% of votes, the alternative with minimum votes will be removed for learners which come up with that alternative, and the votes will be recalculated for the remaining alternatives. As shown in Figure 4.6, since all colors are below 50%, the aqua blue is removed from the results in the left bottom and right bottom models. After that, as the red color exceeds 50%, the ensemble prediction is then red. For the binary choice problem, the Ranked Choice Voting strategy provides the same results as hard voting.

When using voting techniques to assemble results from base models, different base models can be issued with different weights to boost the influence of more accurate learners. In this study, we include a weighted or soft voting strategy, which weights base models based on their performance. The testing set is split into two subsets. One is used for calculating the weights for the models as shown in Equation 4.11, and another one is for testing all the models.

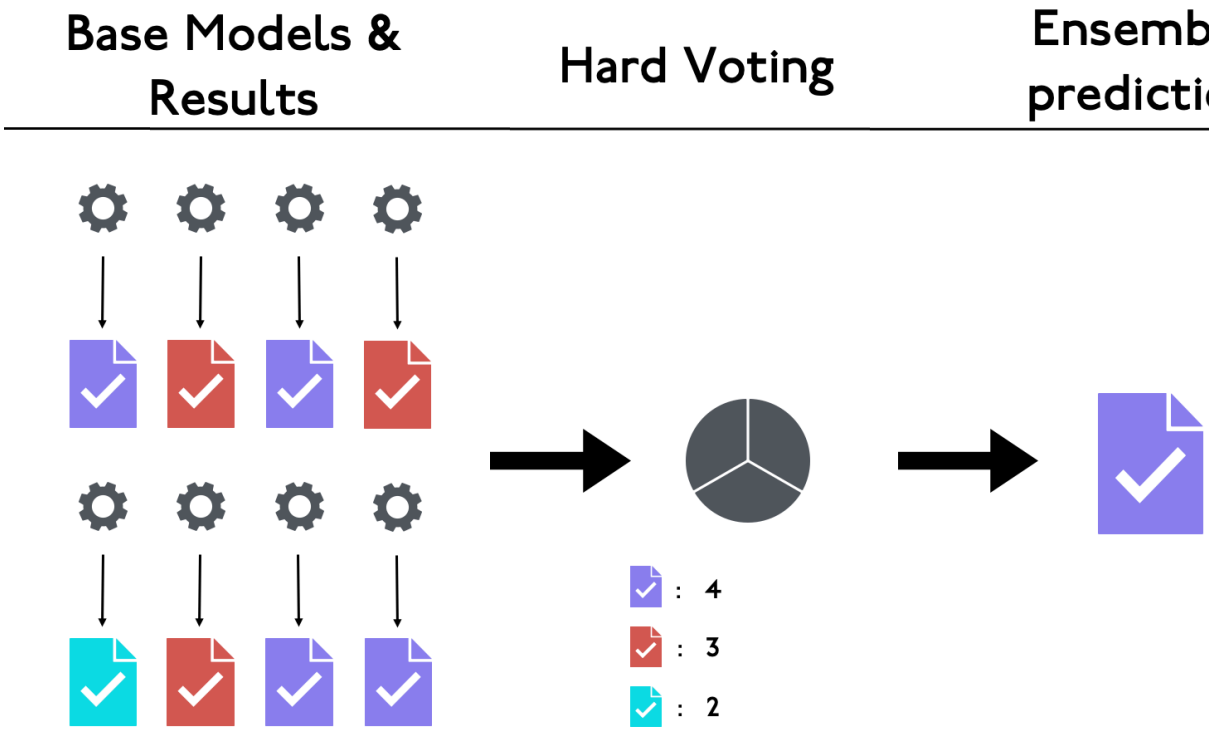


FIGURE 4.4: Schematic of Hard Voting

$$w_i = \frac{p_i}{min(p_i)} \tag{4.11}$$

- Where:
- w_i is the weight of base model i
- p_i is the performance score of model i in the sub-testing set
- $min(p_i)$ is the minimum performance score in all tested models.

4.3.4.2 Stacking

As discussed in chapter 1 and shown in Figure 2.2, the base models (first-level learners) are trained on the same sub-training set. The outputs from these first-level learners are then

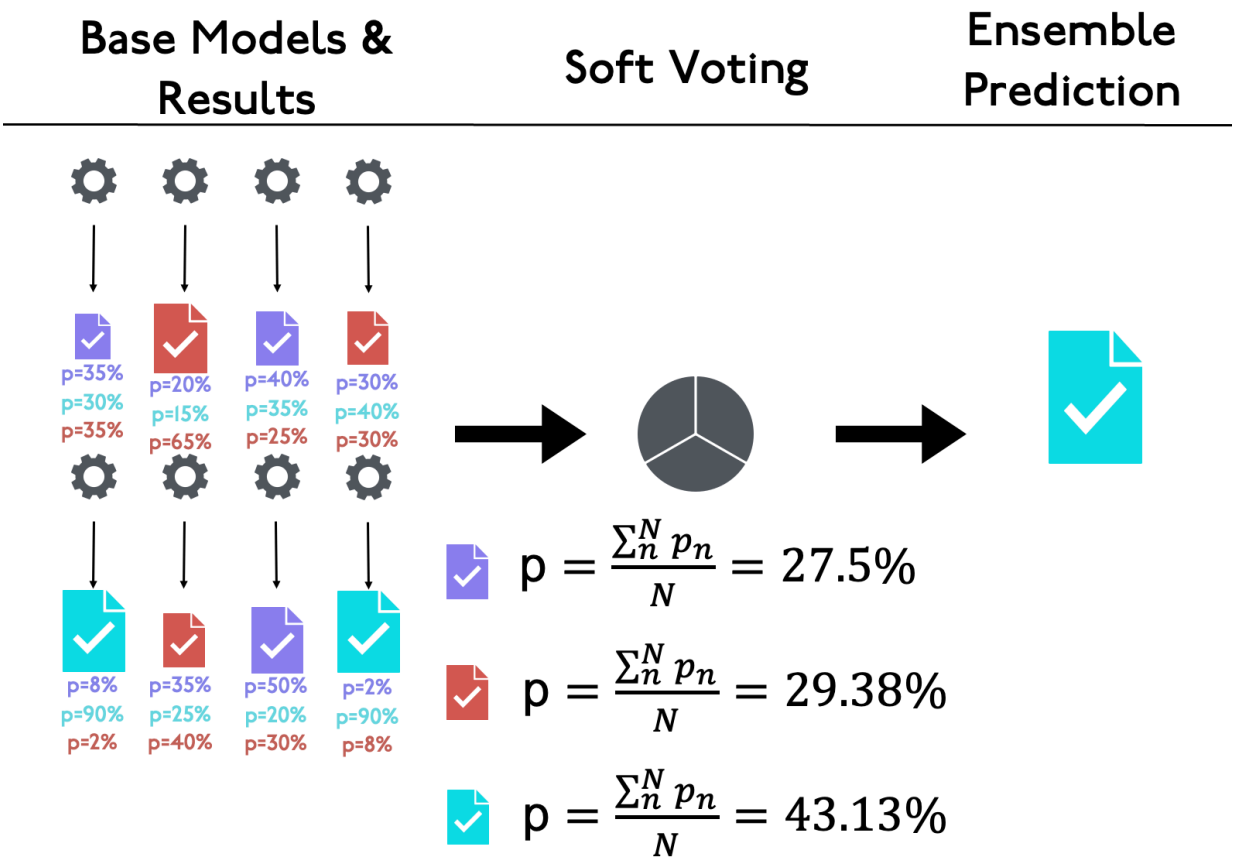


FIGURE 4.5: Schematic of Soft Voting

used as inputs for the second-level learner. The second-level learner is trained using the second sub-training set to avoid over-fitting, and the performance of the two-level ensemble is evaluated on the test set.

4.3.5 Homogeneous Ensemble

For homogeneous ensembles, three pairs of base model and homogeneous ensembles are evaluated:

- MNL and RMNL,
- PSL and RPSL and
- DT and RF.

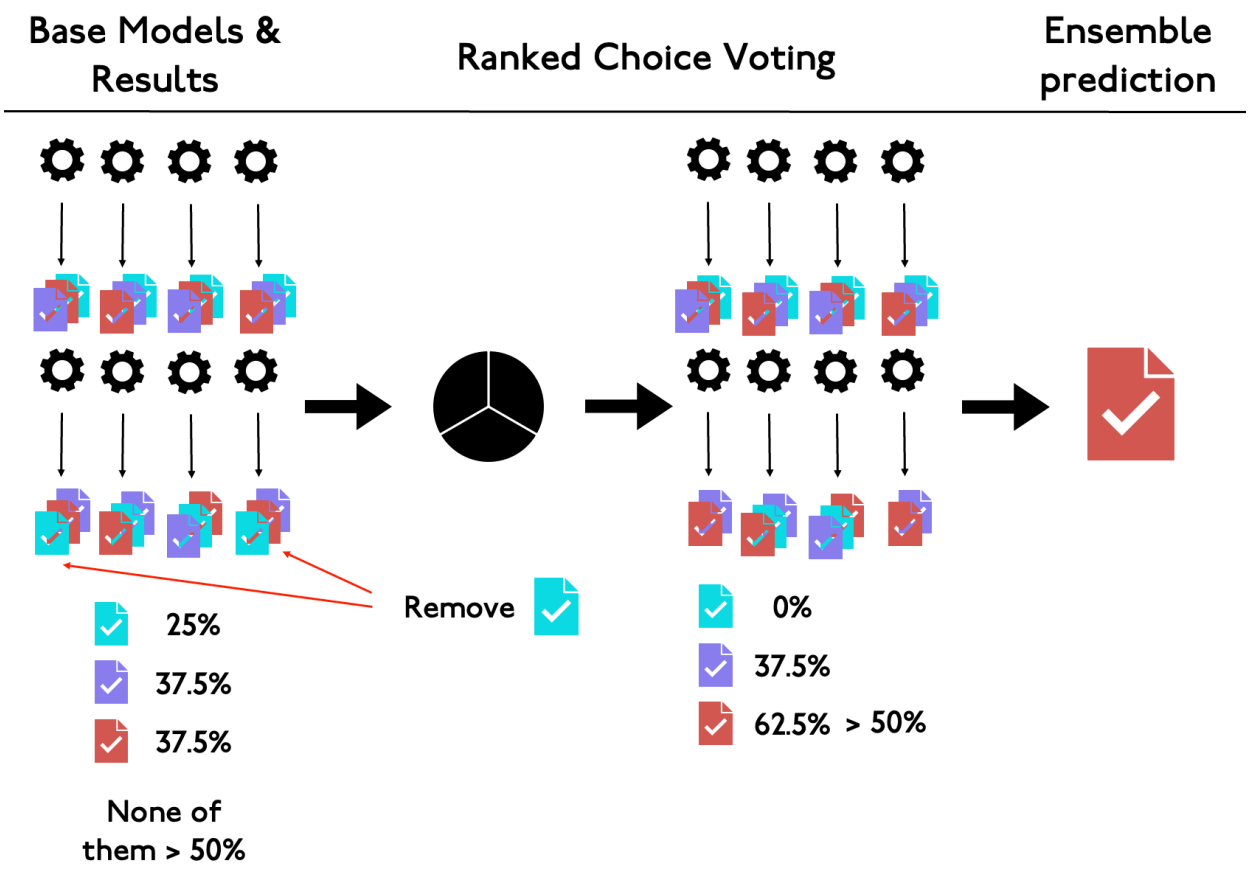


FIGURE 4.6: Schematic of Ranked Choice Voting

All three homogeneous ensembles implement bagging to gain a sub-sample for each base model (MNL for RMNL, PSL for RPSL and DT for RF) and random feature techniques to feed the base model element of all input attributes, and the results from all base models are assembled with a majority (hard) voting strategy. The idea is to show how homogeneous ensembles perform for different kinds of base models.

4.3.6 Heterogeneous Ensemble

For heterogeneous ensembles, all models in subsection 4.3.3 section are taken as base models. Ensemble rules including hard voting (EHV), soft voting (ESV), ranked choice voting (ERV), and stacking are used to combine outputs from base models. Specifically, they are listed as following:

(1) Heterogeneous ensemble using hard voting (EHV)

(2) Heterogeneous ensemble using ranked choice voting (ERV)

(3) Heterogeneous ensemble using soft voting (ESV)

(4) Heterogeneous ensemble using logit meta-learner (ESL)

(5) Heterogeneous ensemble using Random Forest meta-learner (ESR)

(6) Heterogeneous ensemble using AdaBoost meta-learner (ESA)

(7) Heterogeneous ensemble using SVM meta-learner (ESV)

Grid search is also applied to find the local best hyperparameters for Random Forest, AdaBoost, and Support Vector Machine meta-learners.

4.4 Correlation

For heterogeneous ensembles, when the types of base models are different, some base models can be more similar than the others, such as the Random Forest and the Extra Tree. Therefore, the correlation between base models are measured via the resemblence between the predicted probability between two base models. For ensemble using hard voting and ranked choice voting, the predicted alternative route is set to have probability equal to 1, and the rest alternatives have 0 probability. The Jensen-Shannon Divergence (JSD) is applied to measure the correlation between two probability distributions. The formula is shown in Equation 4.12.

$$JSD(P \parallel Q) = \frac{1}{2}\sum_{i} P(i)\log\left(\frac{P(i)}{B(i)}\right) + \frac{1}{2}\sum_{i} Q(i)\log\left(\frac{Q(i)}{B(i)}\right) \tag{4.12}$$

Where:

- P: is a probability distribution.
- Q: is another probability distribution.
- B: is the pointwise mean of the P and Q

		Predicted Condition	
		$Positive$	$Negative$
Actual Condition	$Positive$	TP	FN
	$Negative$	FP	TN

TABLE 4.4: TP: true positive, FP: false positive, TN: true negative, FN: false negative.

4.5 Evaluation

After obtaining the results from ensembles and other tested models, assessing their performance is a key step. There are a variety of ways to evaluate the goodness of fit of models.

4.5.1 Confusion Matrix

The confusion matrix is shown in Table 4.4, and each term is explained as follows:

(1) *True Positive (TP)*: chosen alternative routes which are correctly identified as predicted routes.
(2) *False Positive (FP)*: untaken routes which are incorrectly identified as predicted routes.
(3) *True Negative (TN)*: untaken routes which are correctly identified as predicted-to-be-untaken routes.
(4) *False Negative (FN)*: chosen alternative routes which are incorrectly identified as predicted-to-be-taken routes.

Three criteria are used for evaluating each model's performance:

(1) *Sensitivity (Recall)*: $Sensitivity = \frac{TP}{TP+FN}$ - the percentage of routes actually selected that were predicted to be selected.
(2) *Specificity*: $Specificity = \frac{TN}{TN+FP}$, - the percentage of routes not selected that are predicted correctly.
(3) *Precision*: $Precision = \frac{TP}{TP+FP}$, - the percentage of routes selected that are predicted correctly.

Model	Calculated Probability
Logit	
– Multinomial	Direct output of probabilities
– Path-Size	Direct output of probabilities
Trees	
– Decision Tree	Proportional to training samples in the leaf node
– Extra Tree	Mean predicted probabilities of the trees in the forest
– Random Forest	Same as Extra Tree
– AdaBoost	Weighted mean predicted probabilities of the tree models
Other ML Techniques	
– Support Vector Machine	Platt scaling for calibrated probabilities
– Multilayer Perceptron	Softmax activation function in the output layer
Stacking	
– Logit Models	Probabilities provided by the meta-model
– Random Forest	”
– AdaBoosting	”
– Support Vector Machine	”
Ensemble	
– Hard Voting	Probabilities of predicted classes set to 1, others to 0
– Ranked Choice Voting	Convert deterministic results to probabilistic
– Soft Voting	Mean predicted probabilities of individual models

TABLE 4.5: Calculated probability from tested models

4.5.2 Log-likelihood

In this study, the log-likelihood is applied to measure each models' performance for each random test, and the average log-likelihood from multiple tests is calculated for each model. Table 4.5 summarises the method for calcuating probability.

4.6 Similarity

One of the major difficulties with evaluating route choice models is the high number of choices and the overlap between choices. Traditional discrete choice models often treat each route as a distinct alternative, ignoring the overlapping segments between routes. However, in real-world scenarios, routes within a choice set frequently share common links, which reflects a form of continuity between them. For instance, consider a choice set consisting of three routes: A, B, and C. If route B is chosen, but route A shares many links with route B and none

with route C, treating these routes as purely discrete choices overlooks similarities between *A* and *B*. This discretized continuity implies that the common links between routes *A* and *B* should be acknowledged, as they impact travelers' perceptions and decisions. Ignoring these shared segments may lead to an incomplete understanding of route preferences and the factors influencing route choice. For that reason, we introduce a similarity measure to account for routes that are close but not identical.

The proposed similarity measurement includes three aspects: overlap, attribute similarity, and spatial similarity.

The first aspect is the overlap rate (Ω), which we saw in Path-size Logit. As routes consist of a series of links, the overlap rate measures the ratio of the total length of shared links to the total length of the target route as shown in Equation 4.13. Normally, the target route is the actual route selected by the individual. If the overlap rate between two routes is 1, then these two routes are the same.

$$\Omega = \frac{L_s}{L_t} \tag{4.13}$$

Where:

- L_s: the total length of links shared between predicted route and chosen route
- L_t: the total length of the chosen route

The advantage of overlap rate is that it directly shows what share of the predicted route is the same as the chosen route. However, the overlap rate cannot distinguish how different non-overlapped links are in the predicted and chosen routes. For example, two predicted routes *A* and *B* both include none of the links in the chosen route, but route *A* has the same travel time and travel distance as the chosen route, while route *B* needs more time and distance to reach the destination. In this case, *A* is more similar to the chosen route than *B*, even though the overlap rates between each predicted route and chosen route are both 0.

Therefore, a second aspect, 'attribute similarity' (s), is measured in this study. The attributes of a route are the accumulation of the bundle of attributes of links that comprise a route. When measuring the attribute similarity between two routes, the comparison is based on the non-overlapped links since attributes for shared links are the same. The formula for calculating the attribute similarity is shown in Equation 4.14.

$$s = 1 - \frac{1}{\sum_{x}^{|\mathbf{X}|} w_x} \cdot \sum_{x}^{|\mathbf{X}|} \frac{w_x \cdot |\hat{x_b} - x_b|}{\hat{x_b} + x_b}, \quad \forall \hat{x_b} \neq 0 \tag{4.14}$$

Where:

- w_x: the weight of attribute x
- $\mathbf{X}$: the set of attributes that is used in prediction
- $|\mathbf{X}|$: the number of attributes in $\mathbf{X}$
- x_b: attribute x of non-overlapped parts in chosen route
- $\hat{x_b}$: attribute x of non-overlapped parts of predicted route

As the similarity of two routes in different attributes is aggregated to a single value, the influence of one attribute differs from the others.

Once the attribute similarity between the two routes is considered, the difference between alternative routes, which both have 0 overlap rate with the selected route, can be computed. However, for two alternative routes, which both have the same value of attribute similarity and the overlap rate with the selected route, if one alternative route is one block away from the selected route and another one is ten blocks away from the selected route, the one closer to the selected route is expected to be more similar to the selected route for the individual. The reasons for that include, but are not limited to:

- Travellers are more likely to be familiar with the closest route. (Deviation is an indicator of familiarity, a property of the traveller rather than the route.)
- The closest route is more likely to pass through the region with similar adjacent land use or views. (Deviation accounts for unmeasured attributes).

In this study, the spatial similarity of the two routes is measured based on the deviation between them, which is not a property of route itself, and hence not an 'attribute' as typically measured, but rather a property of the relative position of one route compared to another. As shown in Equation 4.15 and Figure 4.7, the area formed by the two routes is reshaped to a rectangle whose length equals the length of the non-overlapped part in the selected route, and the width represents the average deviation between the two routes. Since the overlap rate and attribute similarity are unitless and between 0 and 1, in order to combine three aspects and to be consistent, the deviation is calculated based on the Equation 4.15.

$$d = \frac{\Lambda}{L_k} \tag{4.15}$$

Where:

- d: the average deviation between predicted route and chosen route
- Λ: the area of the region which is formed by predicted route and chosen route
- L_k: the total length of the non-overlapped links in the predicted route

$$D = \frac{d}{d + L_k} \tag{4.16}$$

If the predicted route is spatially closer to the chosen route, the average deviation d is close to 0, and D is also close to 0. When D is less than 0.5, the average deviation is smaller than the length of non-overlapped part in chosen route. Conversely, when D exceeds 0.5, the average deviation is longer than the length of non-overlapped part.

The overall similarity (S) between predicted and chosen routes includes the influence from all these aspects, shown in Equation 4.17.

$$S = \Omega + (1 - \Omega) \cdot (\alpha \cdot s + (1 - \alpha) \cdot (1 - D)) \tag{4.17}$$

Where:

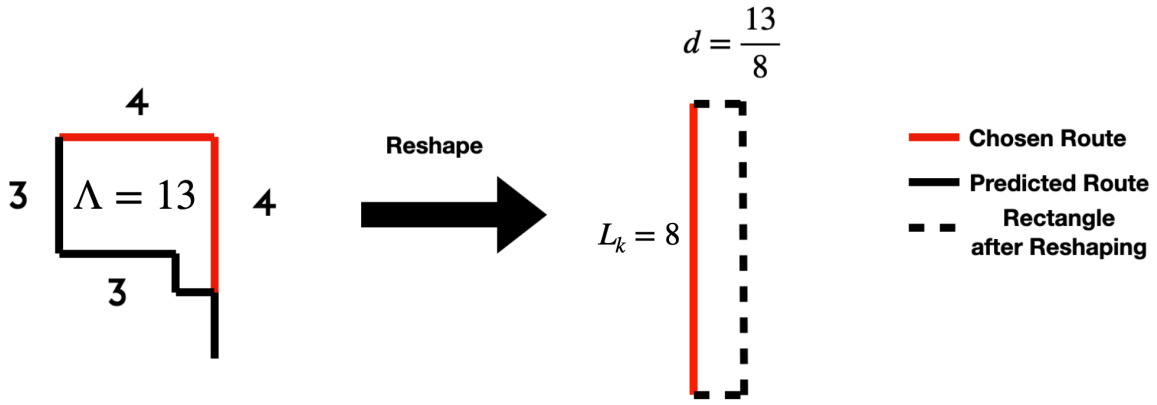


FIGURE 4.7: Average Deviation (d)

- Ω: the overlap rate between predicted route and chosen route
- s: the weighted average of attribute similarity between predicted route and chosen route
- D: the deviation between predicted route and chosen route
- $1 - D$: the spatial similarity between predicted route and chosen route
- α: the weight of attribute similarity in measuring overall similarity
- $1 - \alpha$: the weight of deviation in measuring overall similarity

The first term Ω in Equation 4.17 indicates the proportion of the chosen route which is the same as the predicted route. For the non-overlapped parts $(1 - \Omega)$, the term s measures the similarity of two routes in attributes, and $1 - D$ represents the similarity of two routes in space. In chapter A, there is an example to show how we calculate similarity.

When the predicted route fully overlaps with the chosen route, S equals 1. If none of links are shared between the predicted route and the chosen route, the combination of effects from the last two aspects are combined to show the similarity. However, for the non-overlapped parts, the attributes and deviation can have a different significance of influence on similarity measurement, but as the main purpose of measuring the similarity in this study is assessing the performance of different models, we set equal weights (i.e. $\alpha = 0.5$).

Frejinger and Bierlaire (2007) also argued that overlap does not fully capture the correlation between routes and, thus, introduced a subnetwork to capture perceptual correlation among non-overlapping routes. The similarity measure introduced in this study is different from their subnetwork method in the following aspects.

(1) Different in implementation: The proposed similarity measure can be directly applied to any set of routes with attributes and spatial information, whereas the subnetwork is defined by motorways and main roads which are arbitrarily selected by analysts or identified based on the names commonly used by travellers in interviews. The similarity among routes is influenced by the shared links between those routes and the subnetworks.
(2) Different in purpose: The subnetwork is designed to capture the correlation effects among routes during the modelling process, whereas the proposed similarity measure aims to describe the similarities between routes for several puposes, including assessing goodness of fit.

4.7 Validation

In order to further evaluate the performance of ensembles in predicting traffic flow on freeways, two data sets are used to obtain the OD pairs for all morning commute trips in the Twin Cities, and the final prediction of Vehicle Kilometers Travelled (VKT) on freeway links are compared with the recorded data from loop detectors. The 2010 Origin-Destination Employment Statistics from the Longitudinal Employer-Household Dynamics (LEHD) and

the survey data from 2010 Travel Behavior Inventory (TBI) are used for validating the tested models.

In order to assess the performance of the proposed method, data from 525 loop detectors in the Twin Cities in 2010 are used as ground truth. For each detector, the recorded flow is aggregated at every 5 mins from 6:00 am to 9:00 am for every recorded day. The overall VKT based on loop detector data is compared with that from the proposed methods to test the performance of the various models.

Ensembles are applied to predict travellers' route choice, and the results are compared to other machine learning models and logit models. After obtaining the route choice of each traveller, the number of routes that pass though each freeway link during the 3 hour morning peak is counted and multiplied by the length of the link. The production is the predicted VKT on the set of freeway links for which counts are available and compared to the VKT from those loop detectors. The following describes how we applied the proposed method in both data sets.

4.7.1 Origins and Destinations

The LEHD data set includes all home and workplace locations at the census block level, there is no need to do further analysis in this case.

For the TBI data set, since only 1% of population took the survey, the home and workplace location for the whole population is generated based on the weight ($W_{g,h}$) of each recorded OD pair in the survey and the resident population (R) and employment opportunities (E) at each TAZ. For example, for a workplace TAZ (destination) in the TBI data set, it is connected to one or more home TAZ (origins). The number of trips $N_{g,h}$ from origin g to destination h is based on Equation 4.18.

$$N_{g,h} = \frac{E_h(W_g \cdot R_g)}{\sum_{g=1}^{G}(W_g \cdot R_g)} \tag{4.18}$$

Process	Trips Remaining	Percentage of Dataset
Raw data	18639	100%
Morning Trips	10237	55%
All-day Home to Work Trips	6570	35%
Morning Home to Work (6:00 - 9:00 am)	4557	24%
Morning Home to Work Auto Trips	3646	20%

TABLE 4.6: TBI trip proportion

Once the number of trips $N_{g,h}$ from home zone o_g to workplace zone h is gained, these origins and destinations are randomly distributed within the TAZs, and further matched to the nearest node in the road network.

4.7.2 Modes

Since this study only focuses on the car trips in the morning for commute purpose, trips using other transport mode are removed from the data set. Information about transport mode choice, departure time, and trip purpose are not included in the LEHD data set. Therefore, we estimate the proportion of morning commute (home to work (H2W)) car trips in the TBI survey and apply it to LEHD data set. Overall, 7049 participants in the home interviews survey drive on the road during the morning peak (6:00 - 9:00 am) for various purpose. The Summary of TBI data is presented in Table 4.6.

Based on the results, out of the people who commute to work from home, $\frac{4557}{6570} = 69\%$ travel between 6:00 - 9:00 am, and $\frac{3646}{4557} = 80\%$ of these morning commute trips use cars. For the remaining 20% of morning commute trips, most of them are walking, biking, or public transport trips, but directly multiplying the ratio of car trips by the generated trips from LEHD data set ignores the effect of spatial distribution of public transport service (as operationalized using access to bus stops), which also accounts for the feasibility of walking or using a bicycle. Therefore, we identify OD pairs which have bus stops within a 500 meter buffer zone at both the origin and destination. For these trips we multiply the total number of home-work pairs by a correction factor 0.8 to account for non-auto trips.

4.7.3 Routes

The hybrid method described in section 4.2 is applied to generate route choice sets. For route choice modelling, all the models with the best hyperparameter setting are trained from case studies 1 and 2 and used to predict routes. Since many tested models provide the probability of choosing each alternative route, for each simulation, the final prediction of each probablistic model is drawn based on the probability distribution rather than directly using the route with the highest probability, and 10 simulations are completed for each data set. Once the route choice for all travellers are obtained, they are summarised on each link, and the aggregated values on freeway links with observed traffic counts are multiplied by the length of the link to obtain the predicted VKT. The predicted VKT are compared with the observed VKT which is calculated based on the traffic counts from loop detector stations.

The Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE), and the overall network percentage error based on the sum of VKT on all links with loop detectors (E_{net}) are recorded for each simulation. E_{net} is calculated based on Equation 4.19.

$$E_{\text{net}} = \frac{\sum_{\phi=1}^{\Phi} VKT_{\text{detector}} - \sum_{\phi=1}^{\Phi} VKT_{\text{predicted}}}{\sum_{\phi=1}^{\Phi} VKT_{\text{detector}}} \times 100\% \tag{4.19}$$

Where:

- E_{net}: the network percentage error between prediction and loop detector observation
- Φ: The total number of loop detectors (indexed by phi) included in this study

4.8 Interpretation

Unlike Path-size and Multinomial Logit models, machine learning approaches are often seen as difficult to comprehend, and are commonly referred to as a 'black box.' While the tree structure of a Decision Tree can help people understand the model's decision-making process step by step, most machine learning models lack this transparency. To address this, an explainable AI technique, the SHapley Additive exPlanation (SHAP) value (Lundberg

and Lee 2017), is widely used in many studies and is adopted in this research. SHAP uses a game-theoretic approach to explain machine learning models by calculating the Shapley value from cooperative game theory for each feature, as shown in Equation 4.20. Understanding the SHAP value provides insights into the importance and impact of individual features on the model's predictions. In this study, the model's outcome is considered the 'game,' and the features used in the model are the 'players.' Additionally, this study compares feature importance calculated by Random Forest using Gini importance with the relevant features identified by SHAP values.

4.8.1 Shapley Value

To explain SHAP values, we introduce Shapley values. The Shapley value from cooperative game theory assigns a value to each player in a game representing their contribution to the total payoff Shapley et al. (1953). This value is calculated based on the marginal contributions of a player across all possible coalitions. The Shapley value for a player x in a game with $|\mathbf{X}|$ players and a value function f is determined using the formula shown in Equation 4.20.

$$\psi_x = \sum_{\mathbf{M} \subseteq \mathbf{X} \setminus \{x\}} \frac{|\mathbf{M}|!(|\mathbf{X}| - |\mathbf{M}| - 1)!}{|\mathbf{X}|!} [f(\mathbf{M} \cup \{x\}) - f(\mathbf{M})] \tag{4.20}$$

where:

- ψ_x is the Shapley value for player (feature) x.
- $\mathbf{X}$ is the set of all players (features).
- $\mathbf{M}$ is a subset of $\mathbf{X}$ that does not include feature x.
- $f(\mathbf{M} \cup \{x\})$ is the model's predicted probability for a given instance with the features in subset $|\mathbf{X}|$ along with feature x.
- $f(|\mathbf{X}|)$ is the model's predicted probability for the same instance with only the features in subset $|\mathbf{X}|$.

Shapley et al. (1953) proved that the Shapley value is the only measure to satisfy four axioms to achieve fairness, which are:

Coalition of Palyers (Order does not matter)	Gained prize ($)
Null Player	0
A,B, and C	100
A and B	75
A and C	75
B and C	50
Only A	50
Only B	50
Only C	0

TABLE 4.7: Basic information of players

(1) Efficiency, which states that none of the game value is left over.

(2) Symmetry, which states that two players are considered interchangeable if they contribute equally to all coalitions, and the two players gain the same total values of the game.

(3) Dummy player (null player), which states that if players provide zero marginal contribution to any coalitions then none of the games' total value will be given to them.

(4) Additivity, which states that the overall contributions for a player over two games is the sum of the contributions for the two individual games. The assumption of this axiom is that any game played is independent.

Consider a simple example where three players ($N = \{A, B, C\}$) form a group to attend a competition. The first prize, second prize, and third prize are \$100, \$75, and \$50. In a cooperative game setting, the value a player adds to a coalition is determined by the set of players already present, regardless of the order in which they arrived. The prize that each player and any combination can win are listed in Table 4.7. Considering permutations in the calculation of the Shapley value ensures fairness by averaging the marginal contributions over all possible join sequences. The calculated Shapley value for each player is presented in Table 4.8.

For example, in the first row of Table 4.8, the marginal contribution of the first player, A, equals the prize A can gain by himself (\$50) minus the prize when no player attends (\$0). The marginal contribution of player B equals the prize A and B can obtain together (\$75)

Permutation	Probability	Marginal Contributions of A	Marginal Contributions of B	Marginal Contributions of C
A, B, C	1/6	$50 - 0 = 50$	$75 - 50 = 25$	$100 - 75 = 25$
A, C, B	1/6	$50 - 0 = 50$	$100 - 75 = 25$	$75 - 50 = 25$
B, A, C	1/6	$75 - 50 = 25$	$50 - 0 = 50$	$100 - 75 = 25$
B, C, A	1/6	$100 - 50 = 50$	$50 - 0 = 50$	$50 - 50 = 0$
C, A, B	1/6	$75 - 0 = 75$	$100 - 75 = 25$	$0 - 0 = 0$
C, B, A	1/6	$100 - 50 = 50$	$50 - 0 = 50$	$0 - 0 = 0$

	Calculation
Shapley Value of A	$\frac{50+50+25+50+75+50}{6} = 50$
Shapley Value of B	$\frac{25+25+50+50+25+50}{6} = 37.5$
Shapley Value of C	$\frac{25+25+25+0+0+0}{6} = 12.5$

TABLE 4.8: Shapley value calculation for each player

minus the prize player B would receive alone ($50). Similarly, the marginal contribution of player C is the total prize A, B, and C can win ($100) minus the prize A and B together can win ($75). In the last permutation 'C, B, A', the marginal contributions of A equal the prize that the coalition of A, B, and C can win ($100) minus the prize that the coalition of C and B (order does not matter) can win ($75). The Shapley value of player A equals the average of A's marginal contributions across all six permutations.

4.8.2 SHAP Value

The adaptation of Shapley values to machine learning provides a method for interpreting model predictions. By treating features as players, the model output as the value function, and the prediction task as the game, SHAP uses the concept of Shapley values to explain the contribution of each feature to the prediction (Lundberg and Lee 2017).

The additive feature attribution framework decomposes a model's prediction into the sum of contributions from individual features, as shown in Equation 4.21. The sum of the base value and the SHAP values for all features gives the model's prediction for a specific instance. It should be noted that the base value ϕ_0 is calculated as the expected model output when no features are present.

	Probability of Route		
	$p(Y_1)$	$p(Y_2)$	$p(Y_3)$
Base Value	0.3	0.4	0.3
Predicted Value	0.2	0.5	0.3
No Feature	0.3	0.4	0.3
L_1 only	0.25	0.45	0.3
L_2 only	0.28	0.42	0.3
L_3 only	0.27	0.43	0.3
L_1 and L_2	0.22	0.48	0.3
L_1 and L_3	0.22	0.49	0.3
L_2 and L_3	0.23	0.47	0.3
L_1, L_2, and L_3	0.2	0.5	0.3

TABLE 4.9: Basic information in example of SHAP calculation

$$f(j) = \phi_0 + \sum_{x=1}^{\mathbf{X}} \phi_x \tag{4.21}$$

Where:

- $f(j)$ is the model's prediction for instance j
- ϕ_0 is the base value (often the mean prediction over the training data).
- n is the total number of features.
- ϕ_x is the the contribution of feature x to the difference between the actual prediction $f(j)$ and the base value ϕ_0. It is computed using Shapley values

By using the framework of additive feature attribution and Shapley value within this framework, SHAP ensures consistency and local accuracy in feature attribution (Lundberg and Lee 2017). To help understand how SHAP works, consider a route choice problem that has three alternative routes (Y_1, Y_2, and Y_3) and three features, specifically the length functions ($L(Y_1)$, $L(Y_2)$, and $L(Y_3)$). The base value, prediction for one instance, and the value of using different combinations of features are presented in Table 4.9.

The step-by-step results for calculating the SHAP is shown in Table 4.10.

	Permutation (order matters)						Average
	L_1, L_2, and L_3	L_1, L_3, and L_2	L_2, L_1, and L_3	L_2, L_3, and L_1	L_3, L_1, and L_2	L_3, L_2, and L_1	
			Alternative Route Y_1				
L_1	$0.25 - 0.3 = -0.05$	$0.25 - 0.3 = -0.05$	$0.22 - 0.28 = -0.06$	$0.2 - 0.23 = -0.03$	$0.22 - 0.27 = -0.05$	$0.2 - 0.23 = -0.03$	-0.045
L_2	$0.22 - 0.25 = -0.03$	$0.2 - 0.22 = -0.02$	$0.28 - 0.3 = -0.02$	$0.28 - 0.3 = -0.02$	$0.2 - 0.22 = -0.02$	$0.23 - 0.27 = -0.04$	-0.025
L_3	$0.2 - 0.22 = -0.02$	$0.22 - 0.25 = -0.03$	$0.2 - 0.22 = -0.02$	$0.23 - 0.28 = -0.05$	$0.27 - 0.3 = -0.03$	$0.27 - 0.3 = -0.03$	-0.03
			Alternative Route Y_2				
L_1	$0.45 - 0.4 = 0.05$	$0.45 - 0.4 = 0.05$	$0.48 - 0.42 = 0.06$	$0.5 - 0.47 = 0.03$	$0.49 - 0.43 = 0.06$	$0.5 - 0.47 = 0.03$	0.047
L_2	$0.48 - 0.45 = 0.03$	$0.5 - 0.49 = 0.01$	$0.42 - 0.4 = 0.02$	$0.42 - 0.4 = 0.02$	$0.5 - 0.49 = 0.01$	$0.47 - 0.43 = 0.04$	0.022
L_3	$0.5 - 0.48 = 0.02$	$0.49 - 0.45 = 0.04$	$0.5 - 0.48 = 0.02$	$0.47 - 0.42 = 0.05$	$0.43 - 0.4 = 0.03$	$0.43 - 0.4 = 0.03$	0.031
			Alternative Route Y_3				
L_1	$0.3 - 0.3 = 0$	$0.3 - 0.3 = 0$	$0.3 - 0.3 = 0$	$0.3 - 0.3 = 0$	$0.3 - 0.3 = 0$	$0.3 - 0.3 = 0$	0
L_2	$0.3 - 0.3 = 0$	$0.3 - 0.3 = 0$	$0.3 - 0.3 = 0$	$0.3 - 0.3 = 0$	$0.3 - 0.3 = 0$	$0.3 - 0.3 = 0$	0
L_3	$0.3 - 0.3 = 0$	$0.3 - 0.3 = 0$	$0.3 - 0.3 = 0$	$0.3 - 0.3 = 0$	$0.3 - 0.3 = 0$	$0.3 - 0.3 = 0$	0

TABLE 4.10: Step-by-Step calculation in the simple example

The SHAP values for each feature indicate how much each feature contributes to the difference between the base value and the predicted value for each class. Positive SHAP values indicate a positive contribution towards the prediction, while negative SHAP values indicate a negative contribution. According to the results in Table 4.10, for alternative route Y_1, all three features have negative SHAP values, indicating they contribute to lowering the probability from the base value. For Y_2, all features have positive SHAP values, indicating they contribute to increasing the probability from the base value. Since all SHAP values for Y_3 are zero, there is no contribution to changing the probability from the base value. This simple example, which shows the case for one instance (observation), illustrates how aggregating SHAP values across many instances helps to understand the overall importance of each feature. Given approximately 600 trips (20% of the data) in the testing set, 40 unique routes for each OD pair, 9 features for each alternative route, and 17 models included in our study, it's inefficient to show every route in all instances.

Therefore, we use a Python package called SHAP, which provides a group of explainers for calculating SHAP values. In this study, both the tree explainer and kernel explainer from that package are included. The tree explainer, which offers faster computational speed, is compatible only with tree-based models. For other models, the kernel explainers are used. Each model predicts the probability of choosing a specific alternative route, and the SHAP value of each feature is calculated for each alternative route. In the same model, SHAP values of the same feature can vary across different alternatives. A positive SHAP value means that the feature increases the chance of the model predicting a particular alternative for a person. By analyzing the distribution of SHAP values across all individuals, any trends or patterns can be identified and explored.

When presenting the results in chapter 5, instead of showing the SHAP value for one instance, we pick one alternative route and illustrate the distribution of SHAP values for each important feature across all instances. This distribution reveals the relationship between the feature and its impact on prediction. Features with larger absolute SHAP values are deemed more important for the model's predictions, referred to as local interpretation in this study. Moreover, for each instance, we record the SHAP values for all features for the alternative route with the

highest probability. By calculating the mean of these recorded SHAP values for every feature, we identify the top 10 features with the highest absolute mean SHAP values.

Additionally, in Random Forest, the feature importance is reflected based on Gini importance. For each feature, Random Forest calculates the total reduction in the criterion brought by that feature across all trees in the forest. The importance of a feature is the average (or sum) of the criterion reduction for that feature across all trees. This study finds the top 5 important features based on the Gini importance in Random Forest and compares these results to the findings from using SHAP value.

CHAPTER 5

Results

In this chapter, we first show the performance of the generated choice set. Since the method is developed with the I-35W Bridge Collapse Study (CS2), the detailed step-by-step results come from CS2, and the final results of using same method in the 2010 Travel Behavior Inventory GPS Trajectory Data (CS1) are also presented. Following that, the linear and logit model to explore the importance of input variables, and their results are included. Then, the results of using conventional logit models, popular machine learning models, and a group of ensembles to predict route choice in both case studies are compared and discussed. Finally, results of using the trained model from two case studies to forecast VKT on freeways in two validation data sets are illustrated.

5.1 Overlap

For all route generation algorithms described in chapter 4, the alternative routes are compared with the observed trip, and the overlap rate (Ω) is measured. The results are presented in Table 5.1. If the two generated routes are not exactly the same, they will be identified as two unique routes. As the overlap rate threshold (O) increases, more observed trips are shared in the routes generated from the algorithms. Thus, the capture rate decreases with the increase of the overlap rate threshold for all labels.

Shortest distance path and shortest free-flow time are common labels used in previous studies. The capture rates under various overlap threshold for these labels are presented in Table 5.2 and compared to past studies.

TABLE 5.1: Capture rate for various labels under different overlap threshold

Labels	Unique routes	Capture rate (%) under overlap threshold			
		$O = 70\%$	$O = 80\%$	$O = 90\%$	$O = 100\%$
Shortest distance	1	37	31	27	24
Free-flow travel time					
Freeway preferred factor=0.8	1	55	48	42	26
Minimum free-flow time	1	52	46	39	26
Freeway avoided factor =1.05	1	50	42	34	25
K-minimum free-flow time	20	57	49	38	26
K-freeway-preferred factor=0.8	20	56	47	37	25
Perfectly correlated scenario (M1)					
Freeway preferred factor=0.8	avg. 3.5	58	50	41	22
Shortest time factor=1	avg. 3.7	57	48	39	18
Freeway avoided factor=1.05	avg. 4	56	47	37	18
Perfectly independent scenario (M2)					
Freeway preferred factor=0.8	avg. 7.0	61	52	45	25
Shortest time factor=1	avg. 8.3	60	52	44	22
Freeway avoided factor=1.05	avg. 8.7	60	52	43	22
Minimum left turns	1	45	35	30	25
Shortest distance and least free-flow time path with all freeway factors	avg. 4.3	71	63	51	33
Choice set including all 52 labels in Table 4.1	avg. 107	90	81	71	44

TABLE 5.2: Percentage of observed trips captured by shortest distance path and shortest free-flow travel time path under various thresholds

Overlap threshold (O)	Capture rate in CS1	Capture rate in CS2	Capture rate in (Bekhor et al. 2006)	Capture rate in (Zhu and Levinson 2015)
50%	48% (64%)	47%(59%)	None	None
60%	40% (57%)	42%(54%)	None	None
70%	37% (52%)	37%(48%)	None	None
80%	31% (46%)	32%(41%)	28% (46%)	9% (23%)
90%	27% (39%)	29%(34%)	22% (37%)	5% (16%)
100%	24% (26%)	27%(28%)	20% (34%)	2% (6%)

Note: The percentage outside of the parentheses represents the capture rate for the shortest distance path, while the percentage inside the parentheses represents the capture rate for the shortest free-flow travel time path.

Comparing results with those from Bekhor et al. (2006) and Zhu and Levinson (2015), shortest free-flow path captures more observed routes in all three studies.

As shown in Figure 5.1, the capture rate varies with the weight assigned to freeways. When the freeway factor is low (i.e., freeway time is more expensive than non-freeway time), the capture rate increases and reaches its peak when the freeway factor ranges from 0.7 to 1.11. After this point, it declines across all four overlap thresholds. For a perfect match (overlap = 100%), even though the number of unique routes found based on free-flow travel time is smaller than in the other two scenarios, more observed trips are captured with a factor near 0.7. In general, according to Figure 5.1, freeway-preferred paths capture a higher percentage of observed trips compared to freeway-avoided paths across all three scenarios.

As described in chapter 4, 52 labels are applied, and on average, 107 unique routes are defined for each observed trip. Considering the marginal effect of adding more alternatives to the choice set, for each iteration, only the label that results in the highest capture rate is added to the choice set. The 'best' label here refers to the one that captures the most previously uncovered trips in the choice set. Figure 5.2 shows the cumulative capture rate (the cumulative c) after adding the best label for a total of 10 labels, with a threshold of 80%. The increase in the capture rate tends to be small (around 1%) after using 6 labels.

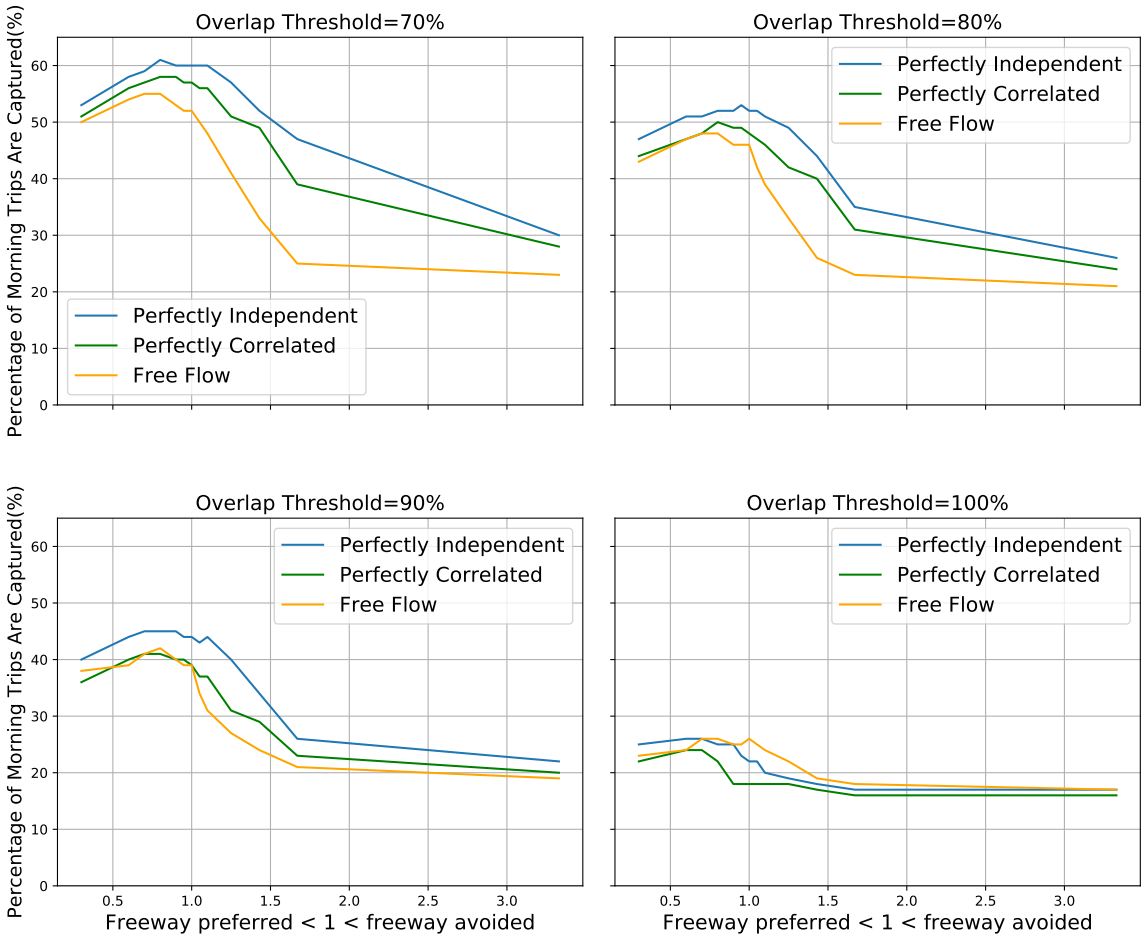


FIGURE 5.1: Comparison of Capture Rate under Different Overlap Thresholds (O) for Each Freeway Factor in 3 Scenarios: Free-flow, Perfectly Independent (M2), Perfectly Correlated (M1)

5.2 Deviation

The results of capture rate for 8 different deviation thresholds (Δ) are presented in Table 5.3. 88% of observed trips have at least one generated alternative route with an average deviation less than 50 meters.

Both overlap and deviation can be used to assess the performance of choice sets. A comparison is made between predicting deviation and predicting overlap to determine the more suitable independent variable for the analysis. To simplify the process, a choice set which includes the

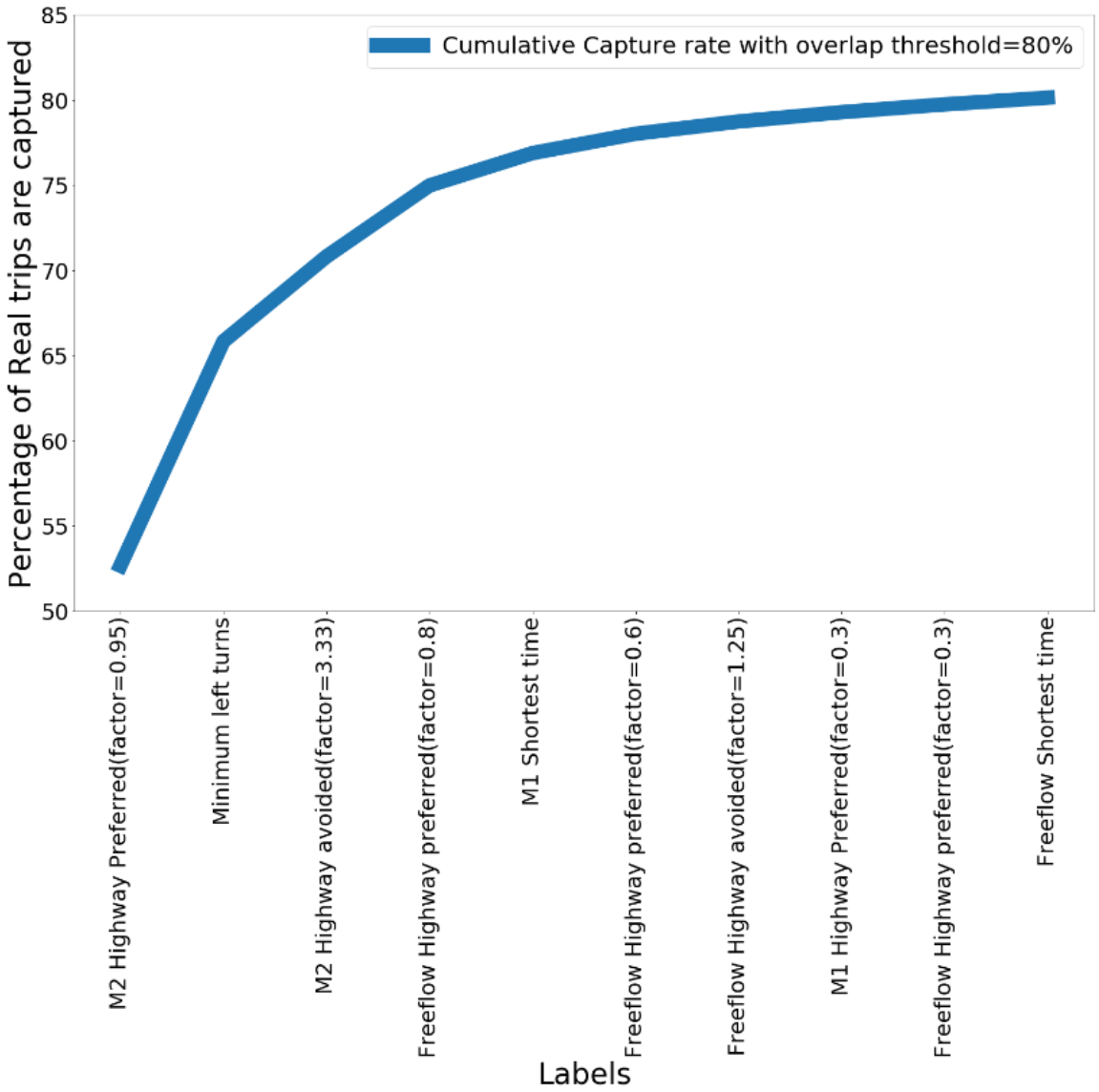


FIGURE 5.2: Cumulative Capture Rate (the cumulative c) of Adding 10 Best Labels in Choice Set

'shortest distance path' and 13 'least free-flow time paths with different freeway factors' is used. With 70% overlap threshold, 70% of observed trips could be captured by this choice set.

According to Breusch-Pagan test ($p = 0.0$ for F-test) and Durbin-Watson test results ($DW = 0.43$), there is heteroskedasticity and auto-correlation in the data set, and the Hausman test ($p = 0.97$) leads us to accept the null hypothesis that there is no correlation between the regressors and the errors. Therefore, a random effects model is implemented when estimating overlap and average deviation. The results is presented in Table 5.4. Overall, the R^2 for

TABLE 5.3: Percentage of observed trips covered by different labels under different deviation threshold

Threshold Δ	Shortest distance	Shortest free-flow time	Shortest time (M1)	Shortest time (M2)	combine all labels in Methodology
5m	24%	31%	36%	39%	62%
50m	35%	48%	53%	55%	88%
100m	42%	53%	60%	62%	93%
200m	50%	61%	67%	70%	97%
300m	63%	68%	72%	75%	98%
400m	68%	73%	77%	78%	98%
500m	69%	77%	80%	81%	98%
800m	78%	84%	84%	85%	100%

overlap is higher than for average deviation with the same independent variables, but they provide a consistent conclusion:

(1) For deviation: longer distances and less freeway percentage are associated with larger average deviation from the observed trip.
(2) For overlap: higher freeway percentage and lower distances, right turns, traffic lights, and bus stop coverage, are associated with high overlap with the observed trip.

5.3 Estimation

5.3.1 Linear Model

The overlap is selected as an independent variable for the linear model, and the selected alternative is defined based on the maximum overlap. As presented in Figure 5.2, with 80% overlap threshold, a choice set which is generated based on the first 10 labels captures 80% of observed trips. This best-10-label choice set averages 40 unique routes for each OD pair and is applied for route choice modelling process. Similar to above, the random effects model is suggested for modelling overlap for the best-10-label choice set, and the path-size term

TABLE 5.4: Linear models predict deviation (d) and overlap (Ω) with the chosen route

β	Deviation (d)		Overlap (Ω)	
constant	51.84 (63.31)		0.50 (0.05)	***
Length	0.05 (0.005)	***	$-5.69e^{-}6$ $(2.14e^{-}6)$	***
Freeway percentage (F)	-452.34 (123.77)		0.178 (0.178)	**
Traffic light	NS		-0.012 (0.005)	**
Right turns	NS		-0.017 (-2.34)	**
Bus stop coverage	NS		-28.32 (12.63)	**
R^2	0.195		0.366	

[1] (standard error)
[2] NS: Not significant
[3] *: significance at 10% level
[4] **: significance at 5% level
[5] ***: significance at 1% level

(Z), introduced in Equation 4.5 is also added as an independent variable. According to the estimated results in Table 5.5, by adding the path-size term, R^2 increases by 0.11. Routes with a larger freeway percentage and average free-flow speed and less trip length and fewer traffic lights are more likely to have a higher overlap with observed trips. The number of left turns and right turns are not significant based on the statistical result, which contrasts with results in the literature. This might be caused by the composition of the choice set. As shown in Table 5.4, the number of right turns is significant when using the simple choice set which is formed by the shortest distance and 13 least free-flow time path. The congested travel times were also tested, but they were not statistically significant.

TABLE 5.5: Linear random effects model with and without path-size term (Z)

β	without Z		with Z	
constant	-0.37 (0.136)	***	-0.31 (0.129)	**
Length	-1.51e-5 (3.18e-6)	***	-9.45e-6 (3.03e-6)	**
Free-flow speed	0.011 (0.0023)	***	0.0054 (0.0022)	**
Freeway percentage (F)	0.237 (0.058)	***	0.23 (0.051)	***
Traffic light coverage	68.68 (14.72)	***	25.86 (14.7)	*
Bus stop coverage	35.8 (14.58)	**	14.69 (13.013)	
Traffic light	-0.0038 (0.0014)	***	-0.004 (0.0013)	**
Bus stop	-0.0022 (0.0005)	***	-0.0005 (0.0004)	
Left turns	-0.0032 (0.0008)	***	-0.001 (0.001)	
Right turns	-0.0006 (0.001)	***	0.0009 (0.0011)	
$ln(Z)$			-0.127 (0.0087)	***
R^2	0.336		0.450	
R^2 (Overall)	0.309		0.439	

(standard error)
NS: Not significant
*: significance at 10% level
**: significance at 5% level
***: significance at 1% level

5.3.2 Logit Model

As described above, overlap is used to determine the selection of routes in choice set. The coefficients β in the utility function Equation 4.6 are assumed to follow the normal distribution, and simulations are replicated 20 times to gain draws from the distribution. Additionally, a Path-size Logit model is applied to compare the performance with the mixed logit model with the Z term. The results are presented in Table 5.6.

In Table 5.6, across the three models, only variables that were identified to be significant were used from the explanatory variables section. The variable 'left turns' has the highest variance inflation factor (VIF) among all variables. Its value equals the standard threshold (10) of VIF. Since the effect of the number of left turns on the route choice is worth exploring, this variable is retained. With the $ln(Z)$ term, log-likelihood, AIC, and BIC improve. For the population average, routes with higher freeway percentage and bus stops and less trip length, traffic lights, left turns, and right turns are more likely to be chosen. It should be noted that, as β is assumed to follow the normal distribution, the coefficients of the mixed logit model presented in Table 5.6 is the mean value. We recognize that there might be some share of the population who have an opposite sign for any given β.

TABLE 5.6: Mixed Logit (MXL) model without vs with Z term vs Path-size Logit (PSL) model

	Mixed logit				PSL	
β	without Z		with Z		with Z	
Length	-0.0013	***	-0.0013	***	-0.00036	***
	(0.0007)		(0.0007)			
Free-flow speed	0.18	***	0.19	***	0.176	***
	(-0.16)		(-0.149)			
Freeway percentage (F)	2.67	***	2.24	***	2.27	***
	(0.15)		(1.83)			
Traffic light	-0.05	***	-0.025	***	0.034	***
	(0.108)		(0.086)			
Bus stop	0.049	***	0.04	***	0.0463	***
	(-0.08)		(-0.092)			
Left turns	-0.03	***	-0.048	***	-0.049	***
	(-0.059)		(-0.048)			
Right turns	-0.06	***	-0.045	***	-0.119	***
	(0.016)		(0.095)			
$ln(Z)$			0.33	***	-0.078	**
			(-0.168)			
Log-Likelihood	-10,675.82		-10,609.644		-12,522.55	
AIC	21,387.64		21,259.289		25,273.1	
BIC	21,503.363		21,387.869		26,006.01	

[1] (sigma of β)
[2] *: significance at 10% level
[3] **: significance at 5% level
[4] ***: significance at 1% level

5.4 Confusion Matrix

Unlike continuous variables, in discrete choice, we not only consider whether the chosen routes are correctly predicted, but also whether the routes not taken are correctly identified. Therefore, we first use a confusion matrix to assess model performance.

5.4.1 Homogeneous Ensemble versus Base Model

For Multinomial Logit, Random Multinomial Logit, Path-size Logit, Random Path Size Logit, Decision Tree, and Random Forest, the average rank across 30 validations in case study 1 and 10 validations in case study 2 are presented in Figure 5.3. Clearly, all three homogeneous ensembles show a higher value in sensitivity. The improvement for Tree-based models is greater than that for the Logit models in both case studies. However, homogeneous ensembles do not outperform their base models for specificity. The Path-size Logit model ranks well in both case studies. Moving to precision, only RF improves the precision of its base model in general.

5.4.2 Heterogeneous Ensembles with Varying Rules

All heterogeneous ensembles include the same type and number of base models. They differ in their ensemble rules, which are the strategy for aggregating results from each base model to a final ensemble prediction as shown in Figure 5.3. Heterogeneous ensembles with stacking strategies, especially using SVM as the meta-learner, generally have lower rankings than voting strategies for sensitivity, specificity and precision. Based on the results, neither RF nor AdaBoost are a good option as a meta-learner, and they are still not as good as the voting strategies in these two case studies.

In case study 1, the heterogeneous ensemble with soft voting (ESV) performs better than those with hard voting (EHV) and ranked choice voting (ERV) in sensitivity, specificity, and precision. Similarly, in case study 2, soft voting (ESV) is better than hard voting (EHV) and

ranked choice voting (ERV) in sensitivity and specificity. However, hard voting (EHV) and ranked choice voting (ERV) are slightly higher than soft voting (ESV) in precision.

The average sensitivity, specificity, and precision of each tested model is shown in Figure 5.3 for CS1 and CS2. A dummy model which selects for the most frequent choice in the training set is also included in both case studies to provide a baseline for all models. Different symbols are used to distinguish between the types of models, as listed below:

- triangle: base models for homogeneous ensemble;
- star: homogeneous ensemble;
- square: base models for heterogeneous ensembles; and
- dot: heterogeneous ensemble

Based on the results of multiple random tests, none of the tested models dominates across sensitivity, specificity, and precision. All models perform better than the dummy model except the Decision Tree, which is worse than the dummy model in a few tests. Extra Tree and Random Forest, as tree-based homogeneous ensembles, place in the top 5 in the sensitivity list in two case studies and the precision list in case study 1. However, they do not perform well in specificity. MNL and PSL generally do very poorly in sensitivity, but PSL is located in the top 5 in the specificity list. For specificity, the Neural Network outperforms the remaining models in case study 1 and is in the top 5 in case study 2.

Comparing the heterogeneous ensembles with the base models, we see some base models perform better than heterogeneous ensembles.

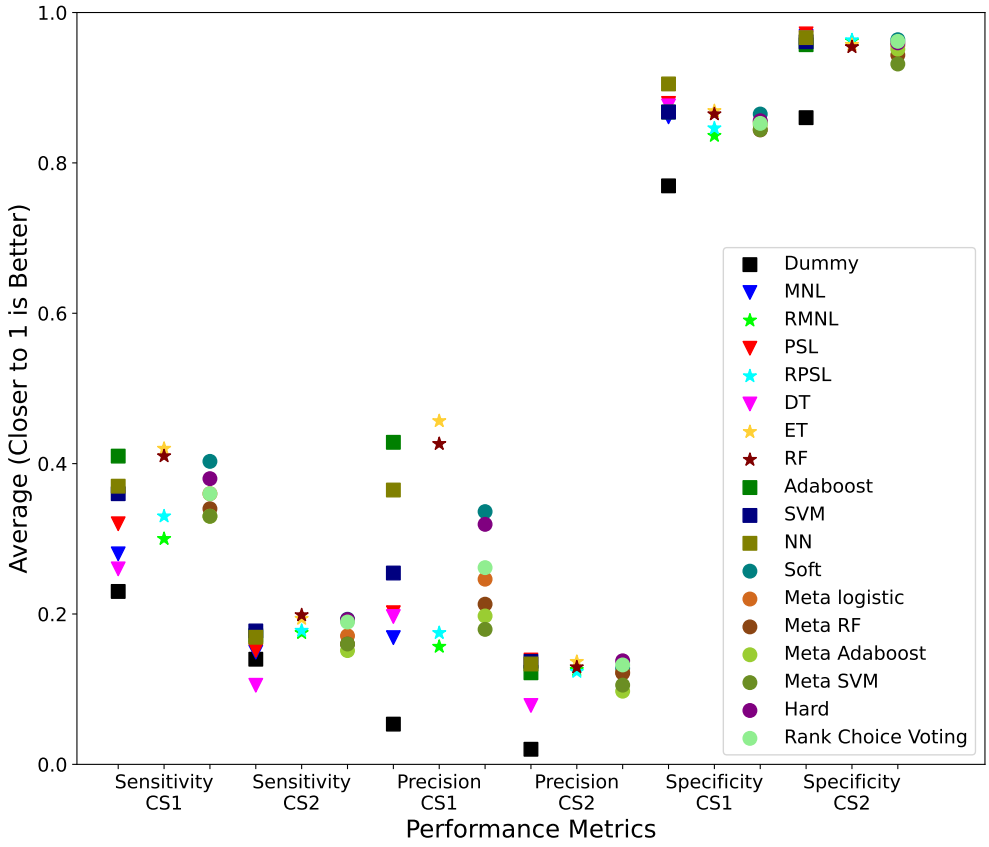
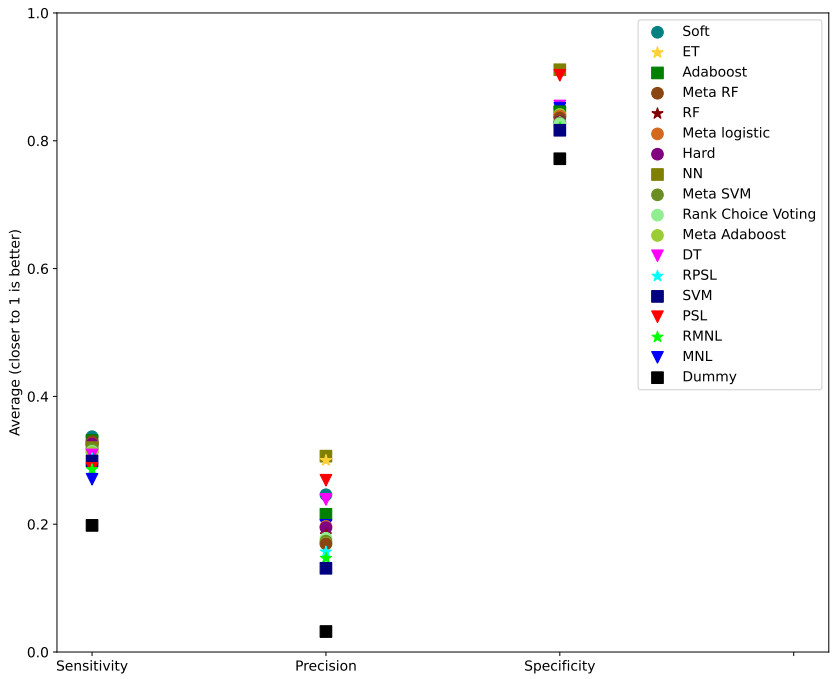


FIGURE 5.3: $Sensitivity = \frac{TP}{TP+FN}$, $Specificity = \frac{TN}{TN+FP}$, and $Precision = \frac{TP}{TP+FP}$

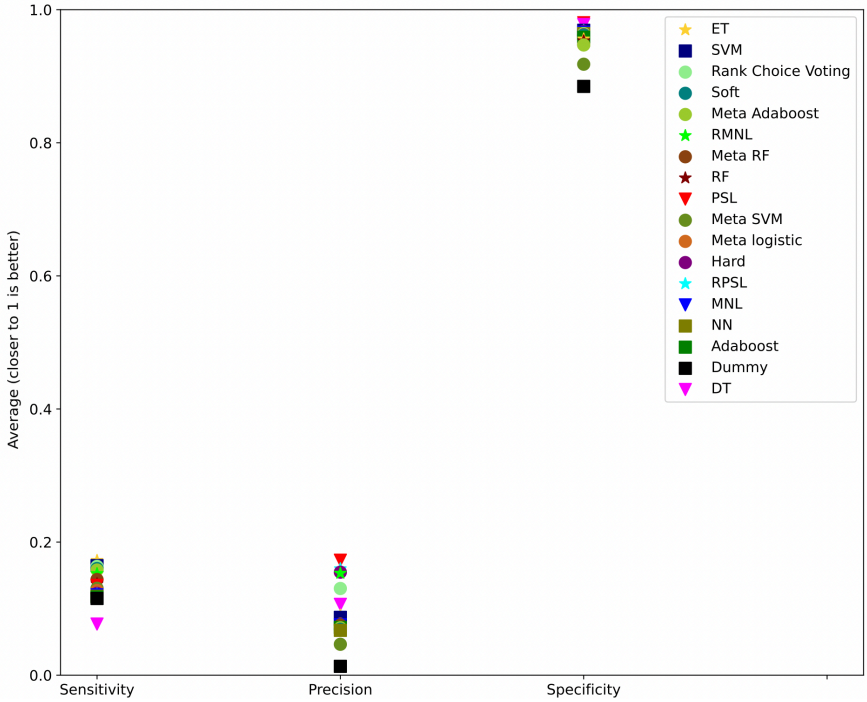
5.5 Cross-Validation

We use the models trained from one case study to model people's route choice in the other case study. The cross-validation results are shown in Figure 5.4.

As expected, model transferability loses some predictive power. Based on the results, using the models trained in one case study to predict the route choice behaviour in another case study performs worse than predicting in the same case study.



(A) Using Models in CS2 to Predict CS1



(B) Using Models in CS1 to Predict CS2

FIGURE 5.4: CS1 and CS2 Models Cross Validation

5.6 Log-likelihood

Whiel the confusion matrix is a popular evaluation criterion, it only focuses on the discrete alternatives and loses the predicted probabilities. Therefore, the total log-likelihood is used for model comparison. For each of the 30 random tests in Case Study 1 and the 10 random tests in Case Study 2, the total log-likelihood is calculated for each model. Box plots, as shown in Figure 5.5, are used to display the distribution of these total log-likelihoods.

In Figure 5.5 soft voting (ESV) has the best result. By comparing the three homogeneous ensembles (RMNL, RPSL, RF) with their base models (MNL, PSL, DT), homogeneous ensembles have less negative log-likelihood, which means the model performance is improved. Ensembles using stacking to combine results show better performance in log-likelihood than with the confusion matrix assessment. Especially for ensembles using logit and SVM as meta-learners, their log-likelihood just follows ESV in both studies. SVM shows a higher performance than other commonly-used machine learning models, the finding aligns with Sun and Park (2017). Extra Tree performs well in the confusion matrix evaluations but not in the log-likelihood measure.

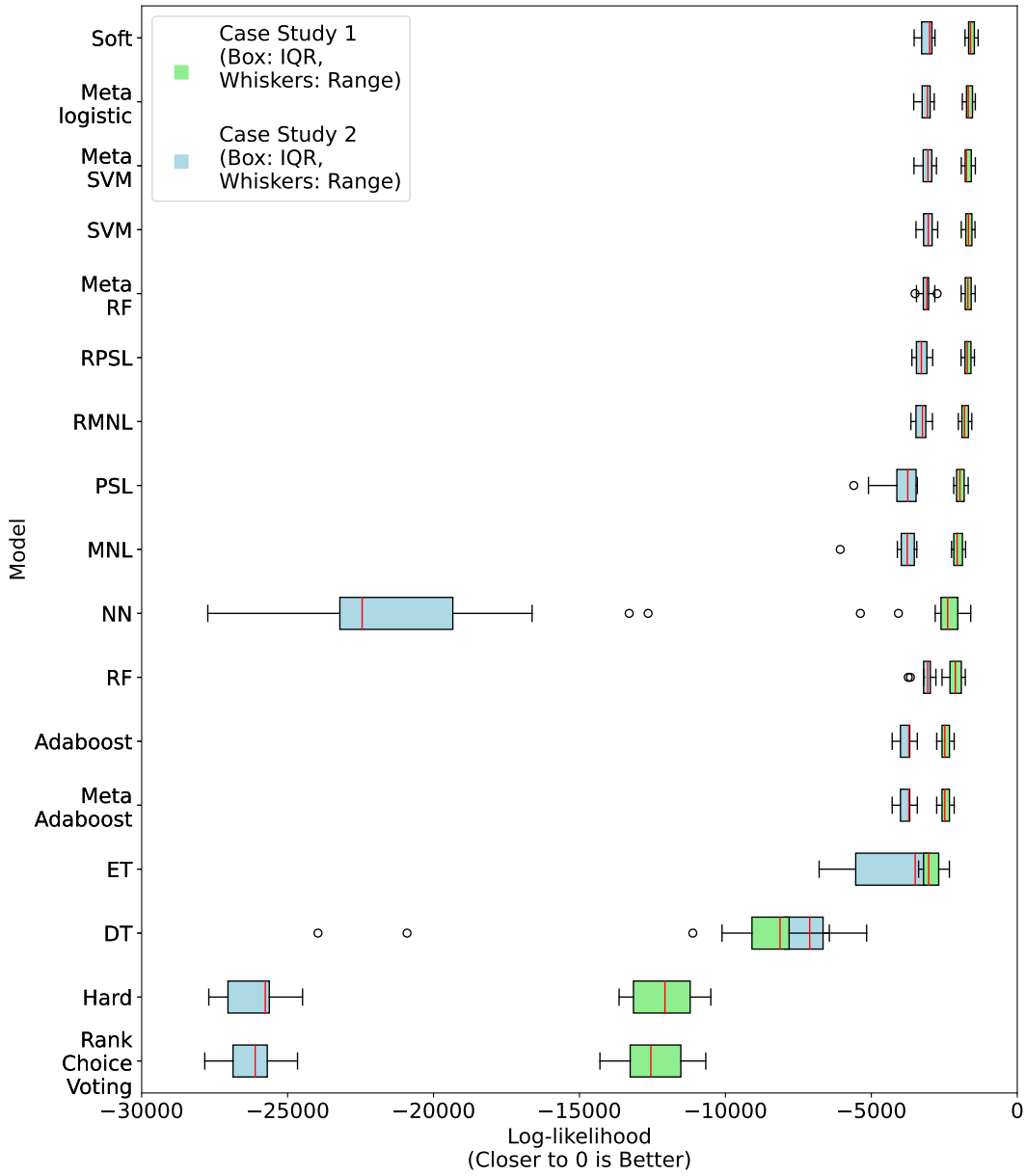


FIGURE 5.5: Log-likelihood in Case Study 1 and 2

5.7 Similarity

The proposed measure evaluates the similarity between the predicted and observed routes. As presented in chapter A, the similarity (S) will not be lower than the overlap measure (Ω) but provides additional credit for similarity in attributes and route location for the non-overlapped route elements.

Since this study aims to use similarity as an evaluation criterion, the attribute and spatial closeness are assumed to have the same level of importance. For each experiment, the average similarity between the predicted and observed routes are calculated, and Figure 5.6 shows the mean of the average similarity in 30 random tests for case study 1 and in 10 random tests for case study 2.

As presented in Figure 5.6, soft voting (ESV) shows the highest similarity. Random Forest and Extra Tree are in top 5 in both case studies, which is similar to the results when using a confusion matrix to assess model performance. Adaboost, which performs poorly based on log-likelihood evaluation and in the middle based on the confusion matrices, is in top 5 based on similarity. The similarity of the MNL and PSL base models in both case studies are higher than that of their homogeneous ensemble counterparts (RMNL and RPSL), but RF outperforms DT in both cases. Similar to the confusion matrix evaluations, heterogeneous ensemble using hard voting (EHV) and ranked choice voting (ERV) have middling performance.

The results from the similarity analysis mostly concur with the results from the confusion matrix but not the log-likelihood, because both similarity and the confusion matrix focus on one alternative (normally the one with the highest probability). The confusion matrix considers whether the predicted route matches the observed route and the similarity assessment shows how similar the predicted route is to the observed route. Neither of these evaluation techniques consider the probability of selecting any other alternative, but the log-likelihood assessment considers how well these models perform considering the probability of each prediction.

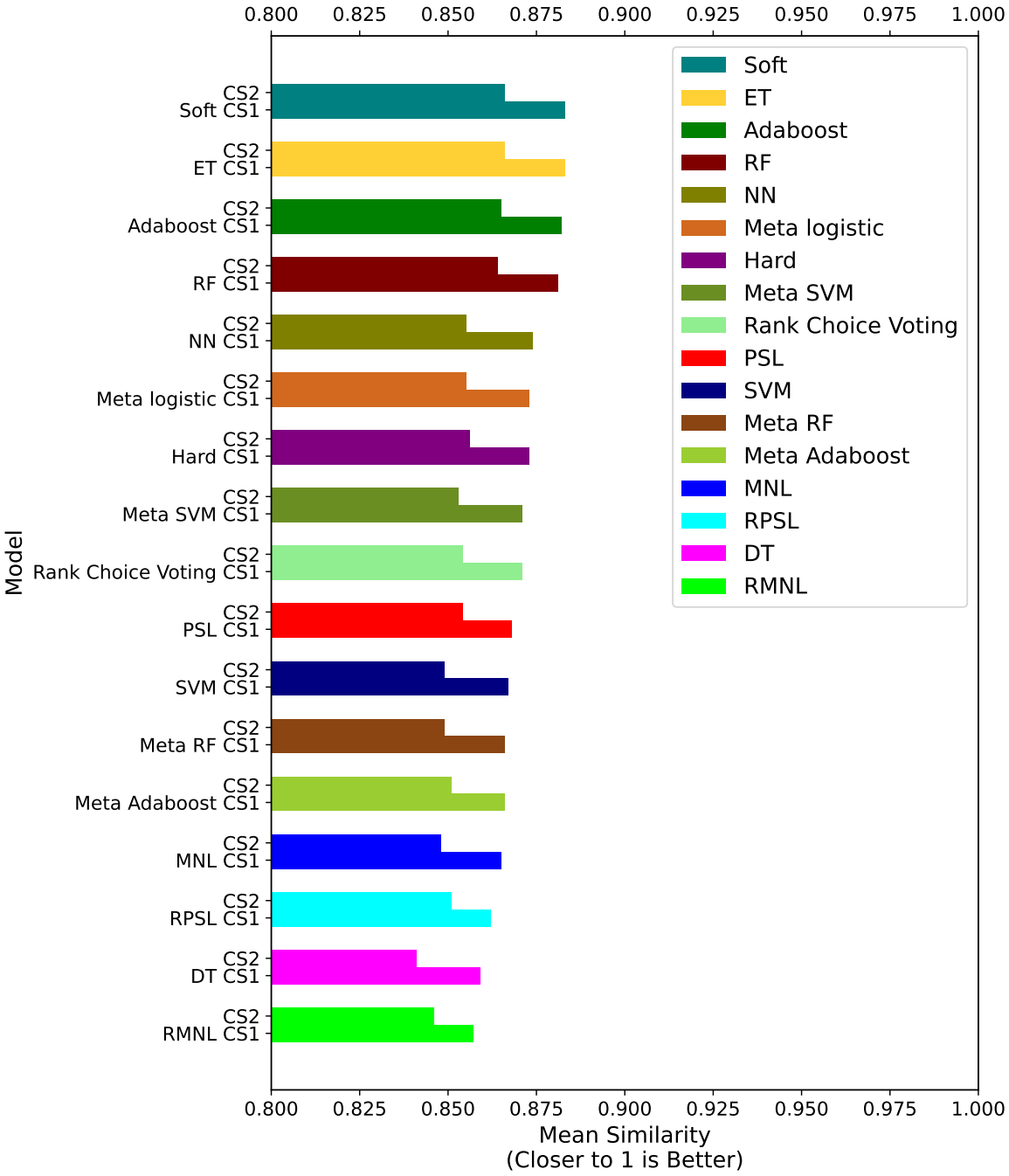


FIGURE 5.6: Mean Similarity of All Models

Model	Validation Set 1 $RMSE$	$MAPE$	E_{net}	Validation Set 2 $RMSE$	$MAPE$	E_{net}
MNL	3108	20.43	**4.56**	4001	28.25	*8.19*
RMNL	2869	19.05	8.03	3792	26.41	11.52
PSL	2947	24.32	19.36	3805	28.42	22.42
RPSL	2862	18.87	7.89	3792	26.41	11.52
DT	2803	19.97	15.57	3719	26.42	18.78
ET	2824	21.03	10.50	3732	26.51	13.90
RF	2812	21.01	10.62	3719	26.55	14.01
AB (AdaBoost)	3317	23.92	11.58	4236	30.76	14.94
SVM	**2712**	20.42	14.70	3717	26.01	17.94
NN	2918	20.14	*5.54*	3809	27.14	9.12
EHV (Hard)	2806	*17.84*	7.42	3754	26.53	**7.18**
ERV (Ranked Choice)	2798	19.03	7.79	*3715*	**25.60**	10.04
ESV (Soft)	*2793*	**16.29**	6.53	**3595**	*25.67*	15.83
Stacking-logit	2973	23.14	19.21	3858	28.18	22.28
Stacking-rf	2867	21.42	16.76	3767	27.04	19.92
Stacking-boost	3319	24.00	11.59	4238	30.78	14.94
Stacking-svm	2815	21.91	17.69	3696	26.79	20.82

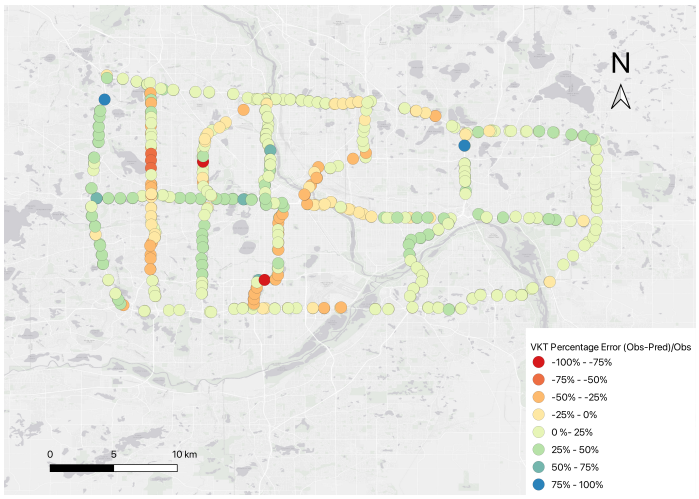
TABLE 5.7: Evaluation of mean predicted VKT on Freeway Links in LEHD Dataset and TBI Dataset. RMSE: Root Mean Square Error, MAPE: Mean Absolute Percentage Error, E_{net}: Network Percentage Error. Best performers denoted in **bold**. Second best denoted in *italics*.

5.8 Loop Detector Validation

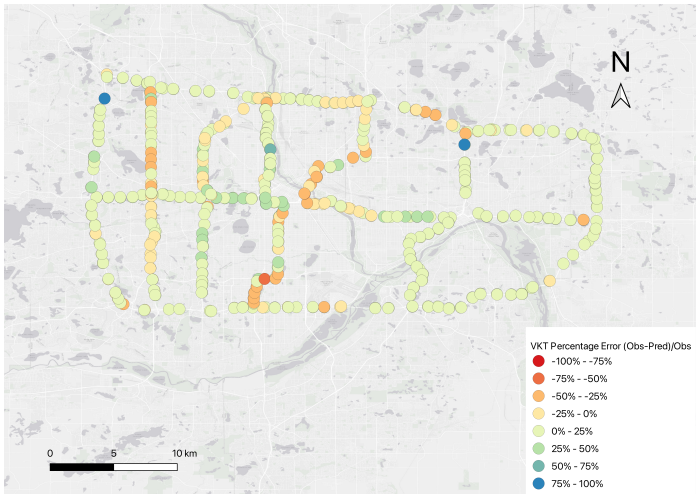
After comparing the models on the testing sets, they are validated using the VKT on freeway links from loop detectors. The models that are trained in case study 1 are used to predict the preferred route choice for each OD pair. Based on the probability of choosing each alternative route, for each simulation, the prediction for each OD pair for each model is randomly drawn from the probability distribution, and 10 simulations are completed for the two data sets. The evaluation results for the LEHD and TBI data sets are presented in Table 5.7.

In both validation data sets, while different measures show the best performance in each validation set, the heterogeneous ensemble using soft voting (ESV) shows good overall performance, and the best in 2 of the 6 columns (and second-best in 2 others).

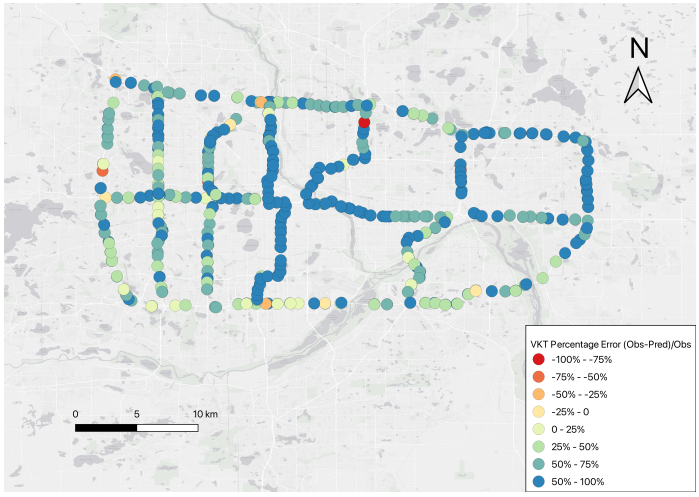
Figure 5.7 shows the percentage error at each loop detector station when using MNL, ensemble with soft voting (ESV), and the shortest path. The results for MNL and ESV are the best cases in the simulation, and a great improvement in accuracy can be found when comparing them to everyone taking the shortest median travel time path.



(A) VKT Percentage Error at Each Loop Detector station for MNL



(B) VKT Percentage Error at Each Loop Detector station for Soft Voting



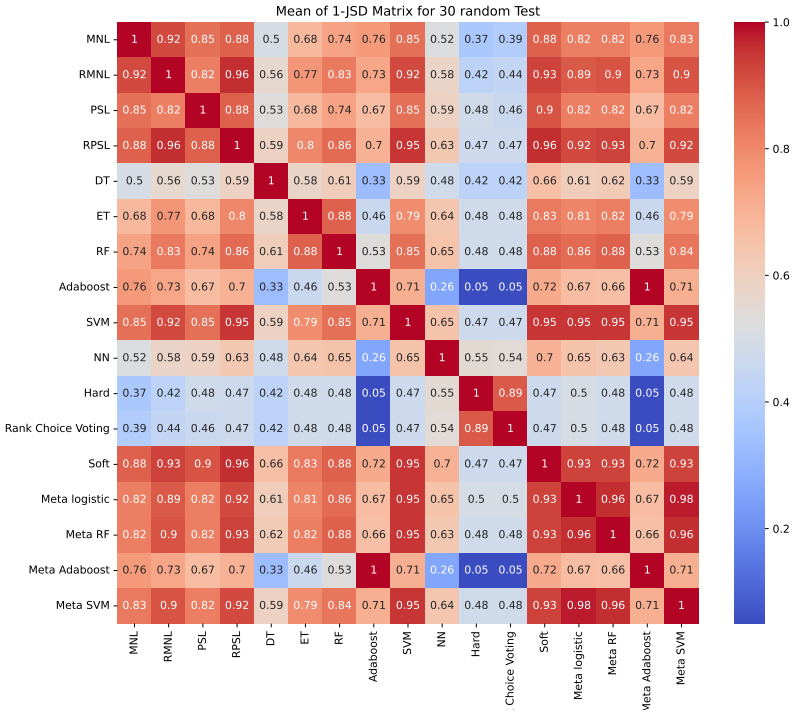
(C) VKT Percentage Error at Each Loop Detector station for Using Shortest Path

FIGURE 5.7: VKT Percentage Error at Each Loop Detector station

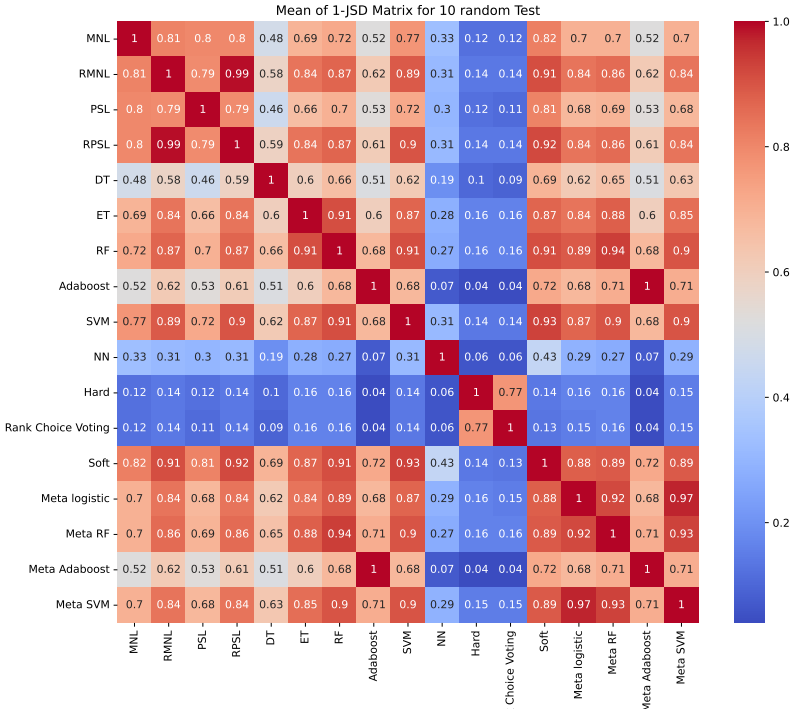
5.9 Cross-Model Correlation

The correlation between base models is assessed by examining the similarity of their predicted probabilities using the Jensen-Shannon Divergence (JSD). Originally, when $JSD = 0$, two distributions are identical, and two distributions are same if $JSD = 1$. However, since we want to explore the correlation between two models, we calculate $1 - JSD$ instead. The results are presented in Figure 5.8.

As expected, Extra Trees and Random Forest are both tree-based models and use similar ensemble techniques, resulting in high similarity between their predicted probability distributions. However, AdaBoost, another tree-based model that uses a different ensemble technique, provides dissimilar probability distributions. The logit models and the homogeneous ensemble based on logit show similar probability distributions, with RMNL and RPSL exhibiting the highest similarity among the tested models. The similarity between the Neural Network's probability distribution and those of other models is relatively low, especially in CS2. Heterogeneous ensembles using hard voting (EHV) and ranked choice voting (ERV) provide similar probability distributions, but since the methods for calculating probabilities for these two models differ from the others, their probability distributions are distinct. Conversely, the probability distribution of the heterogeneous ensemble using soft voting (ESV) shows relatively high similarity with the base models. For each base model, when calculating the similarity of probability distributions between that base model and others, the ensemble using soft voting (ESV) ranks in the top three for all base models. For ensembles using the stacking method, selecting AdaBoost as the meta-learner results in the lowest similarity of probability distributions compared to other models.



(A) Mean of 1-JSD for Tested Models in CS1



(B) Mean of 1-JSD for Tested Models in CS2

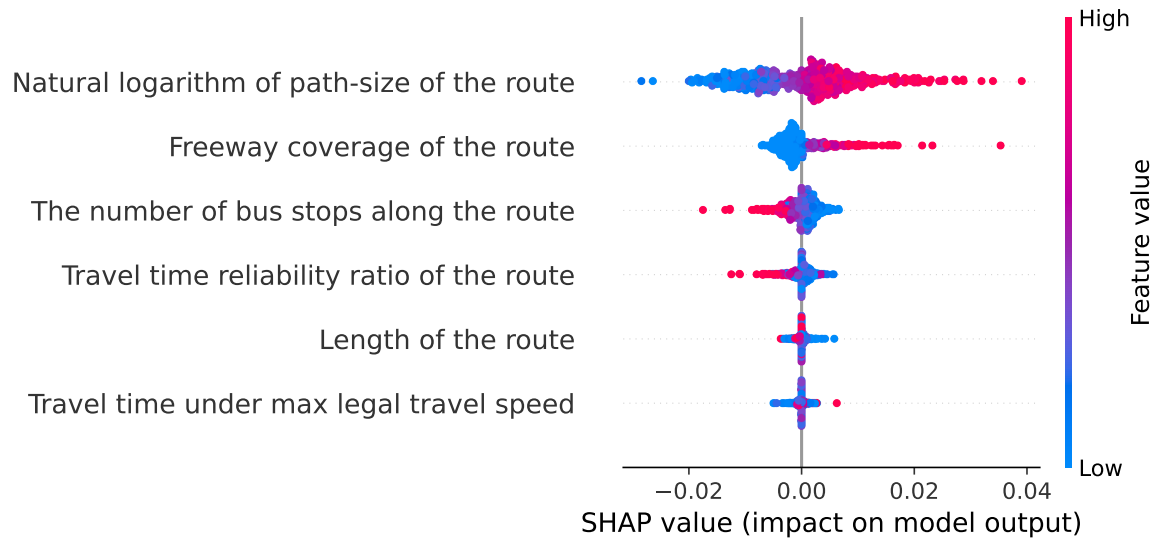
FIGURE 5.8: Mean of 1-JSD for Tested Models

5.10 SHAP Values

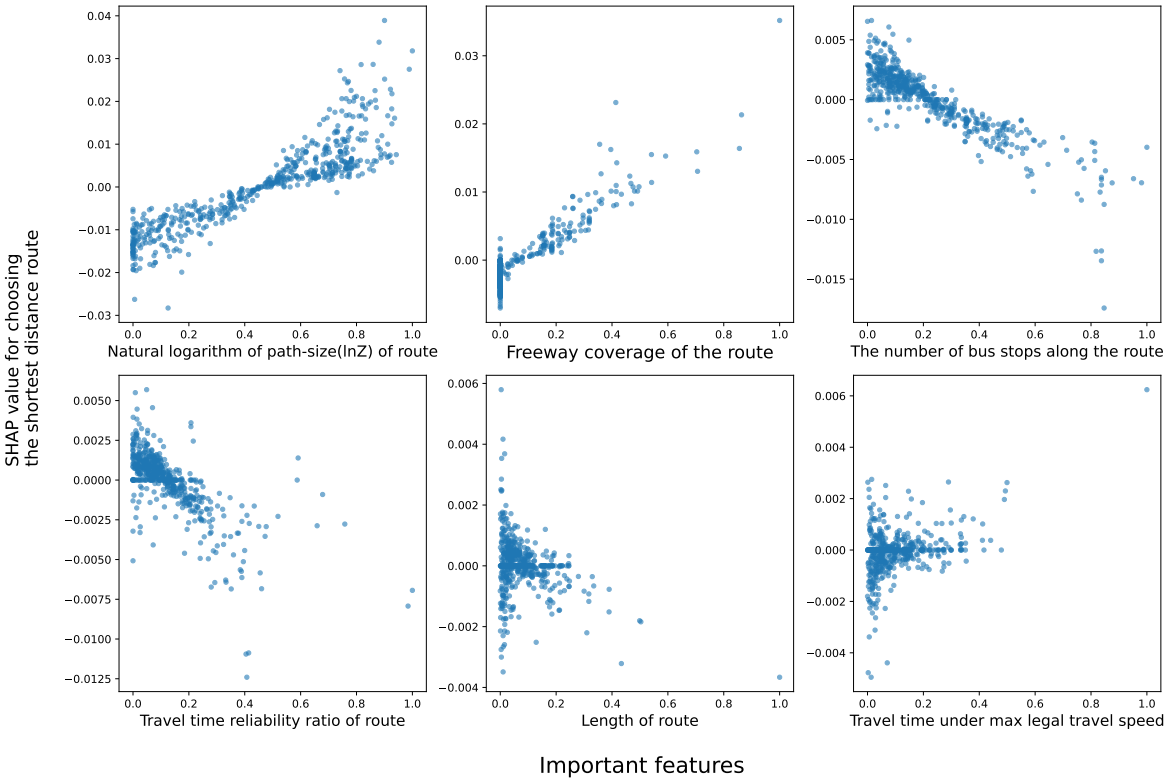
Based on the evaluation and validation results above, ESV generally performs well. Additionally, the best heterogeneous ensemble has been further explored for interpretability using SHAP values, and the results are presented below.

As introduced in section 4.8, the SHAP value for each feature is calculated based on each alternative. To illustrate the relationship between important features and the predicted probability of choosing a route, we select the shortest distance route and present the important features using a summary plot in Figure 5.9a. The vertical axis represents different features, while the horizontal axis shows the SHAP values. Each dot represents an instance, similar to the example in Table 4.10. Features with a wider spread of dots along the horizontal axis, such as the 'Natural logarithm of path-size of the route,' are considered more important to the model based on SHAP values. A positive SHAP value corresponds to an increased probability of choosing the route, while a negative value corresponds to a decreased probability. The color of the dots indicates the magnitude of the feature values, as shown by the color bar on the right. However, it is difficult to observe the exact feature values based on the color, making it challenging to see the relationship between features and model predictions. To address this, the SHAP dependency plots for these important features are also included in Figure 5.9b.

As shown in the top three subplots in Figure 5.9b, the dots follow a discernible trend. The first subplot, from left to right, shows the relationship between the natural logarithm of the path size term ($\ln(Z)$) and the probability of people choosing the shortest distance route. Generally, as $\ln(Z)$ increases, the probability of choosing the shortest distance route also increases. When $\ln(Z)$ is less than 0.5, it tends to discourage choosing the shortest distance route. The middle subplot in the top row indicates that higher freeway coverage increases the probability of selecting the shortest distance route. In the top right subplot, after standardizing the input features, the number of bus stops is normalized to fall between 0 and 1. When the number of bus stops along the route is close to 0, it contributes to selecting the shortest distance path, though its impact (measured by SHAP value) is less significant than that



(A) SHAP Summary of Shortest Path's Features When Choosing Shortest Path



(B) SHAP Dependency Plot of Important Features When Choosing Shortest Path

FIGURE 5.9: Ensemble Using Soft Voting's SHAP Summary of Choosing Shortest Path

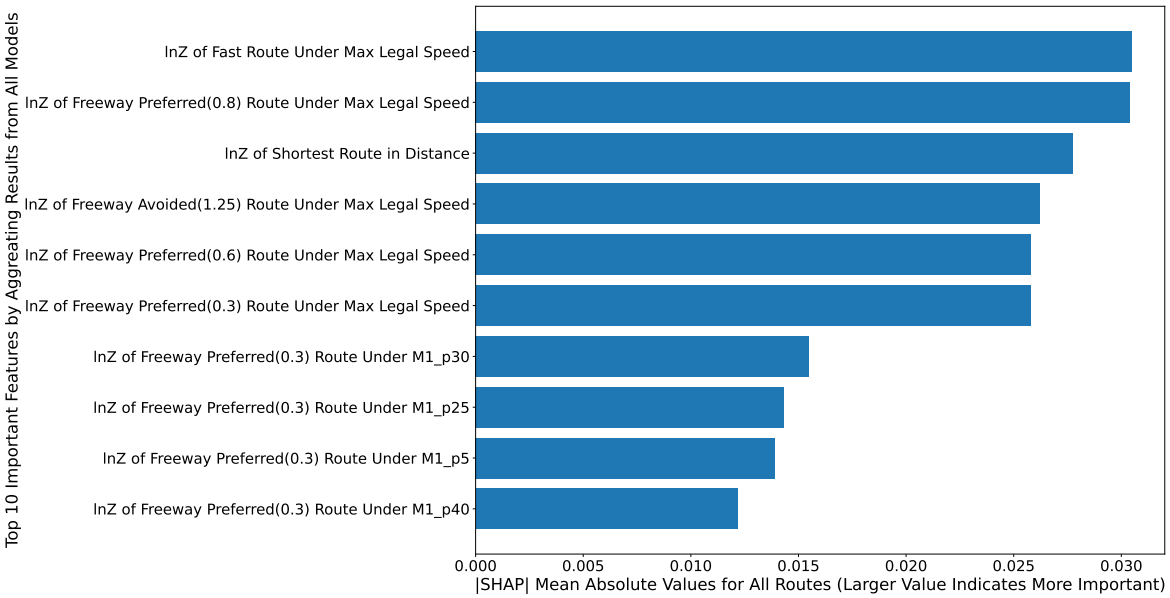
of $\ln(Z)$ and freeway coverage. As the number of bus stops increases, the probability of choosing the shortest distance path decreases.

In the bottom three subplots, the trends of the dots are less clear than the top three subplots. This is also reflected in Figure 5.9a; for features like 'Travel time reliability ratio of the route,' 'Length of the route,' and 'Travel time under max legal travel speed,' these features are less important than the top three features, and the SHAP values include both positive (red) and negative (blue) dots on both sides of the horizontal axis. Although the SHAP values for other features were calculated, they are not presented here, as they are either less important to the model or do not exhibit a clear trend with the prediction.

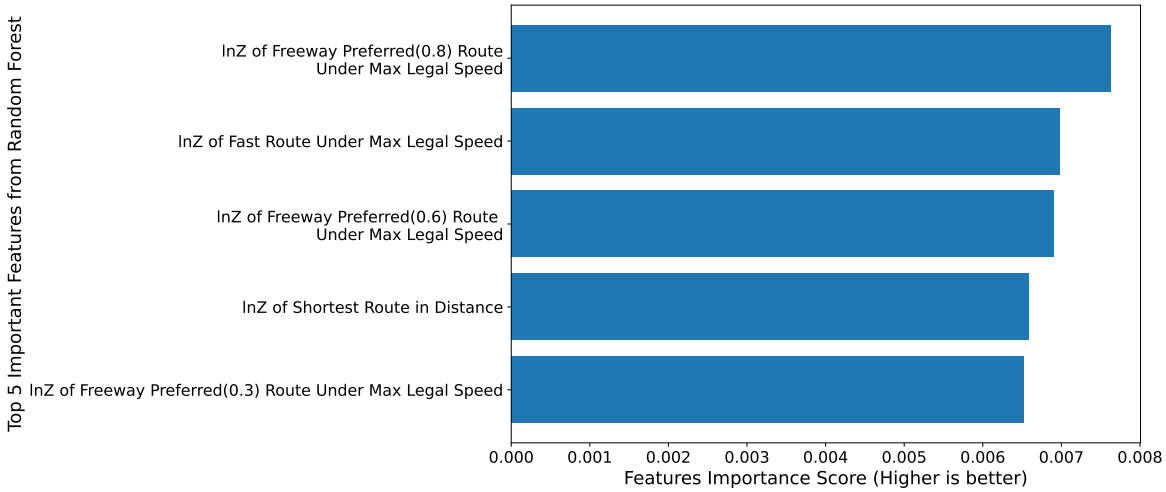
We attempted to summarize the SHAP values of the same feature across all alternative routes, but no clear trend or pattern emerged. This is because the impact of the same feature can vary across different alternative routes, even for the same person, as illustrated in Table 4.10. Additionally, since the features are specific to the alternatives (e.g., the length of a particular route), the important features identified by the model are meaningful for its predictions but may not aid our understanding. Therefore, rather than showing the relationship between features and model predictions through dependency plots, we emphasize the important features for ESV, based on the mean absolute SHAP values. These important features are then compared with those identified using Gini Importance from the Random Forest homogeneous ensemble.

To achieve this, we first identified the alternative with the highest probability for each origin-destination (OD) pair and then determined the top 10 most important features across all tested models. Upon reviewing the top 10 important features for all tested models, we found that path-size terms of different alternatives appeared most frequently (103 out of 150 instances). Specifically, 11 out of 15 models identified the path-size term of the 'fast route under max legal speed' as one of the top 10 important features.

By aggregating the mean absolute SHAP values across all tested models, the overall top 10 important features are shown in Figure 5.10a, with the path-size terms of various alternative routes dominating the list. Without calculating SHAP values, the Random Forest can provide feature importance by calculating the Mean Decrease Impurity (Gini Importance). The results from the Random Forest are presented in Figure 5.10b. By comparing the results in Figure 5.10, the path-size terms of the 'fast route under max legal speed' and 'freeway-preferred route under max legal speed with factor=0.8' emerge as the two most important



(A) Top 10 Important Features for All Models Based on Mean Absolute SHAP Value



(B) Top 5 Important Features Based on Decrease in Impurity in Random Forest

FIGURE 5.10: Important Features

features based on both mean absolute SHAP value and Gini Importance from the Random Forest. Additionally, although the ranking of features differs, both mean absolute SHAP value and Gini Importance highlight the same four important features within the top 5.

CHAPTER 6

Discussion and Conclusions

In this chapter, we discuss choice set generation, continuity vs discreteness of route choice behavior prediction, travelers' route choice prediction and validation, and model interpretation of modelling route choice with ensembles in section 6.1, section 6.2, section 6.3, and section 6.4. In section 6.5, we summarize the contributions of this dissertation. The limitations of this study are discussed in section 6.6 and recommendations for future studies in section 6.7.

6.1 Choice Set Generation

For dealing with a high-resolution road network and GPS trajectories, generating a parsimonious choice set that can capture the most trajectories across an observed trip set is difficult. Many factors influence route choices, and some of them are unknown to the modeller or even to the traveller themselves. Therefore, generating routes that are the same as the observed trips is challenging. Relying on only one choice set generation method is not sufficient because, as more links and nodes are included in the network, the number of possible routes between an OD pair rises. However, the high similarity of the high-resolution graph with the real-life network suggests the analysis results are more useful.

According to the results presented in Table 5.2, alternatives based on the least travel time match more observed trips than those based on the shortest distance. For most drivers, the travel time might be more important than distance when they plan morning trips. Moreover, for both free-flow and congested travel times, multiplying freeway links by a weighting factor ($w_f = 0.8$) when applying the shortest path algorithm improves the capture rate (more than 50% and 40% for 80% and 90% overlap threshold), as more drivers prefer freeways than

surface streets. We also observe that assuming the link travel time is perfectly independent (M2) captures more observed trips than assuming travel time is perfectly correlated (M1) or free-flow travel time. This is because the alternative routes which are found based on the M2 scenario are more similar to observed routes.

A list of freeway time multipliers (weighting freeway times vs. non-freeway times) is applied to freeway links on both free-flow and observed conditions, and factors within 0.8 to 1.0 normally capture more observed trips under almost all overlap thresholds (i.e. freeway time is less onerous than non-freeway time). With the same number of iterations and same multipliers, compared to the result when assuming link travel time is perfectly correlated, more unique routes could be generated if we assume travel time for each link is perfectly independent, and more observed trips could be captured by the alternative routes. By using the best 10 labels to generate, on average, 40 unique routes for each OD pair, around 80% of observed trips could be captured under 80% overlap threshold ($O = 80\%$). Combining all labels in this study, the generated choice set could capture 81%, 71%, 44% of observed trips with overlap threshold set to 80%, 90% and 100% respectively.

Compared with the choice set in the other studies in Table 2.1, we achieve almost the same capture rate (80%) as Bekhor et al. (2006), on a much larger network (108,561 nodes and 277,747 links vs 13,000 nodes and 34,000 links). For the same level of network size, compared with the study from Rieser-Schüssler et al. (2013), we gain a higher capture rate (around 80% vs 75%) with fewer alternative routes for each OD pair in the choice set (40 vs 100). The choice set generation results are improved compared with previous studies. Therefore, instead of simplifying the road network, using the proposed approach can improve the choice set without losing any network fidelity.

6.2 Continuity vs Discreteness in Route Choice

Routes comprise a series of links, and links can be shared between routes. This property distinguishes route choice from other discrete choice problems. This study attempts to make the dependent variable continuous by modeling the overlap and average deviation between

the alternative routes and observed routes using a panel regression model. With the same initial attributes, a random effects model is used to predict the overlap between alternative and observed routes, achieving an R^2 of 0.45. Alternative routes with less distance, a higher percentage of freeway, high free-flow speed, and fewer traffic lights are more likely to overlap with observed trips and be selected. Based on the R^2 of a random effects model, the overlap rate is more easily predicted as an independent variable than the average deviation between alternative routes.

Moreover, the Path-size Logit model and mixed logit model are used to model route choice under a discrete choice framework. Based on the models' statistical evaluation, a mixed logit model with a path-size term shows better log-likelihood, AIC, and BIC than the path-size MNL model and the mixed logit model without that term. Results indicate that routes with a higher freeway percentage and more bus stops, shorter trip lengths, and fewer traffic lights, left turns, and right turns are more likely to be chosen. These findings are similar to the results of the random effects model.

However, from an application perspective, it might not be the 'better' model. As described in the chapter 4, the coefficients $\boldsymbol{\beta}$ in the mixed logit model are assumed to follow a normal distribution, and thus, the coefficients of the mixed logit model in Table 5.6 are the mean values.

For example, for $\beta_{\text{Trafficlight}} = -0.025$ and $\sigma_{\text{Trafficlight}} = 0.086$, it means 39% ($P(x) > 0$) of the population β has the opposite sign to the mean value (-0.025), indicating it is hard to form a confident conclusion. The Path-size Logit model does not consider the heteroskedasticity in the panel data. Testing whether the coefficient $\boldsymbol{\beta}$ in the mixed logit model follows another distribution might be helpful in future studies. For this study, we employ a machine learning approach.

6.3 Traveller's Route Choice Prediction and Validation

According to the results from both case studies, for all evaluation models, no model dominates every validation. It follows our expectations that the 'best' model is hard to find or may not even exist.

Sensitivity is the percentage of routes actually selected that were predicted to be selected. The 'tree-based' homogeneous ensembles and the heterogeneous ensemble with soft voting rank better. The top 5 models in the sensitivity evaluation are either heterogeneous or tree-based homogeneous ensembles. However, the heterogeneous ensembles with different stacking methods do not perform as well as other ensembles. It might be that training meta-learning in stacking requires a larger data set to make an accurate prediction. Sensitivity measures the percentage of actually selected routes that the model correctly predicts. Based on the results for sensitivity, soft voting is a good choice.

Specificity indicates the percentage of routes not selected that are predicted correctly. The Neural Network and PSL models show better performance in general. Similar to sensitivity, heterogeneous ensembles using stacking strategies generally have a worse ranking than those using voting strategies. Considering the low performance of MNL and PSL in sensitivity, these models attend more to getting a high value of true negatives (TN) rather than obtaining a larger value of the more important (and scarcer) true positives (TP). In other words, they tend to avoid missing the untaken routes in all abandoned routes and the type I error.

For precision, which measures percentage of routes selected that are predicted correctly, base models including ET, RF, AdaBoost and NN also perform relatively well in both case studies. In the context of route choice, the false positive (FP) is not a critical issue. This is because FP refers to the routes not taken but incorrectly predicted as chosen. When considering this on a link level rather than a route level, it's possible that untaken routes share some links with the actual taken routes. This makes the consequence of FP less severe than, for example, incorrectly diagnosing a patient with a disease. The significance of the three criteria follows this order: sensitivity is more important than precision, which is more important than specificity.

Beyond confusion matrix evaluations, the similarity measure is introduced to capture correct predictions and quantifies the dissimilarity between incorrect route segments in both space and attributes. Averaging the similarity measure across all random tests in both case studies, the heterogeneous ensemble with soft voting generally performs better. Random forest and Extra Tree models also exhibit high similarity with observed routes. Similar to precision evaluation, except for Random Forest, RMNL and RPSL perform worse than their base models. The ranking of the remaining models aligns with their sensitivity rankings.

Many tested models, including MNL, PSL, SVM, RF, and NN, as well as ensembles with soft voting (ESV), provide probabilities for each alternative. However, converting these probabilistic results into deterministic ones by simply choosing the alternative with the highest probability loses valuable information. Consequently, this study also evaluates the log-likelihood of each model, revealing different results compared to the confusion matrix and similarity measures. ESV emerges as the best model in terms of mean log-likelihood.

In addition, the morning auto-based route choices of all commuters in the Twin Cities are predicted using the same models from the case studies. By comparing the observed Vehicle Kilometers Traveled (VKT) from loop detectors with the predicted VKT from 10 simulation runs in the validation dataset, the ESV generally outperforms other models.

Regarding the performance metrics, RMNL and RPSL exhibit better RMSE and MAPE than MNL and PSL, but RF and DT models show comparable performance in both validation datasets. Based on these validation results, among the four tested meta-learners, SVM emerges as the best meta-learner for this case.

In ensembles employing voting strategies, assuming equal importance for all base models can sometimes lead to the base models negatively impacting overall performance. This issue is particularly pronounced in hard voting (EHV), where the minimum frequency of the most frequent result is $\frac{2}{T}$ for classification into T classes. As the number of classes increases, this minimum frequency becomes smaller, making reliance on the most frequent result a risky strategy.

The ranked choice voting (ERV) strategy offers some advantages, as the frequency value of the most frequent result is always greater than 0.5. This implies that more than half of the base models agree on the same prediction.

ESV generally outperforms other voting strategies because they take into account the probabilities provided by the base models. For instance, if one base model indicates a 99% chance of choosing route *A* and another model suggests a 51% chance for route *B*, in practice, route *A* is more likely to be chosen due to its higher probability. However, if a 'weak' model gives an incorrect prediction with high confidence, it can negatively impact the ensemble's performance. As a result, giving more weight to the more accurate models in the ensemble prediction can mitigate the adverse effects of these 'weak' base models, as discussed in the Appendix and illustrated in Figure B0.1 and Figure B0.2, however this may not be as robust for model-transferability.

6.4 Model Interpretation

SHAP values provide a tool to interpret machine learning models and have been demonstrated to be helpful in understanding route choice behavior for alternative routes. By using SHAP values, features with higher mean absolute SHAP values are more important to the model. When focusing on choosing the shortest distance route, as $\ln(Z)$ and freeway coverage increase and the number of bus stops along the route decreases, the probability of choosing the shortest distance route increases. By aggregating the SHAP values for the selected alternative routes, we identified the important features for the ESV and compared them to the important features based on Gini Importance from the RF model. The top five important features show high consistency. However, these features are specific to the alternatives and are challenging to interpret. Overall, using SHAP values to interpret ensembles and the tested machine learning models has been successful, but the interpretation of machine learning models still remains more difficult than statistical models.

6.5 Contributions

There are three main contributions of this dissertation:

(1) Developing a hybrid route choice generation method without removing any non-cyclic links, on a large, high-resolution network;
(2) Developing a new similarity measure between routes by considering the influence of overlap, attribute similarity, and spatial similarity; and
(3) Developing a framework for applying ensemble techniques to predict individual and subsequently aggregate route choices and assessing the performance of ensembles on different data sets.

For the first component, we present the methodology to combine labeling and link penalty techniques and incorporate both speed limits and congested travel speeds to generate a choice set, as described in section 4.2. Using the 10 labels shown in Figure 5.2, we demonstrate that our method, using fewer alternative routes, captures more observed trips compared to other studies with the same network size in subsection 4.2.3.

The second component involves the methodology to determine how similar two routes are based on their overlap, attribute similarity, and spatial similarity, as described in section 4.6. A simple example involving four routes on a grid network is presented in chapter A, and the similarity between predicted routes and chosen routes is shown in Figure 5.6. When assessing route choice models based on the similarity between predicted routes and chosen routes, the outcomes of the proposed similarity measure largely align with those from the confusion matrix.

The last component covers the methodology for constructing homogeneous and heterogeneous ensembles using various techniques, the evaluation of different route choice models based on diverse criteria, and the validation and assessment of the trained route choice models in predicting all people's commute trips, as described in section 4.3 and section 4.5. According to our results in chapter 5, after reviewing and analyzing the limitations of current choice models, we conclude that predicting individuals' route choices is difficult. However, the

heterogeneous ensemble using soft voting (ESV) achieves better results than the traditional logit model and some popular machine learning models, such as SVM and RF. Through multiple random tests on different datasets, we show that the 'best' route choice model varies with the training and testing data, and no model consistently dominates. By comparing the performance of base models and ensembles, we find that homogeneous ensembles using bagging and feature randomness improve sensitivity and log-likelihood. Additionally, using soft voting to combine results from different types of models outperforms other voting techniques in both prediction accuracy and robustness in network-wide validation. Using logit as a meta-learner improves the log-likelihood and, in some cases, provides high sensitivity and similarity, though performance varies across random tests. Therefore, ESV, which achieves the best performance in sensitivity and log-likelihood, scores near the best in precision, and surpasses other ensemble methods, is recommended for continued study.

6.6 Limitations

This research also has some limitations. First, when testing the performance of the proposed ensembles for route choice, only data sets from the Twin Cities are used. We didn't assess how the trained models perform on data sets from different cities or countries. Second, the LEHD data includes home and workplace locations but does not contain information about the transport mode individuals took or when they departed from home. Therefore, we estimated the proportion of morning car-based commute trips based on the TBI data set. However, if that information were available to us, we expect we could achieve a more accurate prediction of VKT on freeway links.

6.7 Recommendations for Future Research

Future research should compare route-based ensembles and link-based inverse reinforcement learning. Such a comparison could offer valuable insights into the relative strengths and weaknesses of these methodologies, particularly in complex urban environments. In addition, as new models and techniques emerge in the field of route choice modeling and machine

learning, future research should consider incorporating these advancements into ensemble frameworks. Evaluating their effectiveness both as standalone models and as components of heterogeneous ensembles could further enhance predictive accuracy and robustness.

The study aims to explore the potential benefits of ensemble techniques in route choice modeling. The majority of the tasks were completed on a Mac desktop (M1 Chip) and a Windows desktop (13th Gen Intel(R) Core(TM) i5-13600K). The whole modeling process needs a considerable amount of time, which might not be suitable for direct implementation in cases where computation time is the main concern. Future studies should aim to improve the efficiency of the whole modeling process.

The hybrid route choice generation method without removing any non-cyclic links should be evaluated on a large, high-resolution network across various cities, both within the United States and internationally. By testing this method in different urban environments, researchers can assess its adaptability and effectiveness in capturing realistic travel behaviors in diverse contexts. Conducting such studies on a global scale would not only validate the robustness of the method but also highlight the potential difference in choice behavior (e.g., different combination of labels in choice set for different countries). Ultimately, this could lead to the development of more universally applicable route choice models that are better suited to the specific needs of urban planners and decision-makers worldwide.

Future studies should also explore the application of the proposed similarity measure within various route choice models, such as a replacement for the path-size term in Path-size Logit. This measure offers a comprehensive approach to evaluating how similar routes are by considering not only the shared links between routes but also how closely the routes align in terms of their characteristics and spatial positioning. By incorporating this similarity measure into future models, researchers can better capture the subtle differences between alternative routes, leading to more accurate predictions and a deeper understanding of route choice behavior. This approach could be particularly beneficial when evaluating route choices in densely connected networks or urban environments where slight deviations in a route can significantly impact overall travel time and user preference.

Bibliography

Armstrong, J Scott (2001). 'Combining forecasts'. In: *Principles of Forecasting: A handbook for Researchers and Practitioners*, pp. 417–439.

Azevedo, JoseAugusto et al. (1993). 'An algorithm for the ranking of shortest paths'. In: *European Journal of Operational Research* 69.1, pp. 97–106.

Bekhor, S and JN Prashker (2001). 'Stochastic user equilibrium formulation for generalized nested logit model'. In: *Transportation Research Record* 1752.1, pp. 84–90.

Bekhor, Shlomo, Moshe E Ben-Akiva and M Scott Ramming (2006). 'Evaluation of choice set generation algorithms for route choice models'. In: *Annals of Operations Research* 144.1, pp. 235–247.

Bekhor, Shlomo and Carlo Giacomo Prato (2009). 'Methodological transferability in route choice modeling'. In: *Transportation Research Part B: Methodological* 43.4, pp. 422–437.

Bellman, Richard and Robert Kalaba (1960). 'On k th best policies'. In: *Journal of the Society for Industrial and Applied Mathematics* 8.4, pp. 582–588.

Ben-Akiva, Moshe and Michel Bierlaire (1999). 'Discrete choice methods and their applications to short term travel decisions'. In: *Handbook of Transportation Science*, pp. 5–33.

Ben-Akiva, Moshe et al. (1984). 'Modelling inter urban route choice behaviour'. In: *Papers Presented during the Ninth International Symposium on Transportation and Traffic Theory Held in Delft the Netherlands, 11-13 July 1984.*

Bernardi, Silvia, Lissy La Paix Puello and Karst Geurs (2018). 'Modelling route choice of Dutch cyclists using smartphone data'. In: *Journal of Transport and Land Use* 11.1, pp. 883–900.

Bliemer, Michiel CJ and Piet HL Bovy (2008). 'Impact of route choice set on route choice probabilities'. In: *Transportation Research Record* 2076.1, pp. 10–19.

Bovy, Piet H. L and Sascha Hoogendoorn-Lanser (2005). 'Modelling route choice behaviour in multi-modal transport networks'. eng. In: *Transportation (Dordrecht)*. Transportation 32.4, pp. 341–368. ISSN: 0049-4488.

Bovy, Piet H.L and Stella Fiorenzo-Catalano (2007). 'STOCHASTIC ROUTE CHOICE SET GENERATION: BEHAVIORAL AND PROBABILISTIC FOUNDATIONS'. eng. In: *Transportmetrica* 3.3, pp. 173–189. ISSN: 1812-8602.

Bovy, Piet HL, Shlomo Bekhor and Carlo Giacomo Prato (2008). 'The factor of revisited path size: Alternative derivation'. In: *Transportation Research Record* 2076.1, pp. 132–140.

Brathwaite, Timothy and Joan L Walker (2018). 'Asymmetric, closed-form, finite-parameter models of multinomial choice'. In: *Journal of Choice Modelling* 29, pp. 78–112.

Breiman, Leo (1996a). 'Bagging predictors'. In: *Machine Learning* 24, pp. 123–140.

— (1996b). 'Bias, variance, and arcing classifiers'. In.

Broach, Joseph, Jennifer Dill and John Gliebe (2012). 'Where do cyclists ride? A route choice model developed with revealed preference GPS data'. In: *Transportation Research Part A: Policy and Practice* 46.10, pp. 1730–1740. ISSN: 0965-8564.

Cantarella, Giulio Erberto and Stefano de Luca (2005). 'Multilayer feedforward networks for transportation mode choice analysis: An analysis and a comparison with random utility models'. In: *Transportation Research Part C: Emerging Technologies* 13.2, pp. 121–155.

Cascetta, Ennio et al. (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*.

Cheng, Long et al. (2019). 'Applying a random forest method approach to model travel mode choice behavior'. In: *Travel Behaviour and Society* 14, pp. 1–10.

Cheng, Long et al. (2020). 'Applying an ensemble-based model to travel choice behavior in travel demand forecasting under uncertainties'. In: *Transportation Letters* 12.6, pp. 375–385.

Cohn, Nick (2009). 'Real-time traffic information and navigation: An operational system'. In: *Transportation research record* 2129.1, pp. 129–135.

Craig, William (2005). 'White knights of Spatial Data Infrastructure: The role and motivation of key individuals'. In: *URISA journal* 16.2, pp. 5–13.

Cui, Boer et al. (2022). 'All ridership is local: Accessibility, competition, and stop-level determinants of daily bus boardings in Portland, Oregon'. In: *Journal of Transport Geography* 99, p. 103294.

Cui, Mengying and David Levinson (2018). 'Accessibility and the Ring of Unreliability'. In: *Transportmetrica A: transport science* 14.1-2, pp. 4–21.

Darwish, Ahmed Mahmoud et al. (2024). 'Sensitivity evaluation of machine learning-based calibrated transportation mode choice models: A case study of Alexandria City, Egypt'. In: *Transportation Research Interdisciplinary Perspectives* 24, p. 101052.

De La Barra, Tomas, B Perez and J Anez (1993). 'Multidimensional path search and assignment'. In: *PTRC Summer Annual Meeting, 21st, 1993, University of Manchester, United Kingdom*.

Di, Xuan and Henry X Liu (2016). 'Boundedly rational route choice behavior: A review of models and methodologies'. In: *Transportation Research Part B: Methodological* 85, pp. 142–179.

Di, Xuan et al. (2017). 'Indifference bands for boundedly rational route switching'. In: *Transportation* 44.5, pp. 1169–1194.

Ding, Jing et al. (2014). 'Routing policy choice set generation in stochastic time-dependent networks: Case studies for Stockholm, Sweden, and Singapore'. In: *Transportation Research Record* 2466.1, pp. 76–86.

Duncan, Lawrence Christopher et al. (2020). 'Path Size Logit route choice models: Issues with current models, a new internally consistent approach, and parameter estimation on a large-scale network with GPS data'. In: *Transportation Research Part B: Methodological* 135, pp. 1–40.

Eppstein, David (1998). 'Finding the k shortest paths'. In: *SIAM Journal on Computing* 28.2, pp. 652–673.

Ferreira, Artur J and Mário AT Figueiredo (2012). 'Boosting algorithms: A review of methods, theory, and applications'. In: *Ensemble Machine Learning: Methods and Applications*, pp. 35–85.

Fosgerau, Mogens, Emma Frejinger and Anders Karlstrom (2013). 'A link based network route choice model with unrestricted choice set'. In: *Transportation Research Part B: Methodological* 56, pp. 70–80.

Frejinger, Emma and Michel Bierlaire (2007). 'Capturing correlation with subnetworks in route choice models'. In: *Transportation Research Part B: Methodological* 41.3, pp. 363–378.

Frejinger, Emma, Michel Bierlaire and Moshe Ben-Akiva (2009). 'Sampling of alternatives for route choice modeling'. In: *Transportation Research Part B: Methodological* 43.10, pp. 984–994.

Freund, Yoav and Robert E Schapire (1997). 'A decision-theoretic generalization of on-line learning and an application to boosting'. In: *Journal of Computer and System Sciences* 55.1, pp. 119–139.

Gao, Song, Emma Frejinger and Moshe Ben-Akiva (2011). 'Cognitive cost in route choice with real-time information: An exploratory analysis'. In: *Procedia-Social and Behavioral Sciences* 17, pp. 136–149.

Ghanayim, Muhammad and Shlomo Bekhor (2018). 'Modelling bicycle route choice using data from a GPS-assisted household survey'. In: *European Journal of Transport and Infrastructure Research* 18.2.

Golledge, Reginald G (1995). 'Path selection and route preference in human navigation: A progress report'. In: *International conference on spatial information theory*. Springer, pp. 207–222.

Graczyk, Magdalena et al. (2010). 'Comparison of bagging, boosting and stacking ensembles applied to real estate appraisal'. In: *Intelligent Information and Database Systems: Second International Conference, ACIIDS, Hue City, Vietnam, March 24-26, 2010. Proceedings, Part II 2*. Springer, pp. 340–350.

Hadjiconstantinou, Eleni and Nicos Christofides (1999). 'An efficient implementation of an algorithm for finding K shortest simple paths'. In: *Networks: An International Journal* 34.2, pp. 88–101.

Hagberg, Aric, Pieter Swart and Daniel S Chult (2008). *Exploring network structure, dynamics, and function using NetworkX*. Tech. rep. Los Alamos National Lab.(LANL), Los Alamos, NM (United States).

Harvey, Francis, Kevin J Krizek and Reuben Collins (2008). *Using GPS data to assess bicycle commuter route choice*. Tech. rep.

Hillel, Tim et al. (2021). 'A systematic review of machine learning classification methodologies for modelling passenger mode choice'. In: *Journal of Choice Modelling* 38, p. 100221.

Huang, Jie and David M Levinson (2015). 'Circuity in urban transit networks'. In: *Journal of Transport Geography* 48, pp. 145–153.

Ji, Ang and David Levinson (2020). 'Injury severity prediction from two-vehicle crash mechanisms with machine learning and ensemble models'. In: *IEEE Open Journal of Intelligent Transportation Systems* 1, pp. 217–226.

Lai, Xinjun et al. (2019). 'Understanding drivers' route choice behaviours in the urban network with machine learning models'. In: *IET Intelligent Transport Systems* 13.3, pp. 427–434.

Lee, Dongwoo, Sybil Derrible and Francisco Camara Pereira (2018). 'Comparison of four types of artificial neural network and a multinomial logit model for travel mode choice modeling'. In: *Transportation Research Record* 2672.49, pp. 101–112.

Li, Siyuan, Matthew Muresan and Liping Fu (2017). 'Cycling in Toronto, Ontario, Canada: Route choice behavior and implications for infrastructure planning'. In: *Transportation Research Record* 2662.1, pp. 41–49.

Lin, Zijing and Wei (David) Fan (2020). 'Bicycle ridership using crowdsourced data: Ordered probit model approach'. In: *Journal of Transportation Engineering, Part A: Systems* 146.8, p. 04020076.

Liu, Shan et al. (2023). 'AdaBoost-Bagging deep inverse reinforcement learning for autonomous taxi cruising route and speed planning'. In: *Transportation Research Part E: Logistics and Transportation Review* 177, p. 103232.

Lou, Yingyan, Yafeng Yin and Siriphong Lawphongpanich (2010). 'Robust congestion pricing under boundedly rational user equilibrium'. In: *Transportation Research Part B: Methodological* 44.1, pp. 15–28.

Lundberg, Scott M and Su-In Lee (2017). 'A Unified Approach to Interpreting Model Predictions'. In: *Advances in Neural Information Processing Systems* 30.

Mai, Tien, Mogens Fosgerau and Emma Frejinger (2015). 'A nested recursive logit model for route choice analysis'. In: *Transportation Research Part B: Methodological* 75, pp. 100–112.

McNees, Stephen K (1990). 'The role of judgment in macroeconomic forecasting accuracy'. In: *International Journal of Forecasting* 6.3, pp. 287–299.

Menard, Jason, Francis Harvey and Kevin J Krizek (2009). *Improving GPS Data Collection of Human Spatial Behavior*. Tech. rep.

Palmer, Tim (2019). 'The ECMWF ensemble prediction system: Looking back (more than) 25 years and projecting forward 25 years'. In: *Quarterly Journal of the Royal Meteorological Society* 145, pp. 12–24.

Petropoulos, Fotios, Rob J Hyndman and Christoph Bergmeir (2018). 'Exploring the sources of uncertainty: Why does bagging for time series forecasting work?' In: *European Journal of Operational Research* 268.2, pp. 545–554.

Pillat, Juliane, Eileen Mandir and Markus Friedrich (2011). 'Dynamic choice set generation based on global positioning system trajectories and stated preference data'. In: *Transportation Research Record* 2231.1, pp. 18–26.

Politis, Ioannis et al. (2023). 'A Route Choice Model for the Investigation of Drivers' Willingness to Choose a Flyover Motorway in Greece'. In: *Sustainability* 15.5, p. 4614.

Prashker, Joseph N and Shlomo Bekhor (2000). 'Congestion, stochastic, and similarity effects in stochastic: User-equilibrium models'. In: *Transportation Research Record* 1733.1, pp. 80–87.

Prato, Carlo Giacomo (2009a). 'Route choice modeling: past, present and future research directions'. In: *Journal of Choice Modelling* 2.1, pp. 65–100. ISSN: 1755-5345.

— (2009b). 'Route choice modeling: past, present and future research directions'. In: *Journal of Choice Modelling* 2.1, pp. 65–100.

Prato, Carlo Giacomo and Shlomo Bekhor (2006). 'Applying branch-and-bound technique to route choice set generation'. In: *Transportation Research Record* 1985.1, pp. 19–28.

Prato, Carlo Giacomo and Shlomo Bekhor (2007). 'Modeling route choice behavior: How relevant is the composition of choice set?' eng. In: *Transportation Research Record* 2003.2003, pp. 64–73. ISSN: 0361-1981.

Quattrone, Agata and Antonino Vitetta (2011). 'Random and fuzzy utility models for road route choice'. In: *Transportation Research Part E: Logistics and Transportation Review* 47.6, pp. 1126–1139.

Ramming, Michael Scott (2001). 'Network knowledge and route choice'. In: *Unpublished Ph. D. Thesis, Massachusetts Institute of Technology*.

Rasouli, Soora and Harry JP Timmermans (2014). 'Using ensembles of decision trees to predict transport mode choice decisions: Effects on predictive success and uncertainty estimates'. In: *European Journal of Transport and Infrastructure Research* 14.4, pp. 412–424.

Rates, E (2007). *Atlanta Commute Vehicle Soak and Start Distributions and Engine Starts per Day Impact on Mobile Source*.

Ray, Evan L et al. (2020). 'Ensemble forecasts of coronavirus disease 2019 (COVID-19) in the US'. In: *MedRXiv*, pp. 2020–08.

Rieser-Schüssler, Nadine, Michael Balmer and Kay W Axhausen (2013). 'Route choice sets for very high-resolution data'. In: *Transportmetrica A: Transport Science* 9.9, pp. 825–845.

Sahoo, Rosalin et al. (2022). 'A hybrid ensemble learning-based prediction model to minimise delay in air cargo transport using bagging and stacking'. In: *International Journal of Production Research* 60.2, pp. 644–660.

Schmid, Basil et al. (2022). 'Modeling train route decisions during track works'. In: *Journal of Rail Transport Planning & Management* 22, p. 100320.

Shapley, Lloyd S et al. (1953). 'A value for n-person games'. In.

Sharma, Nonita et al. (2021). 'A heterogeneous ensemble forecasting model for disease prediction'. In: *New Generation Computing*, pp. 1–15.

Sheffi, Yosef (1985). *Urban transportation networks*. Vol. 6. Prentice-Hall, Englewood Cliffs, NJ.

Simon, Herbert Alexander (1957). *Models of man: social and rational; mathematical essays on rational human behavior in society setting*. Wiley.

Sobhani, Ana, S Ali Haji Esmaeli and Ahmad Sobhani (2018). 'On the Use of Machine Learning Approaches for Implicit Modeling of Cycling Route Choice: An Application of Machine Learning versus Path Sampling-Logit Model Framework'. In: *IATBR 2018: 15th International Conference on Travel Behaviour Research*.

Sobhani, Anae, Hamzeh Alizadeh Aliabadi and Bilal Farooq (2019). 'Metropolis-Hasting based Expanded Path Size Logit model for cyclists' route choice using GPS data'. In: *International Journal of Transportation Science and Technology* 8.2, pp. 161–175. ISSN: 2046-0430.

Spissu, Erika, Italo Meloni and Benedetta Sanjust (2011). 'Behavioral analysis of choice of daily route with data from global positioning system'. In: *Transportation Research Record* 2230.1, pp. 96–103.

Sun, Bingrong and Byungkyu Brian Park (2017). 'Route choice modeling with support vector machine'. In: *Transportation Research Procedia* 25, pp. 1806–1814.

Tang, Wenyun, David Levinson et al. (2015). 'An empirical study of the deviation between actual and shortest travel time paths'. In.

Tang, Wenyun and David M Levinson (2018). 'Deviation between actual and shortest travel time paths for commuters'. In: *Journal of Transportation Engineering, Part A: Systems* 144.8, p. 04018042.

Tilahun, Nebiyou Y, David M Levinson and Kevin J Krizek (2007). 'Trails, lanes, or traffic: Valuing bicycle facilities with an adaptive stated preference survey'. In: *Transportation Research Part A: Policy and Practice* 41.4, pp. 287–301.

Transport Publications (2013). *TBI (Travel Behavior Inventory)*. Metropolitan Council, St. Paul, MN. (Accessed: 2014-09-30 online at: http://www.metrocouncil.org/Transportation/Publications-Resources-NEW.aspx).

Tribby, Calvin P et al. (2017). 'Analyzing walking route choice through built environments using random forests and discrete choice techniques'. In: *Environment and Planning B: Urban Analytics and City Science* 44.6, pp. 1145–1167.

Van Cranenburgh, Sander et al. (2022). 'Choice modelling in the age of machine learning-discussion paper'. In: *Journal of Choice Modelling* 42, p. 100340.

Wang, Haotian, Emily Moylan and David Levinson (2024). 'Route Choice Set Generation on High-Resolution Networks'. In: *Transportation Research Record* 2678.5, pp. 112–126.

Wang, Haotian, Emily Moylan and David M Levinson (2022). 'Prediction of the Deviation between Alternative Routes and Actual Trajectories for Bicyclists'. In: *Findings*, p. 35701.

Wang, Shenhao, Baichuan Mo and Jinhua Zhao (2020). 'Deep neural networks for choice analysis: Architecture design with alternative-specific utility functions'. In: *Transportation Research Part C: Emerging Technologies* 112, pp. 234–251.

Wardrop, John Glen (1952). 'Road paper. some theoretical aspects of road traffic research.' In: *Proceedings of the Institution of Civil Engineers* 1.3, pp. 325–362.

Wu, Hao, Jinwoo Brian Lee and David Levinson (2023). 'The node-place model, accessibility, and station level transit ridership'. In: *Journal of Transport Geography* 113, p. 103739.

Wu, Hao and David Levinson (2021). 'The ensemble approach to forecasting: a review and synthesis'. In: *Transportation Research Part C: Emerging Technologies* 132, p. 103357.

— (2022). 'Ensemble Models of For-Hire Vehicle Trips'. In: *Frontiers in Future Transportation* 3, p. 876880.

Xiong, Chenfeng, Xuesong Zhou and Lei Zhang (2018). 'AgBM-DTALite: An integrated modelling system of agent-based travel behaviour and transportation network dynamics'. In: *Travel Behaviour and Society* 12, pp. 141–150.

Yamamoto, Toshiyuki, Ryuichi Kitamura and Junichiro Fujii (2002). 'Drivers' route choice behavior: analysis by data mining algorithms'. In: *Transportation Research Record* 1807.1, pp. 59–66.

Yao, Rui and Shlomo Bekhor (2020). 'Data-driven choice set generation and estimation of route choice models'. In: *Transportation Research Part C: Emerging Technologies* 121, p. 102832.

Yen, Jin Y (1971). 'Finding the k shortest loopless paths in a network'. In: *Management Science* 17.11, pp. 712–716.

Zhang, Lei (2011). 'Behavioral foundation of route choice and traffic assignment: Comparison of principles of user equilibrium traffic assignment under different behavioral assumptions'. In: *Transportation Research Record* 2254.1, pp. 1–10.

Zhang, Lei and David Levinson (2008). 'Determinants of route choice and value of traveler information: a field experiment'. In: *Transportation Research Record* 2086.1, pp. 81–92.

Zhang, Lei, Feng Xie and David Levinson (2009). 'Illusion of motion: Variation in value of travel time under different freeway driving conditions'. In: *Transportation Research Record* 2135.1, pp. 34–42.

Zhang, Yunlong and Yuanchang Xie (2008). 'Travel mode choice modeling with support vector machines'. In: *Transportation Research Record* 2076.1, pp. 141–150.

Zhao, Xilei et al. (2020). 'Prediction and behavioral analysis of travel mode choice: A comparison of machine learning and logit models'. In: *Travel Behaviour and Society* 20, pp. 22–35.

Zhao, Zhan and Yuebing Liang (2023). 'A deep inverse reinforcement learning approach to route choice modeling with context-dependent rewards'. In: *Transportation Research Part C: Emerging Technologies* 149, p. 104079.

Zhou, Zhi-Hua (2021). 'Ensemble learning'. In: *Machine Learning*. Springer, pp. 181–210.

Zhu, Shanjiang (2010). *The roads taken: Theory and evidence on route choice in the wake of the I-35W Mississippi River bridge collapse and reconstruction*. University of Minnesota.

Zhu, Shanjiang and David Levinson (2015). 'Do people use the shortest path? An empirical test of Wardrop's first principle'. In: *PloS One* 10.8, e0134322.

— (2018). 'Agent-based route choice with learning and exchange of information'. In: *Urban Science* 2.3, p. 58.

Zimmermann, Maëlle and Emma Frejinger (2020). 'A tutorial on recursive models for analyzing and predicting path choice behavior'. In: *EURO Journal on Transportation and Logistics* 9.2, p. 100004.

Zimmermann, Maëlle, Tien Mai and Emma Frejinger (2017). 'Bike route choice modeling using GPS data without choice sets of paths'. In: *Transportation Research Part C: Emerging Technologies* 75, pp. 183–196. ISSN: 0968-090X.

APPENDIX A

Similarity Measurement on a Grid Network

In the example below, the new similarity measurement is tested on a simple grid network as shown in Figure A0.1. The origin and destination are located at the top left corner and the bottom right corner. Link attributes details are presented in Table A.1. The travel speed is only determined by the type of road. Except for the three attributes in Table A.1, two route-based attributes, the number of traffic lights and the number of left turns, are included in the similarity measurement. In addition, as five attributes are equally weighted in this example, the formula for calculating the attribute similarity simplifies to:

$$s = 1 - \frac{1}{5} \cdot \sum_{x}^{|\mathbf{X}|} \frac{|\hat{x_b} - x_b|}{\hat{x_b} + x_b} \tag{A.1}$$

As shown in Figure A0.2, four predicted routes (PR) that indicate different cases are compared:

- PR1, in Figure A0.2a is designed to show the case which has zero overlap rate but spatially close to chosen route.
- PR2 in Figure A0.2b is used for comparing with PR1 to show the importance of spatial similarity.
- PR3 in Figure A0.2c has zero overlap rate but similar route attributes to chosen route, is designed to show the influence of attribute similarity.
- PR4 in Figure A0.2d has a high overlap rate but also a low attribute similarity and spatial similarity.

According to Equation 4.17, the first step is calculating the overlap rate between predicted routes and the chosen route. The step results are presented in Table A.2. After obtaining the

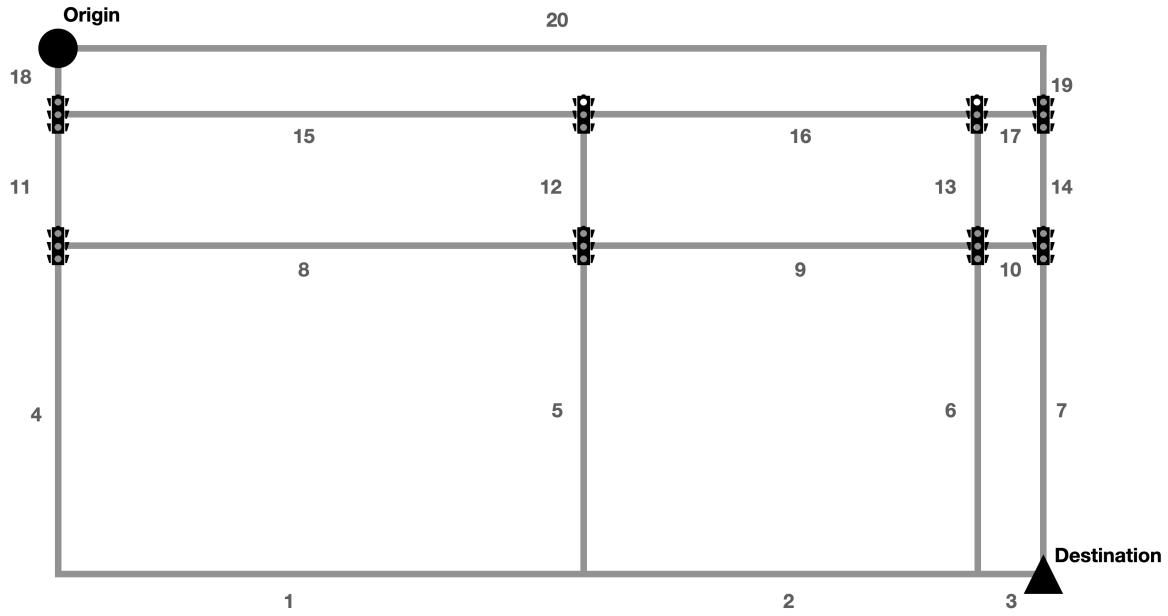
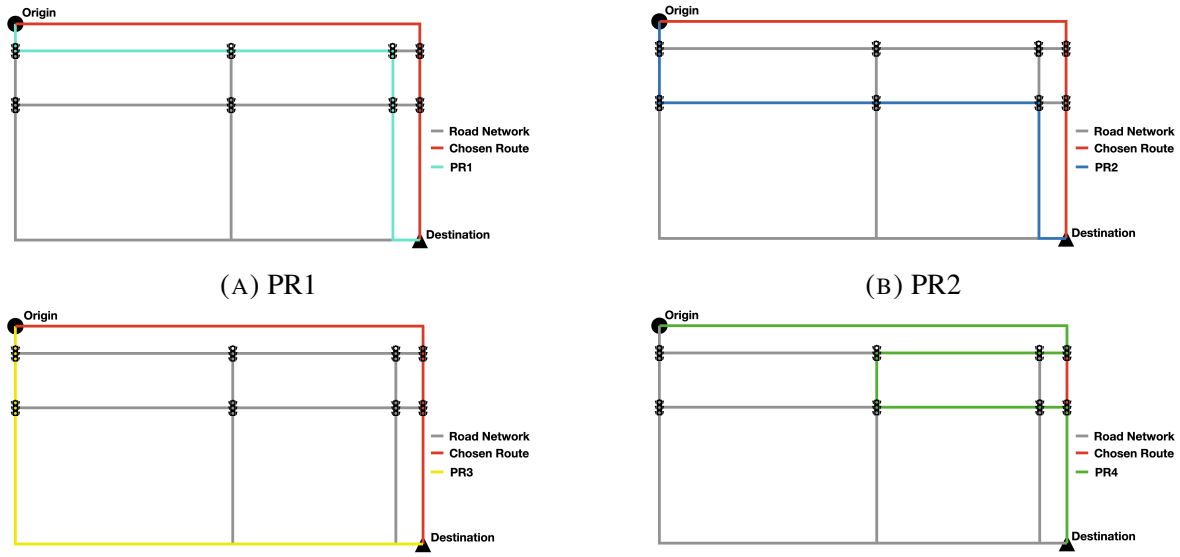


FIGURE A0.1: Simple Grid Network



(A) PR1 (B) PR2

(C) PR3 (D) PR4

FIGURE A0.2: Routes in the Similarity Measure Example

overlap rate, the attribute similarity of the non-overlapped links is calculated for each pair of a predicted route and a chosen route. Since different predicted routes share different links with the chosen route, the attributes of non-overlapped links in the chosen route (x_i) differ

TABLE A.1: Link attributes of the simple grid network

Link ID	Length (km)	Freeway	Travel time (minutes)
1	8	Yes	4
2	6	Yes	3
3	1	Yes	0.5
4	5	No	5
5	5	No	5
6	5	No	5
7	5	No	5
8	8	No	8
9	6	No	6
10	1	No	1
11	2	No	2
12	2	No	2
13	2	No	2
14	2	No	2
15	8	No	8
16	6	No	6
17	1	No	1
18	1	No	1
19	1	No	1
20	15	Yes	7.5

for each case. The step results of each pair of the predicted route and the chosen route are recorded in Table A.2.

The spatial similarity between predicted route and chosen route is calculated after gaining the attribute similarity. The area of the region formed by the predicted route and chosen route, the length of non-overlapped links, the average deviation between the predicted and chosen route, and the spatial similarity are recorded in Table A.2. Finally, the overall similarities between the predicted routes and chosen route are calculated based on overlap rate. attributes similarity and spatial similarity. Results are recorded in Table A.2. Three different weights (α_1, α_2 and α_3) of attribute similarity are tested in this study. They represent 3 different scenarios, which are:

: Scenario 1 (α_1): attribute similarity and spatial similarity are equally important when measureing the overall similarity S between predicted route and chose route

TABLE A.2: Overall similarity calculations in the similarity measure example

Terms	Measurements	PR1	PR2	PR3	PR4
Overlap rate	length of the shared links (L_s)	0	0	0	21
	length of chosen route (L_t)	23	23	23	23
	overlap rate (Ω)	0	0	0	0.91
Attribute similarity	chosen route non-overlapped length (x_1)	23	23	23	2
	predicted route non-overlapped length ($\hat{x_1}$)	23	23	23	16
	chosen route non-overlapped travel time (x_2)	15.5	15.5	15.5	2
	predicted route non-overlapped travel time ($\hat{x_2}$)	22.5	22.5	15.5	16
	chosen route non-overlapped freeway coverage (x_3)	0.65	0.65	0.65	0
	predicted route non-overlapped freeway coverage ($\hat{x_3}$)	0.04	0.04	0.65	0
	chosen route non-overlapped left turns (x_4)	0	0	0	0
	predicted route non-overlapped left turns ($\hat{x_4}$)	2	2	1	2
	chosen route non-overlapped traffic lights (x_5)	2	2	2	2
	predicted route non-overlapped traffic lights ($\hat{x_5}$)	4	4	2	6
	attribute similarity (s)	0.52	0.52	0.8	0.39
Spatial similarity	length of non-overlapped links in chosen route (L_s)	23	23	23	2
	area of the region formed by two routes (Λ)	22	50	120	14
	average deviation (d)	0.96	2.17	5.22	7
	deviation (D)	0.04	0.09	0.18	0.78
	spatial similairty ($1 - D$)	0.96	0.91	0.82	0.22
Overall similarity	overall similarity (S_1) with α_1	0.74	0.72	0.81	0.94
	overall similarity (S_2) with α_2	0.94	0.89	0.82	0.93
	overall similarity (S_3) with α_3	0.61	0.6	0.8	0.93

: Scenario 2 (α_2): spatial similarity is more important

: Scenario 3 (α_3): attribute similarity is more important

By comparing the result of PR1 and PR2 in Table A.2, when the overlap rate and attribute similarity of two predicted routes are the same, the proposed similarity measurement still can correctly indicate the route that is closer to the chosen route. For PR3, as expected, even though it has zero overlap rate and lower spatial similarity than PR1 and PR2 , it still gains a high similarity with the chosen route for cases where attribute similarity is equally or more important. PR4 has high similarity in all three scenarios. Since all these 4 predicted routes are not 100% overlapped with chosen route, if using the metrics derived from the confusion matrix to evaluate them, all of them will be indentified as incorrect prediction, and no further information can be obtained. The proposed similarity can be seen as adding

missing information of the non-overlapped parts in routes to the overlap rate measurement. Note: A route with 100% overlap also has 100% attribute similarity and 0 deviation, so the Ω could be dropped and the model reduced to the following with identical results.

APPENDIX B

Exponentially-Weighted Soft Voting

When using heterogenous ensembles using soft voting, a group of weights including w, w^2, w^3, w^4, w^5 and w^{10} are used to test an exponentially-weighted soft voting strategy. We only include results for w in the main body of this paper, and the results for the rest weights are presented below.

Figure B0.1 shows the sensitivity, specificity, and precision for ensemble using Exponentially-Weighted soft voting with different weights.

Figure B0.2 shows the log-likelihood for ensemble using Exponentially-Weighted soft voting with different weights.

Table B.1 illustrates the performance of ensemble using Exponentially-Weighted soft voting with different weights.

	Validation Set 1			Validation Set 2		
Model	$RMSE$	$MAPE$	E_{net}	$RMSE$	$MAPE$	E_{net}
Soft	**2793**	*16.29*	**6.53**	**3595**	**25.67**	15.83
Soft(w)	*2804*	**16.01**	*10.16*	*3728*	*26.15*	15.49
Soft(w^2)	2814.88	18.82	11.67	3737	26.28	15.02
Soft(w^3)	2826.94	20.41	11.40	3748	26.39	14.76
Soft(w^4)	2836.79	20.53	11.11	3756	26.50	14.48
Soft(w^5)	2843.70	20.56	10.99	3761	26.56	*14.37*
Soft(w^{10})	2856.86	20.63	10.45	3770	26.77	**13.85**

TABLE B.1: Evaluation of mean predicted VKT for Exponentially-Weighted Soft Voting on Freeway Links in LEHD Dataset and TBI Dataset. RMSE: Root Mean Square Error, MAPE: Mean Absolute Percentage Error, E_{net}: Network Percentage Error. Best performers denoted in **bold**. Second best denoted in *italics*.

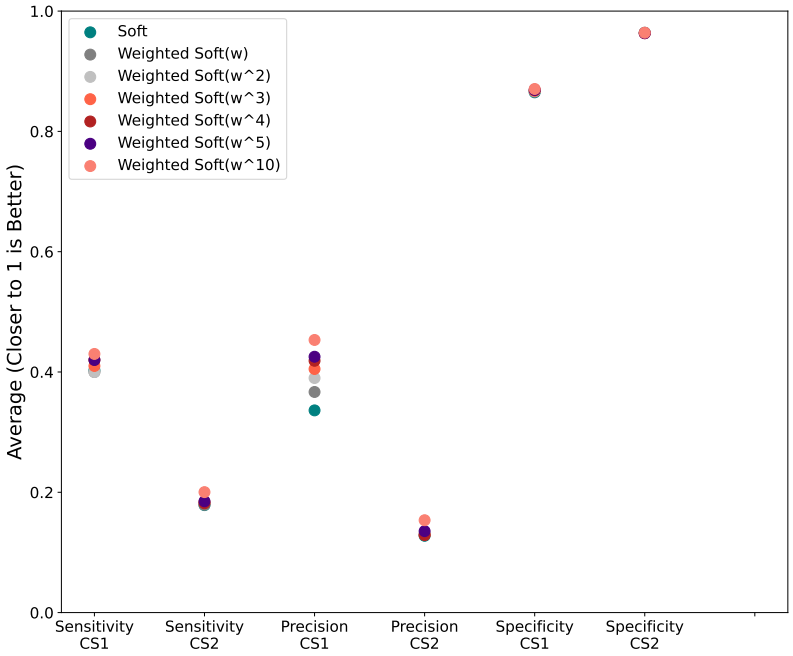


FIGURE B0.1: Average Value of $Sensitivity = \frac{TP}{TP+FN}$, $Specificity = \frac{TN}{TN+FP}$, and $Precision = \frac{TP}{TP+FP}$ for Ensembles Using Soft Voting; triangle: base models also for homogeneous ensemble, star: homogeneous ensemble, square: other base models only for heterogeneous ensemble, and dot:heterogeneous ensemble

The effectiveness of ensembles does not always increase monotonically with the weighting of more accurate learners. In extreme cases, where the weight assigned to one model is significantly higher than the others, the ensemble's final result will be dominated by that model, effectively making the ensemble result identical to the result of that model. The power of w should be adjusted based on the dataset. Future research could focus on finding an effective and efficient method to determine robust weights for exponentially-weighted soft voting.

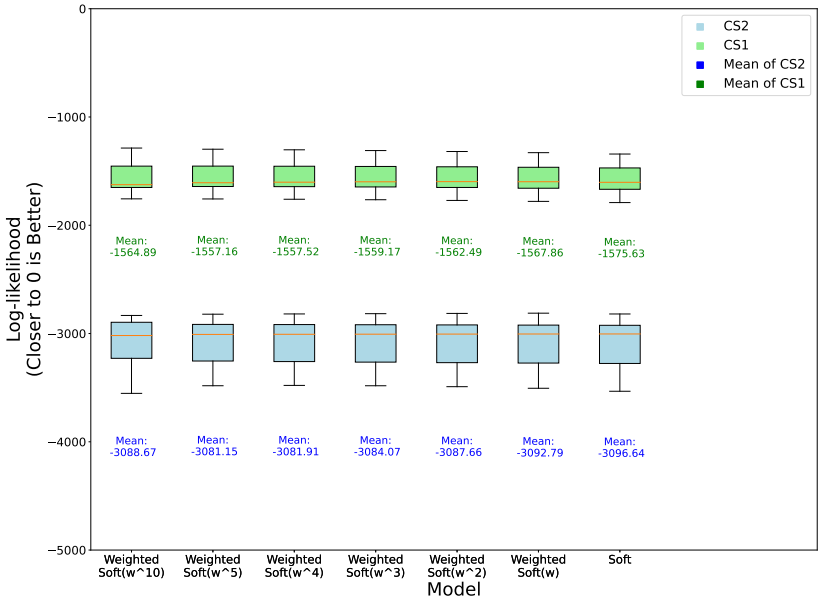


FIGURE B0.2: Log-likelihood for All Soft Voting in Case Study 1 and 2

APPENDIX C

Prediction of deviation between alternatives routes and actual routes for Bicyclists

---

C1 Question

Route choice modelling, especially for bicyclists, quantifies the factors determining route selection and, based on those factors, can be used for facility location and network improvement. For revealed choice datasets, the plausibility of the models is undermined by the challenge of creating realistic choice sets. Moreover, the routes that were actually considered by the traveller may be difficult to identify. Previous studies (Zhu and Levinson 2015; Li et al. 2017) have used ratio of overlap to evaluate the quality of the alternatives, but in dense urban networks, this ratio is low. In this study, we introduce 'deviation' which captures the closeness or average distance between alternative routes and actual trajectories. High values for deviation indicate that the route alternative was not similar to the chosen route, and when all alternatives for a particular trip have high deviation, the choice set has not captured the attribute(s) that are most important to the route choice decision.

The hypotheses are:

(1) Shorter lengths are more likely to result in smaller deviation.
(2) Less traffic signals are more likely to result in smaller deviation.
(3) Lower traffic flow are more likely to result in smaller deviation.
(4) Routes with high percentage of bike trail are more likely to result in less deviation.
(5) Route with high percentage of dedicated bike lane (or on-street bike lane) are more likely to result in less deviation.

C2 Method

C2.1 Data collection

In this study, all GPS data are obtained from Harvey et al. (2008) and Menard et al. (2009), who collected repeated data over 2 weeks from 49 regular bicyclists living in South Minneapolis in 2006. Small GPS dataloggers, which record location and elevation every 2 seconds with roughly 3 meters accuracy, are attached to participants' bicycles. All GPS points which are located in a 100m radius of the participant's home or workplace location, are removed and the recorded home and workplace locations have been randomized within the same radius to protect the privacy of participants. In addition, as recorded GPS for home and workplace were jittered, and participants might not start recording their trips exactly at the home or work place, there is a discrepancy between home or workplace and the start point or end point of some trips. A 250-m radius around the home and workplace is used to filter out origins and destinations of non-commute trips such as shown in Figure C2.1. After filtering, 600 of 831 trips remain.

C2.2 Map matching

As the map data are downloaded from OpenStreetMap version 2021 and the trajectories were recorded in 2006, adjustments based on archival Google Maps StreetView are carried out to reproduce the historical road network before matching GPS data to network. In addition, the 2021 cycleways have been categorized as three different types of cycleway in 2006.

As shown in Figure C2.2, the first type is *dedicated cycleway*, which are approximately 5 ft-wide (1.5 m) and located on the edge of the roadway. *Bike trails* are another facility designed for bicycles but do not share the right of way with cars to provide safer travelling conditions for cyclists. The last type is *shared cycleways*, (typically sharrows) which are shared with other transport modes like buses or cars.

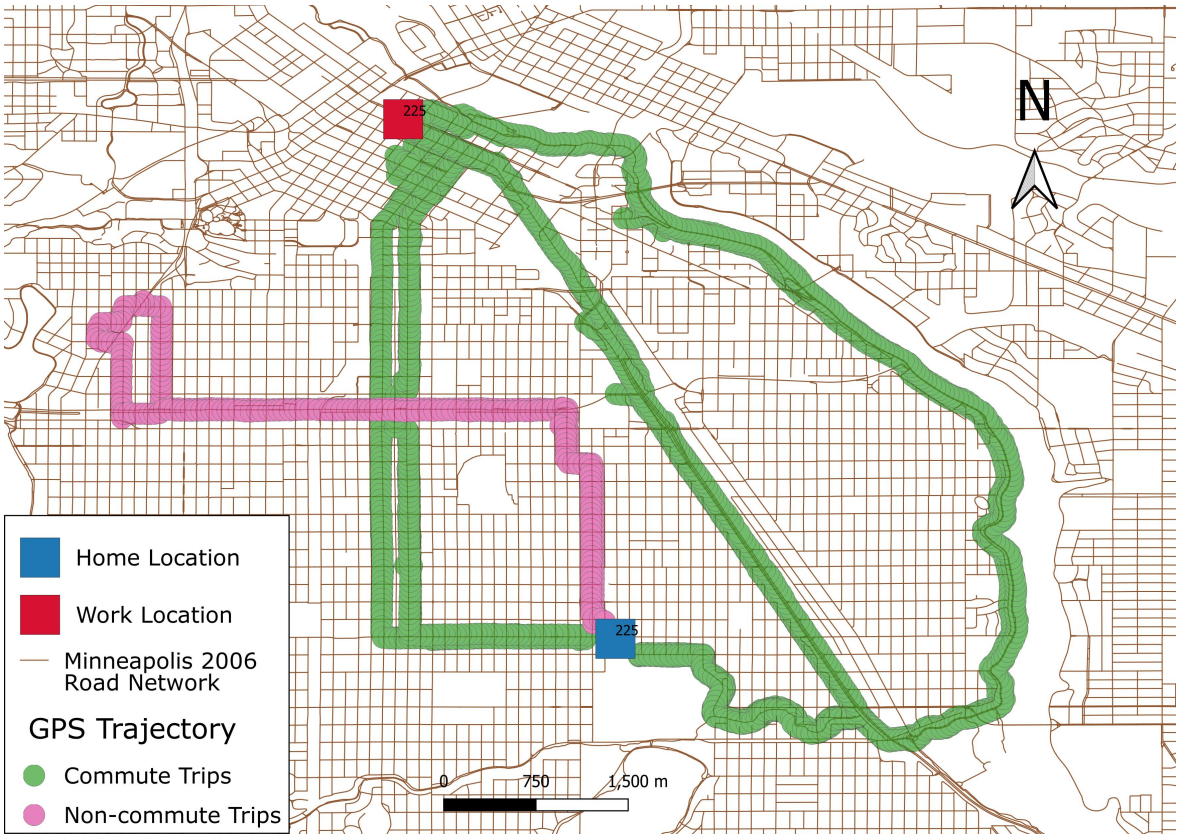


FIGURE C2.1: Example of Commute and Non-Commute Trips

FIGURE C2.2: Examples of three types of cycleways. Source: Google Streetview of Minneapolis.

In this study, we use KD-Tree which is a geometric measure for matching trajectories to the network. To populate network points set, a point was generated every 55 meters along each segment. The basic idea of this method is setting the populated point set as reference layer and trajectory points set as target layer. For each point in target layer, the algorithm finds the closest point in reference layer and records it into the matched trip set.

C2.3 Choice set generation

The size and quality of the choice set influence the route choice modelling (Prato 2009a). To gain a choice set which better captures actual trajectories, a link labelling approach (Ben-Akiva et al. 1984), which generates alternative paths by optimizing different criteria, is implemented. According to previous studies (Broach et al. 2012; Bernardi et al. 2018; Lin and Fan 2020; Zimmermann et al. 2017), distance, traffic flow, and bike facilities are important factors for cyclists planning their trips.

To understand the preference of bike facilities for 49 participants, the constitution of each trip is analysed. For each type of bike facility, the number of trips for which the length percentage of that facility is higher than other road types is listed in Table C.1. Based on the lack of trips dominated by shared cycleways, this facility is not as important as the other two when people choose travel routes. Thus paths with maximized bike trail proportion and paths with maximized dedicated cycleway proportion are included in the choice set. However, paths with maximized percentage of bike trails sometimes result in extremely long routes (approximate 3 times the shortest path length) which might be unrealistic for most commute trips. To include a more realistic percentage of the length on bike trails, a set of factors from 0.1 to 1 with 0.1 increase each time are used to weight the bike trails' length. In addition, as recommended in previous study (Sobhani et al. 2019; Ghanayim and Bekhor 2018; Tilahun et al. 2007), the shortest path and the fastest path are generated for each OD pair.

All generated path are then included into a set C_0, and alternative routes for a OD pair with more than 80% similarity, which is measured by the length of common links, are removed from C_0. The size of C_0 is too large for efficient modeling, and the difference of the size of choice sets between travelers might also create bias. Therefore, for each traveller, 6 paths are defined

- shortest path ($C_{i=0}$),
- path with maximum proportion of dedicated cycleways ($C_{i=1}$),
- path with maximum proportion of on-street cycleways ($C_{i=2}$),
- path with maximum proportion of secondary road ($C_{i=3}$),

TABLE C.1: Length proportion of Bike facilities

Road type Class	Number of Trips, constituted mostly by the type class	Max length proportion
Bike trail	225	0.877
Dedicated cycleway	73	0.705
Shared cycleway	0	0.200

- path with maximum proportion of cycleway ($C_{i=4}$), and
- path with maximum proportion of bike trail but total length within 1.3 times shortest path length ($C_{i=5}$),

have been selected to form choice set C. The reason for containing $Ci = 3$ is that secondary roads are found have a relative high proportion of recorded trips, and fewer turns are needed on secondary roads. For $C_{i=5}$, scenarios with length within 1.1 to 1.6 times shortest path length are tested, and 1.3 gives the highest similarity to observed trajectories.

C2.4 Deviation

Deviation measures the similarity between paths. It can be applied to compare the generated alternative routes and actual trajectories. To measure the dependent variable 'deviation', D, we construct the convex hull, which is the polygon formed by the set of all points in the alternative route and trajectory as shown in Figure C2.3. D equals to the square root of that area.

C2.5 Panel regression model

The alternative route with the lowest D best captures the features of the selected route. Since 600 trips are collected from 49 travelers during a period of time, to control the correlation of the errors due to unobserved variables associated with panel data, a Fixed Effect (FE) regression model is applied to model the deviation D, and the results are compared to pooled

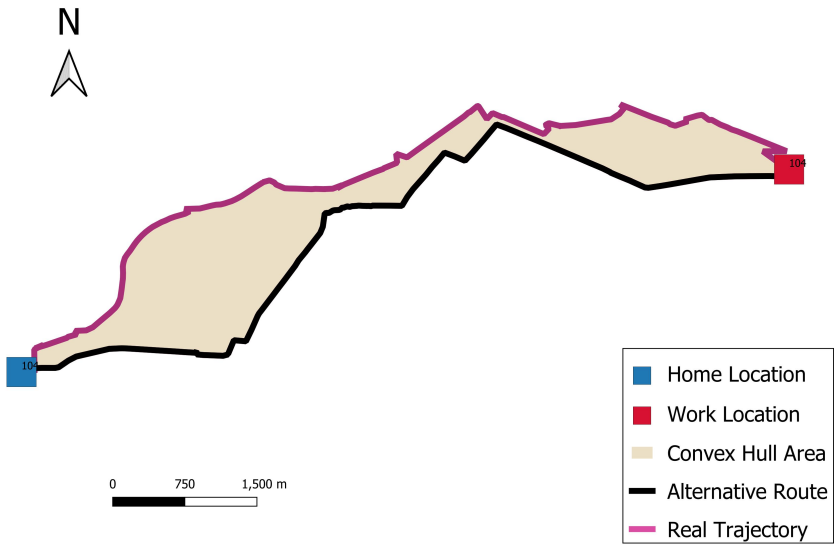


FIGURE C2.3: Convex Hull Example

TABLE C.2: Variables in regression model

Variables	Description	Symbol
Length	Total length of the route	L
VKT	Total vehicle kilometres travelled for the route	V
Traffic light per meter	Total number of traffic light on the route divide by total length of the route	T
Length percentage of bike trail	Total length of bike trail in the route divide by total length of the route	C
Length percentage of dedicated cycleway	Total length of dedicated cycleway in the route divide by total length of the route	B

ordinary least squares (OLS) regression model. The variables in regression models are presented in Table C.2.

TABLE C.3: Outputs for Convex Hull: Deviation between alternative routes and actual route

Variables	Entity and Time Fixed Effects	Pooled OLS
constant	515.23** (250.31)	775.90*** (77.40)
Length	242.03*** (25.35)	230.23*** (6.78)
VKT	-0.0018 (0.0011)	-0.0029 (-0.0007)
Traffic light per km	72.29 (51.823)	32.25 (26.304)
Percentage of bike trail	-724.18*** (190.25)	-832.10*** (120.95)
Percentage of dedicated cycleway	-285.29 (190.25)	-210.07* (122.94)
F-test for Poolability	0.00	0.00
Adjusted R^2	0.349	0.345
Durbin-Watson	2.12	1.56

(standard error)
*: significance at 10% level
**: significance at 5% level
***: significance at 1% level

C3 Findings

Overall, as shown in Table C.3, routes with shorter length and higher percentage of bike trail minimize deviation with actual trips. So the hypotheses 1 and 4 made in section C1 are corroborated. Percentage of on-street bike lane, VKT, and traffic lights per km are not statistically significant at 95% confidence level in this data set.