

Imperial College London
Department of Civil and Environmental Engineering

Endogeneity and consideration-set issues in residential location choice models

Abhilash Chandra Singh

Submitted in part fulfilment of the requirements for the degree of
Doctor of Philosophy at Imperial College London

August 11, 2023

Declaration

I hereby declare that this thesis and the work reported herein were composed by and originated entirely from me. Information derived from the published and unpublished work of others has been acknowledged in the text and references are given in the list of sources.

- Abhilash Chandra Singh (2023)

The copyright of this thesis rests with the author. Unless otherwise indicated, its contents are licensed under a Creative Commons Attribution-Non Commercial 4.0 International Licence (CC-BY). Under this licence, you may copy and redistribute the material in any medium or format. You may also create and distribute modified versions of the work. This is on the condition that: you credit the author and do not use it, or any derivative works, for a commercial purpose. When reusing or sharing this work, ensure you make the licence terms clear to others by naming the licence and linking to the licence text. Where a work has been adapted, you should indicate that the work has been changed and describe those changes. Please seek permission from the copyright holder for uses of this work that are not included in this licence or permitted under UK Copyright Law.

Abstract

Urban land use and transport policies are essential to determine where people live, work and play, enabling sustainable and healthy forms of transportation between destinations. Compact cities are clearly needed to achieve GHG targets, as 70% of the world's population is expected to live in urban areas by 2050 and 33% of total urban GHG emissions in major cities are generated by transport. However, identifying the locations and levels of land use mix and densities along with adequate transport networks and technologies to support equitable, sustainable and healthy living remains challenging. Potential trade-offs between quality of life, pressures on land uses, equitable urban growth, transport efficiencies, and urban temperatures must be addressed transparently to both inform policies and engage society in embracing sustainable solutions. This thesis presents a range of approaches used to model residential location choices within the context of land use and transport models, dealing with multi-faceted problems arising with novel data sources. As part of this effort, we discuss the model scenarios with missing or filtered alternatives during choice-making process by decision makers (consideration-set issues), simultaneous determination of modal accessibilities and trip frequencies, and omitted information during modelling choices (endogeneity issues).

Firstly, residential location choice models are often burdened with missing information for modelling the choice set formation process. The information is missing either because one or more alternatives are unavailable to decision makers or because they considered (sampled) the alternatives endogenously, a consideration process that is inaccessible to the analyst. To the best of our knowledge, this is the first application of probabilistic decision trees to model consideration process in residential location choice using London-based data. We combine probabilistic decision trees with a traditional Mixed Logit choice model to estimate the final choice decision, capturing interdependence between alternatives within specific Transport Fare Zones, thereby accounting for non-compensatory consideration of choice alternatives followed by a compensatory choice decision. The results reveal factors influencing consideration set formation, highlighting the impact of variables such as household income, borough prices, population density, and crime density on location choice probability. Ethnic variation in the effect of crime density on location choice indicates racial and ethnic segregation in London housing. The research confirms the use of a non-compensatory process followed by a compensatory one in decision-making. However, the model has data limitations and

challenges in merging machine learning and econometric approaches.

Secondly, it is widely acknowledged that residential location choice is driven not just by the built environment attributes but also by the level of access to activity opportunities from the location. We simultaneously analyse employment accessibility and household trip frequency across different transport modes in the City of Toronto. Accessibility measures for auto, transit, bicycle, and walk modes were created, along with household trip frequencies for each mode, categorized into low, medium, and high levels. The study uses a multivariate ordered probit model system, considering both observed and unobserved factors' dependence on trip frequencies and accessibility. The results reveal strong associations and direct effects between employment accessibility and household trip frequency. Complementary effects are observed between auto accessibility and transit trips, while substitution effects exist between transit accessibility and auto trips. Policymakers can use these findings to promote transit development in transit-poor neighbourhoods and reduce vehicular congestion in Toronto. The limitations of limited cycling and walking infrastructure in Toronto impact trip utility for high population density neighbourhoods, leading to inequities in transport infrastructure distribution. Moreover, certain population groups are time-impaired, limiting their activity participation and raising equity concerns regarding transport infrastructure distribution in the city.

Thirdly, endogeneity in residential location choice models occurs when any explanatory variable is correlated with the error term. This can happen due to omitted attributes, measurement or specification errors, and simultaneous determination or self-selection. Addressing endogeneity is important because it leads to inconsistent parameter estimates. This analysis develops a control function (CF) based discrete choice model, addressing endogeneity in residential location choice due to housing prices and travel behaviour. Our hypothesis highlights the prominent negative impact of housing prices by accounting for endogeneity, while also addressing the endogeneity with travel behaviour. Utilizing London Travel Demand Survey (LTDS) data, we comprehensively analyse household location choices and their relation to travel behaviour. The CF-MxL model findings underscore housing prices' negative influence on location choices, particularly affecting lower-income groups, as compared to highly subdued effect size of the incorrect MNL model. Additionally, vehicle miles travelled (VMT) significantly impacts location choices, with households having higher VMT showing

aversion to postcodes with a higher population density. Furthermore, we observe a positive association of non-white households with locations infested with a higher crime density. This research offers insights into the interplay between residential location and travel behaviour, isolating and revealing the impact of various postcode-attributes, socioeconomic and demographic characteristics.

This thesis comprises three distinct studies investigating residential location choice models and their interactions with external factors. The first study employs probabilistic decision trees to analyse choice set formation, revealing the impact of variables like income, borough prices, population density, and crime density on location decisions. The second study explores the relationship between employment accessibility, household trip frequency, and transport modes in Toronto, uncovering associations and disparities in infrastructure distribution. The third study addresses endogeneity using a control function-based model, highlighting housing price effects while accounting for travel behaviour endogeneity. These findings contribute insights into residential location dynamics, offering valuable implications for policy and understanding the multifaceted aspects of location choice.

Acknowledgements

I stand on the shoulders of giants, inspired by the relentless pursuit of knowledge exemplified by the great scientists who came before me. Their unwavering dedication to unravelling the mysteries of the universe continues to ignite my passion for discovery.

I am profoundly grateful for the guidance, support, and wisdom bestowed upon me throughout this journey. My deepest appreciation goes to my extraordinary advisor, Dr Aruna Sivakumar, whose brilliance, patience, and unwavering belief in me propelled me to surpass my own expectations. She encouraged me to pursue different research directions and opened up many opportunities for fruitful collaborations. The running theme of this thesis is the exploration (and exploitation) of the residential location choice model and its relationship with travel behaviour. Dr Sivakumar's earlier work on spatial location choice served as a great inspiration for the work presented in this thesis. To the members of my evaluation committee, Dr Dan Graham and Dr Jonas de Vos, I extend my heartfelt thanks for your invaluable insights, constructive feedback, and critical evaluation.

I must also thank (Drs) Eric Miller, Rolf Moeckel, Stephen Boyles, Ram Pendyala, S.L. Dhingra, Tom Mathew, RaaJ Ramsankaran, Gopala Srinivasan, Nagendra Velaga, V.K. Mishra and Raj Gupta for their valuable guidance and feedback in my education and research. Your expertise and commitment to academic excellence have shaped the trajectory of my research, and I am honoured to have had the privilege of learning from you. I am indebted to my colleagues and fellow researchers, (Drs) Ahmadreza Faghieh Imani, Prateek Bansal, Fangce Guo, Jacek Pawlak, Francesco Manca, Nicolo Daina, Hajime Watanabe, Sebastian Astroza, Felipe Dias; and friends from Urban Systems Lab: Han, Xuehao, Tang, Heather, Chenyang and Yuanying, whose camaraderie and intellectual discourse have enriched my scientific growth.

I would like to draw upon the words of Sheldon Cooper: *I have been encouraged, sustained, inspired and tolerated not only by my wife, Apoorva, but also by my parents Mr Dilip Kumar Singh and Mrs Manju Singh and sister, Mrs Neha Singh Ojha, whose unwavering love, encouragement, and support have sustained me throughout this journey.* Finally, I would like to acknowledge financial support from the Pathways to Equitable Healthy Cities grant by the Wellcome Trust [209376/Z/17/Z].

Contents

Declaration	i
Abstract	ii
Acknowledgements	iv
1 Introduction	1
1.1 Motivation and Background	1
1.2 Critical gaps	3
1.3 Aims and Objectives	5
1.4 Contributions	7
1.5 Thesis Outline	8
1.6 Journal and Conference Papers	9
2 Literature Review	10
2.1 Disaggregate residential location choice model	10
2.1.1 Discrete choice models: a brief primer	12
2.1.2 Consideration of alternatives in residential location discrete choice models	14
2.1.3 Residential location choice and travel behaviour	15
2.1.4 Impact of Residential Location on Travel Behaviour, and vice versa . . .	16
2.1.5 Accessibility and Travel Behaviour	17
2.1.6 Summary	19
2.2 Non-compensatory decision rules in discrete choice modelling	21
2.2.1 Summary	24
2.3 Endogeneity in the context of residential location choice models	25

2.3.1	Methods to address endogeneity in DCMs	26
2.3.2	Summary	28
2.4	Overall summary of findings and plan ahead	28
3	Semi-compensatory probabilistic model for residential location choices	30
3.1	Introduction	32
3.1.1	Background and context	32
3.1.2	Representation of disjunction-of-conjunctions by decision trees	32
3.1.3	Contributions of this chapter	34
3.2	Data Description	35
3.3	Methodological framework	39
3.3.1	Problem set up	39
3.3.2	Decision Trees	39
3.3.3	Prediction with a decision tree	42
3.3.4	Ensembles of decision trees: Random Forests	42
3.3.5	Discrete Choice model	44
3.4	Estimation Results	48
3.4.1	Description of model implementation and specification search	48
3.4.2	Consideration set formation results	49
3.4.3	Mixed logit model results	50
3.4.4	Comparative evaluation of discrete choice predictions	57
3.5	Conclusion	58
4	Simultaneity of household trip frequencies and accessibility	62
4.1	Introduction	64
4.1.1	Background and motivation	64
4.1.2	Simultaneity bias in accessibility and travel behaviour	65
4.1.3	Contributions of this chapter	66
4.2	Data description	68
4.3	Methodological framework	74
4.3.1	Model formulation	75
4.3.2	Conceptual framework	77

4.4	Estimation results	80
4.4.1	Exogenous variable effects	80
4.4.2	Endogenous variable effects and correlation effects	89
4.5	Conclusions	94
5	An instrumental variable model for addressing endogeneity in residential location choice	97
5.1	Introduction	99
5.1.1	Background and motivation	99
5.1.2	Methods to address endogeneity in residential location discrete choice models	100
5.1.3	Contributions of this chapter	101
5.2	Data Description	102
5.3	Methodological framework	106
5.3.1	Discrete choice model accommodating Control Function (CF)	106
5.3.2	Challenges	109
5.3.3	Testing Validity of Instruments	109
5.4	Results and discussion	110
5.4.1	Description of model implementation and specification search	110
5.4.2	Model fit and validity of instruments	112
5.4.3	Comparative analysis of corrected and endogenous parameters outcomes	113
5.5	Conclusion	117
6	Conclusion	119
6.1	Summary	119
6.2	Discussion of results	121
6.2.1	Accessibility measurement using this research	123
6.3	Limitations and future research directions	125
7	Bibliography	127
A	Appendix: Chapter 3	143
A.1	Probabilistic predictions	143

A.2	Heterogeneity in a population's non-compensatory rules	143
A.3	Context-dependent preference heterogeneity	144
A.4	Uncertainty in estimation	145
A.5	Monotonicity constraints	145
B	Appendix: Chapter 4	146
C	Appendix: Chapter 5	163

List of Tables

3.1	Description of Sample (n=4,866)	37
3.2	Borough choice frequency distribution (n=4,866)	38
3.3	Matching variable names to their description	52
3.4	Residential location choice model results	54
4.1	Description of Sample (n=42,718)	70
4.2	Trip frequency observations (%) by each mode (n=42,718)	70
4.3	Exogenous Variable Effects on Modal trip frequencies	85
4.4	Exogenous Variable Effects on Modal trip frequencies (continued)	86
4.5	Exogenous Variable Effects on Modal Accessibilities	87
4.6	Exogenous Variable Effects on Modal Accessibilities (continued)	88
4.7	Endogenous Variable Effects	92
4.8	Correlation Effects and Goodness-of-fit	93
5.1	Description of Sample (n=4,866)	103
5.2	Outer postcode choice frequency distribution (n=4,866)	105
5.3	First stage housing price regression model	113
5.4	First stage vehicle miles travelled regression model	113
5.5	Control function choice model result compared with simple MNL model	114
B.1	Single adult and no kids: Exogenous Variable Effects on Modal trip frequencies	147
B.2	Single adult and no kids: Exogenous Variable Effects on Modal Accessibilities	148
B.3	Single adult and no kids: Endogenous Variable Effects	149
B.4	Single adult and no kids: Correlation Effects and Goodness-of-fit	150

B.5	Single or couple adults with one or more kids: Exogenous Variable Effects on Modal trip frequencies	151
B.6	Single or couple adults with one or more kids: Exogenous Variable Effects on Modal Accessibilities	152
B.7	Single or couple adults with one or more kids: Endogenous Variable Effects . . .	153
B.8	Single or couple adults with one or more kids: Correlation Effects and Goodness-of-fit	154
B.9	Multi adults with no kids: Exogenous Variable Effects on Modal trip frequencies	155
B.10	Multi adults with no kids: Exogenous Variable Effects on Modal Accessibilities .	156
B.11	Multi adults with no kids: Endogenous Variable Effects	157
B.12	Multi adults with no kids: Correlation Effects and Goodness-of-fit	158
B.13	Multi adults with one or mode kids: Exogenous Variable Effects on Modal trip frequencies	159
B.14	Multi adults with one or mode kids: Exogenous Variable Effects on Modal Accessibilities	160
B.15	Multi adults with one or mode kids: Endogenous Variable Effects	161
B.16	Multi adults with one or mode kids: Correlation Effects and Goodness-of-fit . .	162
C.1	First stage model to estimate residuals for price endogeneity	163
C.2	Control function (for price endogeneity) choice model result	164
C.3	First stage model to estimate residuals for travel behaviour endogeneity	165
C.4	Control function (for travel behaviour endogeneity) choice model result	166

List of Figures

2.1	Cycle of stipulated choice-making process, adapted from De Vos and Ettema (2020)	18
3.1	London Boroughs map	46
3.2	Model implementation	47
3.3	Variable Importance Plot for consideration of alternatives	51
3.4	Comparative evaluation of discrete choice predictions	59
4.1	Map of Greater Toronto region with modal accessibilities, clockwise (from top right) the levels of transit, walk, bicycle and auto accessibility (green for lowest category 1, up to red for highest category 5)	72
4.2	Map of Greater Toronto region with modal trip frequencies, clockwise (from top right) the levels of transit, walk, bicycle and auto trips (green for lowest category 1, up to red for highest category 5)	73
4.3	Hypotheses Regarding Relationships Among Endogenous Variables	79
4.4	Endogenous Variables Effects (the highlighted blue are the significant effects, with values signifying the coefficients, and ‘low/medium/high/very high’ in parentheses signify the level of accessibility for which the effect is quantified . . .	90
5.1	London Postcodes map	104
6.1	Diagrammatic illustration of enhanced multidimensional interpretation of accessibility	124

Chapter 1

Introduction

1.1 MOTIVATION AND BACKGROUND

Urban land use and transportation sectors can play a crucial role to combat climate change, contributing towards the aim to limit global warming to below 2°C as per the Paris Agreement. However, even radical scenarios like Draconian resettlement may fall short in achieving necessary CO_2 reductions, emphasizing the need for balanced land use and transportation policies. The challenge is, therefore, to provide a sustainable and equitable way of promoting growth in urban areas. Urban land use is central to an extremely important allocation decision made for households, namely residential developments. Residential development is one of many urban resource categories that may be considered to have an impact on spatial structure of cities (Kim, 1994; Knox and Pinch, 2000; Wu, 2004; Kim et al., 2005). To understand changes in spatial structure of cities, we model the residential location choices of the citizens. The purpose of residential location choice models is to quantify the contribution of certain factors, such as real estate prices, rent prices, dwelling attributes, accessibility to jobs and healthcare services, school quality and safety concerns, on residential location-relocation decisions (Waddell, 2009; Hurtubia et al., 2011; Farooq and Miller, 2012). These models help to design policy measures deterring urban sprawl, promoting mixed-use of destinations and impede racial segregation in urban areas (Bayer et al., 2005; Guo and Bhat, 2007). These policy measures vary in their design and purpose, which are dependent upon the level of detail involved in modelling choices, giving rise to aggregate and disaggregate approaches.

The aggregate approach of modelling residential locations is used to predict the spatial structure and urban sprawl of population in cities, usually involving gravity-type model

formulations (Wilson, 1970), while **this thesis will focus on the disaggregate approaches of modelling residential locations by discrete choice models based on microeconomic random utility theory, as it is by far the most predominant approach in practice** (Lerman, 1976; McFadden, 1977). The residential location discrete choice models could also be aggregated to support predictions of urban landscape structuring over time (Waddell et al., 2003). In addition to spatial distribution of population, the spatial structure of cities is also affected by the movement of population in cities. Residential and land use destinations can be considered as the origins and destinations of trips and may therefore require availability of transport services to support the movement of people. Accurate prediction of spatial structure relies on understanding of residential location and travel behaviour interactions (Ortuzar and Willumsen, 2011).

The association between travel behaviour and residential location choices can be bidirectional, if not truly endogenous. The effect of built environment on travel behaviour has been widely studied by several researchers (Cao, Mokhtarian, and Handy, 2009; Singh et al., 2018), thus making the case for residential location choices being able to explain travel behaviour. Observed travel behavioural differences may primarily be attributed to the residential location characteristics, access to transport facilities and the interaction between these, out of which accessibility has gained considerable traction in recent times (Miller, 2018; Wu and Levinson, 2020). Therefore, in an ever-changing spatial structure of urban landscape, coupled with complex interdependencies with travel behaviour, **the behaviourally realistic modelling of residential location choices is one of the greatest challenges.** While there has been a lot of research on modelling residential location choice (see Chapter 2 for description and references), there are several challenges in developing these models in practice. Specifically, this thesis aims to look at issues related to missing information and endogeneity, which are some of the most critical challenges faced by current residential location choice models.

Addressing these challenge is imperative as it directly influences the precision and reliability of residential location choice models. The incorporation of behaviourally realistic decision-making into the modelling framework requires more effective approaches (Swait, 2001). This entails the development of methodologies that accurately capture how individuals make choices based on specific attributes, without presuming compensatory trade-offs between diverse factors (Jedidi and Kohli, 2005). We must also accommodate limited cognitive

capacities of decision-makers (Zhu and Timmermans, 2010). Furthermore, the complexity arising from empirical scenarios where decision-makers confront a substantial array of alternatives remains a relatively unexplored area (Guevara, 2010). To ensure the practical applicability and robustness of residential location choice models in real-world contexts, strategies must be devised to navigate and comprehend such intricate decision processes. Additionally, the interplay between a household's residential location decisions and their observed travel behaviour poses challenges (Handy et al., 2005). Effective disentanglement and quantification of both direct and indirect relationships between these factors are crucial to unveil the underlying mechanisms of complementary and substitution effects influencing decision-making (Singh et al., 2018). Finally, the identification and mitigation of endogeneity in residential location choice models (Guevara, 2015), particularly concerning travel behaviour variables, remains underexplored, and warrants the development of robust methodologies for accurate diagnosis and bias correction. **This thesis aims to produce practically useful tools and methods for addressing the challenges of missing information and endogeneity in the context of residential location choice modelling.**

The following section 1.2 described the critical gaps for this research, with section 1.3 providing the aims and objectives, followed by section 1.4 for major contributions arising from this research, section 1.5 for an outline of the chapters of this thesis, and section 1.6 for the list of publications resulting from this research.

1.2 CRITICAL GAPS

The literature review (Chapter 2) identifies several significant research gaps in the field of residential location choice modelling. Firstly, there is a lack of methodologies to address the inherent dependence between alternatives during consideration set formation (Caplin et al., 2019). Incorporating non-compensatory decision-making behaviour into the modelling framework is also needed, requiring approaches that accurately capture attribute-based choices without assuming trade-offs (Cantillo and Ortúzar, 2005). Additionally, handling the empirical scenario where decision-makers face a large and diverse set of alternatives remains unexplored (Bovy, 2009). The review also highlights the challenge of simultaneity endogeneity, emphasizing the need to untangle the complex relationships between residential location choices and observed travel behaviour (Mokhtarian and Cao, 2008). A research gap is

noted in addressing omitted-variable and measurement endogeneity, particularly in the context of travel behaviour variables (Goetzke and Weinberger, 2012). The review underscores the importance of innovative approaches to detect and correct endogeneity biases. The influence of non-compensatory decision-making processes on residential location choice modelling is highlighted, noting challenges and limitations of existing methods (Brathwaite et al., 2017). Furthermore, the impact of endogeneity on residential location choice modelling is emphasized, prompting the need for advanced techniques like instrumental variables and control functions (Guevara, 2015). The control function approach is particularly highlighted for addressing endogeneity due to multiple endogenous variables.

These challenges are discussed in much greater detail in Chapter 2, but the key gaps in the literature in the context of missing information and endogeneity are:

- Dealing with missing/sampling of alternatives: This gap emphasizes the need to address the inherent dependence between alternatives when forming consideration sets. The challenge lies in avoiding biases that can arise from the way alternatives are included or excluded. How can we develop methodologies that accurately reflect decision-makers' choices when faced with a multitude of options? Incorporating non-compensatory decision-making behaviour is crucial here. This means capturing how individuals make choices based on specific attributes, without assuming that different factors can compensate for each other. Additionally, in real-world scenarios, decision-makers often encounter a vast and diverse set of alternatives. How can we develop strategies to handle such complex decision processes, ensuring that the resulting models are both practically applicable and robust?
- Dealing with simultaneity endogeneity: This gap highlights the interplay between a household's residential location decisions and their observed travel behaviour. Understanding the directionality of effects between these two factors is key. How can we determine if a change in residential location leads to changes in travel behaviour or vice versa? Furthermore, it's important to differentiate between direct relationships (where one factor directly influences the other) and indirect relationships (where one factor indirectly influences the other through a third factor). Identifying these relationships can help uncover complementary and substitution effects—how changes in one factor

affect the other. Developing methodologies and empirical applications to tease out these complex relationships is essential for accurate and robust modelling.

- Dealing with omitted-variable and/or measurement endogeneity: This gap revolves around the challenge of addressing endogeneity in residential location choice models, particularly concerning travel behaviour variables. Endogeneity arises when there's a bi-directionality between variables, leading to biased estimates. Detecting this endogeneity is the first step. How can we identify cases where endogeneity is affecting our model outcomes? Once identified, the challenge is to account for and correct this endogeneity. This ensures that the model results accurately represent the causal relationships between variables, rather than being distorted by hidden factors. Developing methods to detect, account for, and correct for endogeneity is crucial for achieving unbiased and reliable results in residential location choice analyses.

1.3 AIMS AND OBJECTIVES

Recent advances in data availability and the rapid development of machine learning methods provide a compelling impetus to address these critical gaps in residential location choice models. These opportunities stem from the potential of new data sources to offer richer insights into decision-making processes, enabling more accurate consideration set formations and a deeper understanding of the interplay between residential choices and travel behaviour. Novel econometric frameworks, optimisation methods and machine learning techniques, with their capacity to handle large and complex datasets, present avenues to capture heterogeneity and endogeneity effects. This convergence of data availability and advanced analytical tools offers a fertile ground for the development of innovative methodologies that can substantially enhance the accuracy, reliability, and practical utility of residential location choice models. As urban environments continue to evolve, harnessing these opportunities is essential for informed urban planning and policymaking, ensuring that the resulting models align with the complexities of real-world decision processes.

This thesis aims to address issues in residential location choice models from a methodological, empirical and policy perspective, thereby developing models which not only advance an understanding of the current issues but also develop deployable solutions for the

policymakers. This research provides new insights into the modelling of residential location choices and its relationship with travel behaviour, particularly solving for endogeneity and choice-set issues using econometric and machine learning techniques. The key objectives of the thesis, and associated research questions, are as follows:

1. The first objective is to develop a probabilistic consideration set formation model of residential location choices. This arises from the need to capture the inherent uncertainty and complexity in how individuals or households filter and choose among a plethora of alternatives when making a decision about where to live. By developing a probabilistic framework, we intend to understand the heterogeneity in formation of consideration sets under uncertainty, providing a more realistic and comprehensive understanding of the decision-making process in residential location choices.
 - (a) We propose a semi-compensatory decision tree-based framework to estimate and predict probabilistic consideration sets of alternatives, accommodating heterogeneity within each decision maker.
 - (b) We utilise these consideration sets to estimate discrete choice models and assess the realistic impact of various built-environment, socioeconomic and demographic variables.
2. The second objective is to develop a comprehensive framework for simultaneous determination of employment accessibility and household trips, by specific modes of drive, public transport, bicycle and walk. Many household decisions are centred around the location of employment opportunities, including travel behaviour. Understanding the relationship between employment accessibility and households' travel behaviour is crucial to discern the nuanced patterns of how households engage with their built environment. A modal distinction provides valuable insights into heterogeneous constraints and behaviours, enabling the design of targeted interventions that cater to diverse mobility needs.
 - (a) We develop a conceptual framework to test hypotheses of directionality of effect between (modal) employment accessibility and (modal) household trips, as a bundle of choices made simultaneously rather than sequentially.

- (b) We formulate a methodological framework to test and estimate the direct and indirect relationships and uncover the complementary and substitution effects at play.
3. The third objective is to adapt a control-function method to address endogeneity in residential location choice models due to housing prices and vehicle miles travelled (VMT). By developing a robust control function-based discrete choice model, we seek to disentangle the endogenous relationships between residential location choice and factors such as housing prices and travel behaviour (VMT). This aims to provide accurate and unbiased estimates of the impact of these factors on residential location decisions, enhancing our understanding of the complex interplay between economic considerations and transportation behaviour in shaping where households choose to live.
- (a) We propose adequate instruments that may be obtained from data to deal with omitted variable and measurement endogeneity.
 - (b) We develop a Control Function model to assess and address the endogeneity due to travel behaviour (fraction of trips made by public transport) using instruments such as public transport accessibility levels.
 - (c) We implement over-identification Refutability tests for discrete choice models to test for exogeneity and strength of instruments.

1.4 CONTRIBUTIONS

This thesis makes significant contributions to our understanding of residential location choices and their connection to travel behaviour. Firstly, we introduce an innovative approach that uses machine learning decision trees to handle the challenge of considering a multitude of options when choosing a place to live. These decision trees help us better understand the factors that influence decisions and allow us to filter out options that don't meet specific criteria. By combining this approach with traditional choice models, we can capture both the non-compensatory aspects of decision-making and the trade-offs people consider. This is particularly useful when there are many variables at play, making it hard to determine a clear choice. Secondly, we create a model that simultaneously examines how accessible jobs are from a household's location and how people travel to work using different modes of transportation like

cars, transit, bicycles, or walking. We account for various factors like household size and type, which can influence travel behaviour. By considering both observed and unobserved factors, we reveal how these factors interact and affect trip frequency, even when controlling for known variables. Lastly, we propose a comprehensive Control Function method to tackle endogeneity due to multiple variables, particularly related to location attributes and travel behaviour. This model addresses the challenge of understanding how personal preferences and external factors impact residential choices and travel decisions. In essence, this thesis introduces three novel models that enhance our ability to accurately estimate and understand the complex interplay between where people live and how they travel.

1.5 THESIS OUTLINE

The remainder of the thesis consists of five chapters. Chapter 2 describes the literature focussed on disaggregate residential location choice models, followed by establishing consideration of alternatives, endogeneity between residential location and travel behaviour, and the modelling approaches. Chapter 3 addresses the first objective to develop a probabilistic consideration set formation model of residential location choices, and proposes a novel implementation of a probabilistic decision tree based framework to address consideration of alternatives in a large alternative scenario, followed by description of the empirical data set used in this analysis. This chapter looks at how decision trees follow disjunction-of-conjunctions protocol, allowing estimations of non-compensatory decision process, followed by a model set-up combining decision trees-based random forests with a discrete choice model and discusses the estimated results. Chapter 4 addresses the second objective to develop a comprehensive framework for simultaneous determination of employment accessibility and household trips, by specific modes of drive, public transport, bicycle and walk, and proposes a multivariate probit model framework. We utilise Toronto metropolitan data set for empirical application and describe the methodological approach taken, followed by describing both exogenous and endogenous variable effects. Chapter 5 addresses the third objective to address endogeneity between residential location choice and travel behaviour of a household, and proposes a model framework with an instrumental variable model for addressing endogeneity between residential location and travel behaviour. In this chapter, we discuss the formulation of valid instruments, describe the Control Function

methodology, and results. Chapter 6 summarizes the contributions of this dissertation and provides insights for future research directions.

1.6 JOURNAL AND CONFERENCE PAPERS

1. Singh, A.C., A. Faghieh-Imani, A. Sivakumar, Y. Xi and E. J. Miller. A joint analysis of accessibility and household trip frequencies by travel mode (*R&R with Transportation Research: Part A*).
2. Singh, A.C., and A. Sivakumar. Semi-compensatory probabilistic model for residential location choices. 7th International Choice Modelling (ICMC), Reykjavik, Iceland, May 2022.

Additional work undertaken during PhD:

1. Sivakumar, A., A. Gough, A.C. Singh, and F. Guo. Incorporating the impacts of air pollutants and exposure to crime into accessibility-based planning: A London Case Study (*under development*).
2. Singh, A.C., H. Bouscasse, A. Sivakumar. Psychosocial Factors associated with Intended Use of Automated Vehicles: A Latent-Class and Latent-Variable Analysis. 9th Symposium of the European Association for Research in Transportation (hEART), Lyon, France, February 2021.

Chapter 2

Literature Review

The first section will provide an introduction to the disaggregate residential location choice model, the role discrete choice methods play in modelling the residential location, consideration or filtering of alternatives, and the relationship between residential location and travel behaviour. Then we summarize the findings of the research literature and the gaps identified. The next section makes a deep dive into the non-compensatory decision rules or protocols in discrete choice models, followed by a section on endogeneity in the context of residential location choice models. We conclude with a summary of findings, identifying the gaps and outlining the objectives formulated.

2.1 DISAGGREGATE RESIDENTIAL LOCATION CHOICE MODEL

The dynamics of household residential location are a fundamental driving force behind the intricate fabric of urban areas. This choice profoundly influences key aspects such as employment distribution, economic growth, social structure, spatial segregation, and the configuration of transportation systems. Consequently, accurate modelling of behavioural patterns underlying residential location choices assumes paramount significance for urban planners, policymakers, and researchers alike (Liao et al., 2015). The foundations of residential location modelling trace back to the seminal contributions of von Thünen in 1826, where he laid the groundwork for understanding the impact of transport costs on land-use distribution. His work introduced the concept of bid-rent, highlighting landowners' tendency to lease properties to the highest bidder, thus establishing the foundation for subsequent discussions on residential locations. Building on this, Alonso (1964) extended the bid-rent

concept to residential settings, contextualizing it within a mono-centric city framework focused on employment opportunities. Here, individuals and households make informed choices based on a utility function, integrating considerations such as goods expenditure, land plot size, and proximity to the city centre. Simultaneously, Lowry (1964) introduced the gravity model to residential location analysis, proposing that households distribute themselves among zones based on the attractiveness of those zones and the number of workers residing and employed within them. Residential attractiveness was closely linked to the availability of developable land within each zone.

With the advent of the current version of the discrete modelling framework by McFadden in 1978, the landscape of residential location studies underwent a transformation. This framework's strength lies in its capacity to quantify the influence of diverse residential location characteristics, taking into account their intricate interaction with household attributes. It empowered researchers to move beyond simplistic mono-centric distance-based descriptions and instead incorporate a wide array of attributes characterizing residential alternatives, thus better reflecting the rich diversity inherent in urban environments. The progression of research in this field has been closely intertwined with advancements in data availability and granularity (Schirmer et al., 2014). While earlier studies primarily relied on aggregated census data, the past few decades have witnessed a surge in disaggregated data sources, offering unprecedented insights into individual units, the built environment, and its attributes (Guo and Bhat, 2007). This shift has engendered a new wave of disaggregate models that may consider buildings units as viable choice alternatives, incorporating granular attributes to refine the understanding of residential location preferences. Nonetheless, this evolution has led to some challenges. The diverse hypotheses and attributes, coupled with a wide-array of methodological development, tested across cities, countries and continents, have made cross-study comparisons complex, hampering our ability to draw generalizable conclusions about household preferences. Thus, a need arises to explore a common set of underlying relationships of residential location that can facilitate a more unified approach to modelling residential location choices. For an overview of modelling of residential location choice with respect to the variety of variables involved, we refer the reader to the review by Schirmer et al., (2014).

2.1.1 Discrete choice models: a brief primer

The realm of discrete choice modelling has undergone a fascinating evolution, challenging conventional economic paradigms and embracing the multifaceted nature of human decision-making. This journey is marked by pivotal contributions that illuminate the intricate interplay between rationality, psychology, and the dynamic factors shaping our choices. The inception of discrete choice modelling was rooted in the standard economic model of the early 1950s, where consumers were presumed to be driven by the maximization of stable preferences over goods. However, Daniel McFadden's work exposed the limitations of this model. He unveiled the tautological nature of self-interest maximization and acknowledged the role of random elements in influencing economic behaviour. This led to a departure from the linear dependence on economic variables and an embrace of the Random Utility Maximization (RUM) model, as inspired by psychologists such as L.L. Thurstone (1927) and Jacob Marschak (1960). This model recognised the impact of experiences on preferences and introduced the Independence from Irrelevant Alternatives (IIA) axiom. McFadden's contributions flourished further with the introduction of the conditional logit or multinomial logit (MNL) model, a notable exemplar of RUM. This model's elegance lay in its connection to consumer theory, linking unobserved preference heterogeneity to demand distribution. Meanwhile, Moshe Ben-Akiva (1972) enriched this landscape with nested MNL and other RUM models, crafting a micro-econometric lens for analysing individual decisions. This intricate dance between economics and psychology unveiled the importance of psychological factors, such as perception and preference, alongside the traditional rationality-based approach.

Concurrently, the foundation of behavioural economics was laid by luminaries such as Herbert A. Simon, who championed bounded rationality and introduced heuristics in decision-making. Economists Daniel Kahneman and Amos Tversky added to this with prospect theory, redefining rational choice under uncertainty. These psychological insights have found resonance in the domain of residential location choice models, where individuals balance an array of factors such as proximity to work, schools, and amenities. Discrete choice modelling empowers us to comprehend the intricate mechanisms driving our decisions. Within this framework, consumers choose from mutually exclusive alternatives, factoring in utility, constraints, and unobserved characteristics. The heterogeneity of consumers' taste templates and perception mechanisms

enriches this landscape. This modelling journey underscores that our understanding of choice is an evolving odyssey, a synthesis of economics and psychology, paving the way for a nuanced exploration of human decision-making. Below, we outline the mathematical outline of two prominent methods in discrete choice modelling:

Multinomial Logit (MNL)

The Multinomial Logit (MNL) model is a widely used discrete choice model that assumes independence of irrelevant alternatives (IIA) and a linear utility function. For a decision maker choosing from a set of J alternatives, the probability of choosing an alternative i is given by the MNL choice probability formula:

$$P(i) = \frac{e^{V_i}}{\sum_{j=1}^J e^{V_j}} \quad (2.1)$$

where V_i is the representative utility of alternative i .

The log-likelihood function for the MNL model, assuming independent and identically distributed (IID) choices, can be written as:

$$\mathcal{L}_{\text{MNL}}(\beta) = \sum_{n=1}^N \sum_{i=1}^J y_{ni} \log \left(\frac{e^{V_{ni}}}{\sum_{j=1}^J e^{V_{nj}}} \right) \quad (2.2)$$

where y_{ni} is an indicator variable that takes value 1 if the decision maker n chooses alternative i , and 0 otherwise.

Mixed Logit

The Mixed Logit model extends the MNL model by allowing the coefficients of the utility function to be random variables with a known distribution. This allows for capturing preference heterogeneity among decision makers.

The choice probability for the Mixed Logit model is given by:

$$P(i) = \int \frac{e^{V_{ni}(\beta)}}{\sum_{j=1}^J e^{V_{nj}(\beta)}} f(\beta) d\beta \quad (2.3)$$

where $f(\beta)$ is the density function of the random coefficients vector β , and $V_{ni}(\beta)$ is the representative utility of the alternative i for the decision maker n with coefficients β .

The log-likelihood function for the Mixed Logit model is:

$$\mathcal{L}_{\text{MXL}}(\theta) = \sum_{n=1}^N \log \left(\int \frac{e^{V_{ni}(\beta)}}{\sum_{j=1}^J e^{V_{nj}(\beta)}} f(\beta) d\beta \right) \quad (2.4)$$

where θ represents the parameters of the density function and any other fixed model parameters.

2.1.2 Consideration of alternatives in residential location discrete choice models

Residential location choice is a complex cognitive decision involving a large number of alternatives. In many discrete choice model applications, such as residential and employment location choice, subsets of alternatives may be inaccessible or unknown to subgroups of the population (Haab and Hicks, 1997). Even though an analyst may observe the choice of residential location for the decision maker, she may not have access to the considered or filtered alternatives. Without revealed preference data, this consideration of alternatives cannot be empirically identified (Capurso et al., 2019). This identification issue arises either because one or more alternatives are unavailable to the decision maker or because she considered (or sampled) the alternatives endogenously, a consideration¹ process that is inaccessible to the analyst.

Disentangling the consideration set formation from the choice model is not merely an esoteric econometric issue. Earlier research on consideration set generation has indicated the important impact of consideration effects on consumer choice (Roberts and Lattin, 1991; Ben-Akiva and Boccara, 1995) by indicating that failure to properly specify the choice set considered by consumers can lead to biased choice model parameters, a lack of robustness in parameter estimates, and violations of the independence from irrelevant alternatives assumption (Williams and Ortuzar, 1982; Shocker et al., 1991). Methodologically, the incorporation of consideration effects can lead to a more accurate prediction of the choice process being modelled (Gensch, 1987; Swait, 2001). The recognition of consideration effects can help determine the relative effects of policy relevant variables on consideration and choice, and thus aid in a comprehensive understanding of the impacts of policy actions. The important point to note here is that regardless of the relative utility of an alternative compared to other alternatives in a traveller's

¹We define the *consideration process* as an intermediate step for filtering a few alternatives from the universal choice set, also known as *choice set formation process*. If this consideration process is explicitly accounted for, then the resulting subset of alternatives will hereafter be called *consideration set*.

choice set, the alternative will not be chosen if it is not first considered (Andrews and Srinivasan, 1995).

An analyst may assume a single-step and fully compensatory behavioural process (Bierlaire et al., 2010) to describe a majority (if not all) of choice decisions. Almost all discrete choice models used in practice and research are based on compensatory decision processes (Martinez et al., 2009), with utility-maximization being the most common example. Such single-step fully compensatory behavioural models hypothesize that the decision makers compensate by making trade-offs between attributes and across all alternatives (Paleti, 2015). However, counter to prevailing practices, behavioural economists (Manski, 1977) and psychologists have presented much evidence that individuals frequently depart from standard notions of utility maximization and rationality. Observations that challenge the traditional notions of rationality and utility maximization have led to a growing body of research on alternative models of human decision-making. These models explicitly incorporate the possibility of non-utility maximizing behaviour and are known as non-compensatory decision rules (Araña and León, 2009).

2.1.3 Residential location choice and travel behaviour

Land use patterns refer to the way urban developments are spatially distributed, providing opportunities for various activities, while, the transport network facilitates the movement of the population by providing the necessary resources for travel patterns (Ortuzar and Willumsen, 2011). Land use and transport models (ILUT) capture the interdependence between land use patterns and travel demand (Moore and Thorsnes, 1994). Early models, such as gravity models (Wilson, 1970 and 1971), estimated interaction between zones based on size and distance. Econometric models, such as the Lowry Model (Lowry, 1964), addressed limitations of gravity models. Random utility theory based discrete choice models (DCMs) allowed sensitivity analysis for transport modes and residential location decisions (McFadden, 1978). Microsimulation models, such as UrbanSim (Timmermans, 2003) and ILUTE (Farooq et al., 2008), utilized agent-based modelling techniques to capture complex urban systems. The SILO model (Moeckel, 2017) focused on transport choices and incorporated budget constraints and modular modelling. The complexity of both land use and transport models has increased to accommodate behavioural realism and available data (Miller, 2006).

While a detailed overview of ILUT models is beyond the scope of this research, it is important to appreciate the nature of relationship detailed in these frameworks. In summary, these models estimate and predict individual and household choices that influence land use and travel demand. These include short-term choices such as mode choice and destination choice, medium-term choices such as vehicle and bicycle ownership and long-term choices such as residential and work location. While the long-term choices (residential and employment location) subsequently influence land use patterns, the short-term and medium-term choices (vehicle ownership, mode, destination) influence travel demand (Haque et al., 2019). There is a plethora of studies reviewing ILUT models and the features of various models, with a notable summary synthesized by Hawkins and Narul Habib (2019). It can be argued that residential location and travel behaviour are ideal candidates to understand and estimate the relationship between long-term and short-term choices that govern the land use and transport. However, many residential location choice models lack modularity and integration with travel behaviour models. Considering the relationship between residential location and travel behaviour, joint frameworks that evaluate their impact and provide realistic outcomes are well warranted. Developing and testing models within these frameworks can lead to innovative and realistic modelling approaches.

2.1.4 Impact of Residential Location on Travel Behaviour, and vice versa

Households residing in urban or mixed-use neighbourhoods tend to prefer non-motorized and public transportation modes over driving, however, those residing in suburban or rural neighbourhoods are shown to drive more as compared to other modes (Handy, Cao, and Mokhtarian, 2005; Mokhtarian and Cao, 2008; Cao, Mokhtarian, and Handy, 2009). These variations in travel behaviour and the frequency of driving may be attributed to the differences in travel distances influenced by the built environment provided by the residential location (Krizek, 2003; Scheiner and Holzrau, 2013; Aditjandra et al., 2016; Oakil et al., 2016). It is important to note that the choice of residential neighbourhood is likely to have an impact on travel patterns, even after accounting for self-selection effects (as studied by Cao et al., 2009 and discussed in subsequent sections). Thus, changes in both residential (re)location and the built environment (such as land use mixing) may impact travel behaviour.

While the studies cited above account for built environment attributes, such as job

opportunities, education, healthcare, and lifestyle benefits, it can be argued that travel attitudes and behaviour may also impact residential location choices (De Vos and Ettema, 2020). Other residential location studies have included factors such as housing or rental costs, dwelling attributes, proximity to work, family and friends, and neighbourhood safety and quality (Schirmer et al., 2014). There is evidence that questions the significance of travel behaviour in residential location choices (Ettema and Nieuwenhuis, 2017; Wolday et al., 2018), supported by inconsistencies between travel attitudes and the chosen residential neighbourhood (Schwanen and Mokhtarian, 2005; De Vos et al., 2012). While the impact of factors such as housing costs and proximity to work, have a more noticeable impact on residential location decision-making, the effect of travel behaviour (attitudes and patterns) may not be insignificant (Lin et al., 2017). Figure 2.2 (derived from De Vos and Ettema, 2020) provides the directionality-hypotheses tested for statistical significance.

However, even in the studies that either discuss the need for disentangling this relationship, the focus is limited to the criteria of statistical association and non-spuriousness, without any testing for time precedence and directional effects, as described necessary for robust inferences (Singleton and Straits, 2005). Further, probing into literature about bi-directionality structure reveals that out of 38 studies analysing the self-selection problem as reviewed by Cao, Mokhtarian, and Handy (2009), only two attempted to identify mechanisms beyond statistical correlations. Næss (2015) states:

“In contemporary correlation-based research on land use and transport, qualitative properties of the built environment as well as the humans whose actions are conditioned are put in the background. Theories explaining why correlations exist between built environment characteristics and travel behaviour are rarely exposed, let alone reflected on.”

2.1.5 Accessibility and Travel Behaviour

Travel behaviour exhibits a complex relationship with built-environment attributes, which has been extensively studied in the literature (Cervero and Kockelman, 1997; Frank and Engelke, 2001; Bagley and Mokhtarian, 2002; Cervero and Duncan, 2003; Krizek, 2003; Shay and Khattak, 2005; Lee and Moudon, 2006; Copperman and Bhat, 2007; Vance and Hedel, 2007; Brownstone and Golob, 2009; Cao et al., 2009; Cao, 2010; Zhang et al., 2012; Hong et al.,

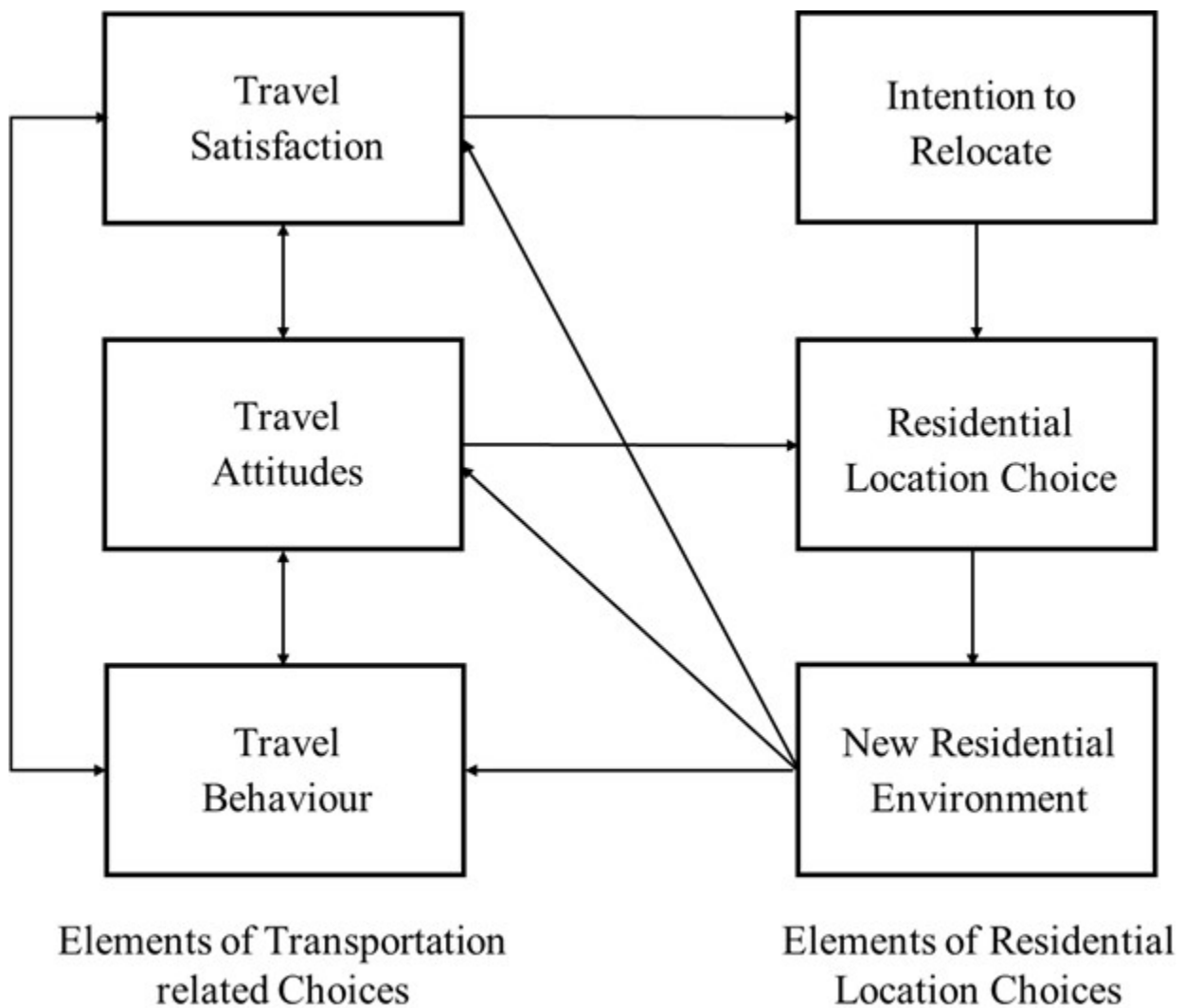


FIGURE 2.1 Cycle of stipulated choice-making process, adapted from De Vos and Ettema (2020)

2014; Liao et al., 2015; Næss et al., 2018). The neighbourhood-level accessibility, measuring the impact of the immediate built environment on travel behaviour, may be homogenous across different travel modes, but the influence becomes disproportionate beyond the immediate vicinity. This can be observed through variations in accessibility levels across different travel modes. While several earlier studies have focused on immediate neighbourhood characteristics such as the availability of facilities, land use density, and street network features (Handy, Cao, and Mokhtarian, 2005; Mokhtarian and Cao, 2008; Cao et al., 2009; Singh et al., 2018), there are relatively fewer studies that assess the impact of built environment beyond the immediate neighbourhood (Krizek, 2003; Ewing and Cervero, 2010). In fact, one can argue that focusing on network-level accessibility can lead to a wider understanding of the factors influencing travel behaviour.

The network-level accessibility is defined as the capacity to reach valued destinations where opportunities, activities, and individuals are located. It is typically quantified based on travel impedance and the number of reachable opportunities (Xi, Miller, and Saxe, 2018). Ewing and Cervero (2010) observed that accessibility to destinations strongly affects household travel behaviour. Accessibility to job opportunities within an isochrone is the most common measure of accessibility used in regional planning (Boisjoly and El-Geneidy, 2017) and can be found in the analysis of land value (Du and Mulley, 2012; Iacono et al., 2008), commuting distance (Levinson, 1998; Manaugh et al., 2010), social equity (Manaugh and El-Geneidy, 2012; Foth et al., 2013) and mode share (Foth et al., 2014; Owen and Levinson, 2015). Therefore, we can hypothesise that the observed levels of network-level accessibility (represented by cumulative opportunities approach measured by the number of jobs reachable within the isochrones) by different travel modes have a substantial impact on household travel behaviour. We must investigate the direct and indirect effects between travel behaviour and the level of accessibility by empirically analysing the direction and extent of their mutual influence, focusing on multi-modal household trip frequencies and accessibility.

2.1.6 Summary

Firstly, there is a notable lack of methodologies aimed at mitigating the inherent dependence between alternatives during the consideration set formation process. Addressing this challenge is crucial to enhance the accuracy and reliability of residential location choice models. Secondly,

there exists a need to more effectively incorporate non-compensatory decision-making behaviour into the modelling framework. This entails developing approaches that capture how individuals make choices based on specific attributes without assuming compensatory trade-offs between different factors. Lastly, the empirical scenario of decision makers facing a large and diverse set of alternatives remains a relatively unexplored area. Developing strategies to handle such complex decision processes is essential to ensure the practical applicability and robustness of residential location choice models in real-world contexts.

We also identify a significant research gap concerning the handling of simultaneity endogeneity in the context of residential location choice modelling. A crucial challenge lies in establishing a clear and differentiated understanding of the intricate interplay between a household's residential location decisions and their observed travel behaviour. Specifically, there is a need to develop methodologies that can disentangle and quantify both direct and indirect relationships between these factors, thereby uncovering the underlying mechanisms of complementary and substitution effects that influence the decision-making process. Addressing this research gap is essential for advancing our comprehension of the complex dynamics between residential choices and travel behaviour, ultimately contributing to more accurate and robust residential location choice models.

Finally, the literature surrounding residential location and travel behaviour reveals a critical research gap related to the management of omitted-variable and/or measurement endogeneity within the framework of residential location choice models. The identification and mitigation of endogeneity, particularly concerning travel behaviour variables, present significant challenges that warrant exploration. Detecting such endogeneity in the context of residential location choice models remains an underdeveloped area of research, calling for innovative approaches that can accurately diagnose its presence. Furthermore, addressing this research gap involves devising robust methodologies to effectively account for and correct endogeneity biases, ensuring the attainment of unbiased and reliable outcomes in residential location choice analyses. Resolving this gap is pivotal in advancing the accuracy and validity of residential location choice models, offering substantial contributions to urban planning and policy decision-making.

2.2 NON-COMPENSATORY DECISION RULES IN DISCRETE CHOICE MODELLING

Non-compensatory rules involve filtering (selecting or rejecting) alternatives based on a specific attribute, without necessarily considering all the attributes of all the alternatives (Hauser et al., 2009). Such rules also do not allow positive attributes to compensate for negative attributes of the same alternative. Non-compensatory models are better suited to capturing the limited cognitive resources of decision makers and are considered more behaviourally realistic (Swait, 2001). In discrete choice methods, conjunctive or disjunctive rules (and their variations) are typically employed to model non-compensatory decision processes (Gilbride and Allenby, 2004). A conjunctive rule dictates that an individual only considers alternatives that meet all of a given number of requirements, while a disjunctive rule dictates consideration of alternatives that meet at least one of a given set of requirements. Disjunction-of-conjunctions protocol allows any alternative that meets at least one of a given set of conjunctive conditions, where each condition may differ in the number of requirements that compose the conjunction. Algebraically, then, disjunctions-of-conjunctions are expressed as sums of products with no constraints on the number of elements in each product. The limited application of non-compensatory decision processes in choice modelling can be attributed to difficulties in estimation (Stüttgen et al., 2012; Kohli and Jedidi, 2017). Researchers circumvent this limitation by linking disjunctions-of-conjunctions to decision trees (Brathwaite et al., 2017; Gutiérrez-Vargas et al., 2023), which draws upon well-established methods for estimating decision trees. With the ability to estimate disjunctions-of-conjunctions, it becomes possible to estimate these strategies in combination with compensatory procedures used in standard discrete choice models.

The literature describes three different approaches to decision-making. The *first approach* deals with predicting the probability of an individual's choice decision employing a single-stage model that generates a choice set implicitly, known as the reduced-form formulation (Swait, 2001b). It could also be assumed that there is only a reject or accept state, and the systematic utilities are compared to estimated thresholds to determine the 'state' of an alternative (Zhu and Timmermans, 2010). A drawback of this approach is that it can become unquantifiable for cases with more than two alternatives. There is uncertainty about the outcome when multiple alternatives are categorized as 'accepted,' and the procedure to select a specific choice from among these acceptable alternatives remains ambiguous. The decision tree models (Arentze

and Timmermans, 2004, 2007) predict the probability of a given alternative using decision trees combined with standard discrete choice models. These analyses were motivated mostly by an attempt to estimate the effect of discrete variables on one's systematic utilities using a non-parametric function that is adept at detecting interactions. In their combination of the decision tree with a standard discrete choice model, the decision tree is estimated based only on the explanatory variables that are originally discrete, and then a dummy variable for each output node of the tree is added to the systematic utilities of the various alternatives. The coefficients of these dummy variables are then estimated, along with the usual parameters of one's choice model.

The *second approach* employs non-compensatory decision rules to derive a consideration set, followed by a compensatory choice model. A vast majority of models have relied on the Probabilistic Independent Availability Logit (PIAL) model (Swait, 1984, 2009), which uses non-compensatory decision rules to determine whether each alternative will be present in an individual's choice set. PIAL models try to explain how people choose between different options, based on the attributes of each option. However, there is often uncertainty about how people use these attributes to make decisions. This uncertainty creates randomness in the model, which can be difficult to deal with. PIAL models assume that the probability of choosing one option is independent of the probability of choosing any other option, but this assumption can make the model very complex. To overcome this complexity, researchers have developed different methods for estimating PIAL models. This complexity does not arise in the two stage framework prescribed by Manski (1977). Similarly, Gilbride and Allenby (2004) employ a two-stage model with conjunctive and disjunctive rules. Their method estimates the probability of an option being chosen based on its attributes, and then uses a Bayesian estimation method to explore the space of possible attribute thresholds. This helps to reduce the number of possible choice sets that need to be considered. However, most models since Gilbride and Allenby have taken a different path, using frequentist methods instead of Bayesian methods (Cantillo and Ortuzar, 2005). These methods can also be effective, but they may be more complex and computationally intensive.

We also discuss various methods used to overcome the computational difficulties of PIAL models, including enumeration, simulation-based optimization, Bayesian estimation, and subset-conjunctive rules. In a frequentist model, Cantillo and Ortuzar (2005) used

standard maximum likelihood estimation to estimate a PIAL model. They enumerated all possible consideration sets, which incurred estimation difficulties due to the curse of dimensionality. The advantage of this procedure is that it allowed to parameterize the attribute thresholds as a function of socioeconomic variables and choice conditions, resulting in more behaviourally interpretable thresholds. Jedidi and Kohli (2005) also used a PIAL model that allowed for individual heterogeneity using latent classes and accommodated uncertainty in the number of requirements that need to be satisfied. Their formulation involved full enumeration of all possible choice sets under each possible set of criteria and number of requirements. Swait (2009) later developed a two-stage choice model called a k-Mix model, which is also a PIAL model, with the ability to use conjunctive or disjunctive rules to place alternatives in a ‘dominance’ or ‘rejection’ state.

The *third approach* elicits information from individual decision-makers, which allows the analyst to specify the decision-maker’s preference structure (Hess and Daly, 2014). When the thresholds used by an individual are observed, this allows their consideration set to be specified with certainty (Kaplan et al. 2009, 2012). Zolfaghari et al. (2013) also estimate a model according to this approach but allowed for the possibility that individuals do not use all the attribute thresholds, similar to Jedidi and Kohli (2005) model. The authors account for uncertainty by fully enumerating all possible combinations of criteria. This leads to a formulation similar to a subset-conjunctive rule, with uncertainty over the number of criteria that must be met. While most models mentioned so far focussed on either conjunctive or disjunctive rules, Jedidi and Kohli (2005) allowed for subset-conjunctive rules, but none of the models allowed for disjunctions-of-conjunctions. The decision tree framework has the potential to address this issue by using sets of conjunctive conditions that are mutually exclusive, ensuring that no observation is ever described by more than one condition (Breiman et al., 1984).

Our review highlights the limitations of the first and second approaches, such as the reduced-form approach’s lack of clarity when there are more than two alternatives and the curse of dimensionality in PIAL models. In contrast, the third approach is considered advantageous in that it can elicit a decision-maker’s preference structure, which can lead to greater behavioural interpretation of estimated thresholds. Lastly, we note that some researchers have tried to minimize the number of possible consideration sets by collecting explicit consideration set information from decision-makers. The issue is that the

aforementioned models primarily represent only conjunctive or disjunctive rules. Only the model by Jedidi and Kohli (2005) allowed for subset-conjunctive rules, and none of the models allowed for disjunctions-of-conjunctions. The decision tree models described later avoid this issue by using sets of conjunctive conditions that are all mutually exclusive, thus ensuring that no observation is ever described by more than one condition. With the help of decision trees, a model framework can probabilistically formulate a consideration set in a non-compensatory protocol, followed by a compensatory decision rule for choice decisions. This could very well account for the unbiased mechanism by which individuals explore and evaluate a large set of alternatives in residential location choice scenarios.

2.2.1 Summary

From this literature review, several key insights emerge regarding non-compensatory decision-making processes and their application in residential location choice modelling. Non-compensatory rules, which involve selecting or rejecting alternatives based on specific attributes without considering all attributes, are considered more behaviourally realistic due to their alignment with limited cognitive resources. While the reduced-form approach predicts individual choice probabilities using a single-stage model by implicitly generates a choice set, its effectiveness diminishes when faced with more than two alternatives due to uncertainties in handling multiple acceptable choices. While we can collect explicit information from individual decision-makers to ascertain their choice set, it is however impractical using observed or revealed preference data and challenges remain in incorporating diverse decision-making rules beyond conjunctive or disjunctive forms. Finally, Manski's two-stage framework determines whether alternatives are included in the choice set based on non-compensatory criteria. However, complexities arise due to uncertainties in individuals' attribute utilization, challenging model estimation. There is the necessity for a model framework that probabilistically establishes an individual's consideration set, followed by a compensatory choice mechanism. Such a framework is imperative to comprehensively address the unbiased exploration and evaluation of diverse alternatives in residential location choice scenarios.

2.3 ENDOGENEITY IN THE CONTEXT OF RESIDENTIAL LOCATION CHOICE MODELS

From a statistical or econometric perspective, an endogeneity problem occurs when there is a correlation between an exogenous variable and the error term in the model (while, in non-endogenous models, the error term is non-correlated with the exogenous variables, such that the error term in model is due to all unobserved factors in dependent variable that is not due to the observed exogenous variables in model). In other words, endogeneity or non-orthogonality occurs when an independent variable is correlated with the error term. It is important to understand and decouple endogeneity as it may cause model misspecification, thereby rendering the estimated coefficients of a linear or a nonlinear model as biased and inconsistent (Kitamura et al., 2009). Firstly, endogeneity in this broad sense can be caused by omitted variables, termed as *unobserved or omitted variable heterogeneity*. This can be dealt with the availability of the omitted variables or proper substitutes (instruments). Secondly, there are other possible sources of endogeneity that may be difficult to address, such as simultaneity endogeneity – such that, X causes Y, but Y also causes X – and self-selection endogeneity.

Endogeneity can be observed in several types of discrete choice models used in land use and transport analysis: residential location choice (Guevara and Ben-Akiva 2006; Guevara 2010), travel behaviour decisions (Buehler, 2008; Kitamura et al., 2009; Sivakumar and Pinjari, 2012), mode choice (Fernández-Antolín et al., 2016), EV demand (Helveston, 2016), airline ticket choice (Lurkin et al., 2017), passenger booking timing (Wen and Chen, 2017), machine learning models of route choice (Guevara et al., 2017), and mobility data collection (Zegras et al., 2018). Addressing endogeneity is important because it leads to inconsistent parameter estimates. These estimates can yield wrong policy implications, such as value of travel times can be overestimated due to confounding effect of omitted comfort in the mode choice example.

Endogeneity poses a significant challenge in the realm of residential location discrete choice models, stemming from attributes such as price and the relationship between residential location and travel behaviour. At its core, endogeneity arises when these attributes are jointly determined with the choice of residential location, creating a complex feedback loop that challenges traditional model assumptions. In the case of price, for instance, housing costs may not only be influenced by location preferences but also directly impact the choice of where to reside. Similarly, the connection between residential location and travel behaviour,

notably vehicle miles travelled (VMT), is bidirectional - where individuals' chosen locations can influence their travel patterns, and conversely, travel options and accessibility can shape the attractiveness of different residential areas. These dynamics challenge the exogeneity assumption, necessitating solutions such as control function approaches to ensure robust inference in residential location choice models. By addressing endogeneity, these models can offer insights into the true drivers of residential decisions and contribute to more accurate policy recommendations in urban planning and transportation management.

2.3.1 Methods to address endogeneity in DCMs

An important consideration to model development is future forecasts and policy design based on the relationship between residential location and travel behaviour. Forecasts for policy design may be seriously misinformed if the model estimation is inconsistent. An inconsistent estimation approach may invalidate model analysis and has been noted to be unavoidable in models of residential location choice (Guevara and Ben-Akiva, 2006) because of endogeneity, unless it is addressed methodically. However, given the substantial simultaneity endogeneity between residential-location and travel behaviour attributes, it is fitting to develop models which could jointly model these as a bundle of choices made simultaneously rather than sequentially. We refer the reader to Guevara (2015) for a description and evaluation of five methods to correct for endogeneity in discrete-choice models: the use of Proxys (PR); the two steps Control-Function (CF) method; the simultaneous estimation of the CF method via Maximum-Likelihood (ML); the Multiple Indicator Solution (MIS); and the integration of Latent-Variables (LV). A large literature employs the BLP method (Berry et al., 1995), which uses a random coefficient logit (RCL) model with structural error terms to allow for unobserved product characteristics, and which handles endogeneity of prices and market shares. Substitution patterns are determined by a random coefficient specification of the distribution of unobserved preference heterogeneity, which means that complementarity is ruled out. The BLP method involves a non-linear, non-convex optimization problem as well as the simulation and numerical inversion of the demand function.

The control function (CF) may be the most appropriate approach to address endogeneity in residential location DCMs (Guevara, 2015). Generally, CF approach is a two-stage (sequential) procedure, but can also be simultaneously estimated (Guevara and

Ben-Akiva, 2009). A latent variable (LV) approach may also be employed to model omitted attributes using indicator variables (Rutz and Watson, 2019). Let's consider a residential location model with endogeneity due to travel behaviour variable q . Assuming the true linear utility as:

$$U_{ij} = \beta_x x_{ij} + \beta_q q_{ij} + e_{ij} \quad (2.5)$$

where, β_q and β_x are parameters to be estimated, x_{ij} and q_{ij} are explanatory variables of the model with e_{ij} as the exogenous error term. The variable q_{ij} may be endogenous, either due to omission or correlation with the error term. Now, let us consider that the k^{th} term of x_{ij} is correlated with q_{ij} as follows:

$$x_{kij} = \gamma_0 + \gamma_z z_{ij} + \gamma_q q_{ij} + \phi_{ij} \quad (2.6)$$

where ϕ_{ij} is the exogenous error term, z_{ij} is the exogenous attribute working as an instrument and independent of e_{ij} . For the model to be identifiable, there must be (at least) as many IV as endogenous variables in the model (Guevara and Ben-Akiva 2012). Following the assumption that variable q_{ij} may be endogenous, the specification needed to treat endogeneity would be:

$$x_{kij} = \gamma_0 + \gamma_z z_{ij} + \theta_{ij} \quad (2.7)$$

where the error term θ_{ij} contains both ϕ_{ij} and q_{ij} . If valid, such that z_{ij} is truly exogenous, then θ_{ij} will capture the entire part of x_{kij} that is endogenous. Therefore, residential location discrete choice model corrected using CF approach would have utility:

$$U_{ij} = \hat{\beta}_x x_{ij} + \beta_{\hat{\theta}} \hat{\theta}_{ij} + \tilde{e}_{ij} \quad (2.8)$$

In practice, CF follows two stages:

1. Obtaining residuals $\hat{\theta}_{ij}$ by applying Ordinary Least Squares (OLS) regression to x_{kij} on z_{ij}
2. To estimate discrete choice model considering $\hat{\theta}_{ij}$ and x_{ij} attributes

We may obtain consistent estimators $\hat{\beta}_x$ for the β_x up to a scale (Guevara and Ben-Akiva 2012). There could also be a loss of efficiency due to two-stage estimation, unless the

error terms are homoscedastic and non-autocorrelated (Rivers and Vuong, 1988). We may also face the issue that the standard errors cannot be derived from the information matrix, hence requiring bootstrapping in the two-stage process. However, "sequential CF method is robust to misspecifications of the distributional assumptions of the model" as discussed by Guevara (2015).

2.3.2 Summary

This literature review highlights the critical impact of endogeneity on residential location choice modelling. The intricate interplay between residential location choice and attributes such as price or travel behaviour creates a challenging feedback loop that traditional models struggle to address, leading to biased estimates and inconsistent results. The review underscores the necessity for innovative methods, such as instrumental variables and control function approaches, to disentangle the complex relationships and produce robust models. Our review highlights the strength of CF methods, particularly dealing with residential location discrete choice model's endogeneity due to multiple endogenous variables such as alternative-attributes (such as price) and household travel behaviour attributes (such as vehicle miles travelled).

2.4 OVERALL SUMMARY OF FINDINGS AND PLAN AHEAD

In this chapter, we started with an overview (section 2.1) of disaggregate residential location choice models, implementation of discrete choice methods, consideration of alternatives, the relationship between residential location and travel behaviour, and the role of accessibility in this relationship. This was followed by the review of advances in non-compensatory decision rules and protocols in discrete choice models (section 2.2) and endogeneity in DCMs (with limited methodological review) and that in residential location due to travel behaviour (section 2.3). Each of the section summarized its findings and how they help identify and motivate the research gaps previously described in Chapter 1.

These research gaps can be addressed by objectives outlined here. Firstly, a probabilistic model is proposed for the formation of consideration sets in residential location choices. This entails a semi-compensatory decision tree-based framework to estimate and predict unique consideration probabilities for each decision maker. Subsequently,

consideration sets tailored to individual decision makers are developed, enabling the estimation of discrete choice models that evaluate the impact of various variables. Secondly, a comprehensive framework is introduced to simultaneously determine employment accessibility and household trips using different transportation modes. This involves testing directional effects and employing a methodological framework to analyse direct and indirect relationships, identifying complementary and substitution effects. Thirdly, the research addresses endogeneity issues between residential location choice and travel behaviour. Proposed solutions include utilizing appropriate instruments to manage omitted variable and measurement endogeneity, implementing a Control Function model to address endogeneity arising from travel behaviour, and employing over-identification Refutability tests for exogeneity assessment and instrument strength evaluation. These objectives collectively strive to enhance the accuracy and reliability of residential location choice models, incorporating complex decision-making dynamics and accounting for potential endogeneity concerns.

Further specific literature review directly addressing each research gaps and objectives are presented in the individual chapters, such as Section 3.1.2 provides review of representation of disjunction-of-conjunctions by decision trees, Section 4.1.2 provides review of simultaneity bias in accessibility and travel behaviour, and Section 5.1.2 provides review of the methods to address endogeneity in residential location discrete choice models. These satellite literature review sections continue the theme of these individual problem statements and provide a further deep-dive into each of these research gaps.

Chapter 3

Semi-compensatory probabilistic model for residential location choices

Abstract

This chapter presents a novel approach to address the consideration of alternatives in residential location choice by employing probabilistic decision trees. By incorporating non-compensatory decision-making protocols, the decision trees accurately predict a household's consideration set, providing insights into the factors influencing residential location choices. The proposed model framework combines probabilistic decision trees with a traditional discrete choice model to capture both non-compensatory consideration and compensatory choice decisions, using data from the 2018-2019 London Travel Demand Survey. The results shed light on the factors contributing to consideration set formation, demonstrating the relative impact of socioeconomic and demographic factors on residential location choices. The probabilistic decision tree predictions allow for the identification of alternatives that are most likely to be evaluated by decision-makers, enhancing the accuracy of predictions. The analysis also reveals the interdependence among alternatives within specific Transport Fare Zones, uncovering correlations between alternatives. Furthermore, the study uncovers intriguing findings, such as the negative impact of housing prices on choice probabilities for lower-income single-person households, which increases as income rises. It also highlights the varying effects of population density, crime density, and racial ethnicity on location choices, providing valuable insights into housing segregation patterns in London.

Overall, this research contributes to a deeper understanding of residential location considerations and provides policymakers with valuable information to promote equitable residential neighbourhoods. The rest of the chapter is presented with an introduction to the consideration of alternatives in residential location choice models, followed by an overview of use of decision trees in representing disjunction-of-conjunction protocols, description of the data, methodological framework, estimation results and conclusions.

3.1 INTRODUCTION

3.1.1 Background and context

Residential location choice is a multifaceted cognitive process, with discrete choice models commonly employed to estimate and predict these decisions. However, these models can be hindered by missing data on consideration of alternatives. This empirical blind spot emerges due to either unavailable alternatives or an endogenous consideration process inaccessible to the analyst. The conventional models, estimated using incomplete data, yield biased outcomes. To surmount this challenge, an analyst may assume a fully compensatory behavioural process to describe the majority of choice decisions. Nevertheless, empirical evidence suggests that households often deviate from conventional utility maximization, therefore motivating alternative models of human decision-making that allow for non-compensatory behaviour. Non-compensatory decision rules, which filter alternatives based on specific attributes without requiring full consideration of all attributes, present a more behaviourally realistic approach. Such models reject the notion of positive attributes compensating for negative ones within the same alternative. Although non-compensatory models better reflect the cognitive limitations of decision makers, their limited application in choice modelling arises due to estimation complexities. We can overcome this hurdle by linking non-compensatory rules to decision trees, facilitating estimation.

3.1.2 Representation of disjunction-of-conjunctions by decision trees

Decision trees¹ are a sequence of "if-then" statements that aim to forecast a certain variable (Loh, 2011). To achieve this, these statements must split the explanatory variable space into a range of mutually exclusive areas that span the entire space of explanatory variables, thereby creating the output nodes (Lemon et al., 2003). When making predictions about a decision-maker, the "if" statement determines the region/output-node that the decision-maker is in, and the corresponding "then" statement provides the desired prediction. In the context of discrete choice, such predictions might include the probability that a particular alternative is being considered, or the probability with which an alternative is chosen. Deterministic choice sets may be elicited in certain residential location choice scenarios, where housing options are excluded based on income constraints (Kaplan et al., 2012; Zolfaghari et al., 2013; Bhat, 2015). However, for all other cases, we must employ a "probabilistic" or "soft" decision tree,

where the output is a probability of consideration. Given a decision tree, it is possible to compute probability statements that are similar to those in two-stage choice models with non-compensatory rules (Gilbride and Allenby, 2004; Cantillo et al., 2006). When using a decision tree to model alternative consideration, the probabilities of interest can be represented as a disjunction-of-conjunctions or a summation of products. This decision protocol is more general than the conjunctive or disjunctive rules that are commonly used in non-compensatory choice models.

The trees used in decision-making can vary significantly. Some algorithms, such as Classification and Regression Trees (CART) and C4.5, aim to create partitions that are as homogeneous as possible in terms of the dependent variable by growing a large tree and then pruning it using cross-validation to minimize the test error. In other words, they focus solely on maximizing the accuracy of predictions (Breiman et al., 1984; Quinlan, 1993). On the other hand, more recent decision trees, such as Fast and Accurate Classification Tree (FACT), Quick, Unbiased and Efficient Statistical Tree (QUEST), Generalized, Unbiased, Interaction Detection and Estimation (GUIDE), and Conditional Inference Trees algorithm (CTREE), incorporate statistical tests into the tree-growing process (Loh and Vanichsetakul, 1988; Loh and Shih, 1997; Loh, 2002; Hothorn et al., 2006). For example, FACT and QUEST use parametric models at the end leaves of the tree, while GUIDE and CTREE introduced statistical tests to aid in the tree-growing process. In the ‘Model-Based’ (MOB) recursive partitioning algorithm, structural breaks of the fitted model’s score functions are used. After the split variable is selected based on tests for parameter stability, the MOB algorithm chooses the split point by maximizing the sum of log-likelihood functions in the emerging subgroups. However, unlike the GUIDE algorithm, the MOB algorithm requires the declaration of different variables to be used as splitting variables and as regressors (i.e., to be included in the parametric model defined at the end leaves) (Zeileis et al., 2008).

The discrete choice community is increasingly interested in data-driven methods, as indicated by a recent literature review by Hillel et al. (2021) that found more than 70 articles using machine learning techniques. Several researchers have used decision trees to model travel mode choice, including Karlaftis (2004), who developed a multivariate recursive

¹This subsection provides an overview of the literature representing disjunction-of-conjunctions by decision trees. This follows the previously described Section 2.2 “Non-compensatory decision rules in discrete choice modelling” from Chapter-2

partitioning algorithm that outperformed traditional MNL models in predictive power. Tang et al. (2015) and Yang et al. (2021) also used decision trees to model travel mode choice but did not focus on the interpretability of their models. Arentze and Timmermans (2007) developed the parametric action decision tree (PADT) algorithm, which grows the tree in two steps, using discrete attributes to produce the tree and continuous attributes to model utility at the end leaves. Brathwaite et al. (2017) used a Bayesian model tree to model bicycle mode choice, accommodating different types of non-compensatory behaviour. The MOB algorithm is another decision tree method that can include a parametric model at the end leaves and is able to include random coefficients in the model specification, making it more flexible than previous decision trees. Cockx and Canters (2020) applied the MOB algorithm to revealed preference data for residential location choice in Belgium and identified heterogeneous preferences in residential location. However, they used an MNL model and did not compare the performance of the MOB algorithm with other commonly used discrete choice models or perform simulations to assess its performance. Gutiérrez-Vargas et al. (2023) provide another application of the MOB algorithm in a discrete choice context, using a MIXL model at the end leaves of the MOB tree structure. showed that the MOB-MIXL model obtained a better model fit than the MIXL model fitted to the whole sample. Additionally, the authors compared the MOB-MIXL results with other models commonly used in discrete choice applications, such as LC and LC-MIXL models. Using the MOB algorithm as a variable selection step, they were able to outperform the latent classes with constant-only allocation models in terms of information criterion. Furthermore, they observed that the parameter estimates produced by the LC and LC-MIXL models were much more different across the classes than those produced by the MOB algorithm.

3.1.3 Contributions of this chapter

This chapter proposes to address the residential location consideration problem by employing probabilistic decision trees (also known as soft or fuzzy decision trees) to predict an household's consideration set. The decision trees resolve the ambiguity in deriving exact conditions for each observation by using disjunctions-of-conjunctions decision rules (Brathwaite et al., 2017). Our model framework combines probabilistic decision trees with a traditional choice model to account for non-compensatory consideration of choice alternatives followed by a compensatory

choice decision. To the best of our knowledge, this is the first application of probabilistic decision trees to model consideration process in a residential location choice context. This method ensures that no observation is ever described by more than one conjunctive condition. We estimate the probability statements as a summation of products, the output of which is the consideration probability of an alternative. These consideration probabilities across all alternatives are then utilized to form consideration sets, conditional on which a Mixed Logit (MxL) choice model is estimated. The data for this analysis is sampled from 2018-19 London Travel Demand Survey (LTDS). LTDS is a continuous household survey of the London area with an annual sample size of approximately 8,000² households. It captures information on households, people, trips, and vehicles and therefore allows for detailed analysis of household residential location choice (including socio-demographic factors) and its relationship to travel behaviour. This research study further aims to build on these initial findings by (1) making probabilistic predictions with higher accuracy, (2) representing heterogeneity in a population's non-compensatory rules, (3) accommodating large numbers of alternatives, and (4) alleviating the independence in consideration set formation. The contribution of this chapter is twofold: 1. in addition to a deeper understanding of factors that contribute to residential location choice, this analysis will also provide us factors that lead to consideration of neighbourhoods based on classes described by income level, vehicle ownership and household structure, among other socioeconomic and demographic factors; 2. this information can be utilized for policies that can lead to equitable residential neighbourhoods and provide a deeper perspective into causes of residential segregation in London.

3.2 DATA DESCRIPTION

The data used for this study is a sample taken from the 2018-2019 London Travel Demand Survey (LTDS). The LTDS is a household travel survey collecting travel information from approximately 8,000 households in the London metropolitan region. The LTDS collects household and household socio-economic and demographic information in addition to one day travel diary data (for each person above the age of five) regarding the trips made in a day, such as travel mode, start time, etc. After removing the households with incomplete and

²The original LTDS data is cleaned and additional data for price, accessibility and crime is merged, as described in the Data Description section, which results in a reduced number of 4,866 households available for this analysis

inconsistent attributes, the remaining 4,866 households residing in the London were employed from the data. These households reside in the 31 boroughs³ of London. The dependent variable of interest in this analysis is the residential location choice at the aggregate level of the Borough. Household, person, vehicle and trip data from LTDS were used to formulate one aggregate data set of location choice and travel behaviour at the household level.

This household data was supplemented with housing price data from Bennett et al. (2023) who utilized spatio-temporal models to correlate mortality and house prices for small geographies. Further, we added public transport accessibility data obtained from Public Transport Accessibility Levels (PTALS). The PTALS measure the accessibility of a point to the public transport network in Greater London by taking into account the walk access time and service availability. It is a way to measure the public transport network density at any location in Greater London. Each area is assigned a grade between 0 and 6b, where 0 means very poor access to public transport and 6b indicates excellent access. The methodology was created by the London Borough of Hammersmith and Fulham in 1992 and has been thoroughly reviewed and tested. The PTAL development group led by the London Borough has agreed that it is the most appropriate methodology to use across London.

Table 3.1 provides an overview of the descriptive characteristics of the sample. The independent characteristics of the households and households constitute the top part of the table, with an average number of household members being 2.49. Similarly, we provide the average (Mean), standard deviation (S.D.), minimum (Min) and maximum (Max) of the variables. The annual household income variable comprises categories 1 to 10, such that: 1: less than £5,000; 2: £5,000 - £9,999; 3: £10,000 - £14,999; 4: £15,000 - £19,999; 5: £20,000 - £24,999; 6: £25,000 - £34,999; 7: £35,000 - £49,999; 8: £50,000 - £74,999; 9: £75,000 - £99,999; 10: £100,000 or more. The dependent variables of interest is the residential location choice of Borough. At each of these aggregation levels for residential location choice, we have alternative specific attributes: Natural-log of estimated price of a 2-bedroom, Natural-log of population density (per squared-km), crime incidents density (per squared-km), and PTAL average values.

³After cleaning for missing information and data with fewer than 10 observations for any specific borough, we were left with 31 boroughs in London, out of a total of usual 33 boroughs which form the London metropolitan region.

TABLE 3.1 Description of Sample (n=4,866)

Socio-economics and demographic variables distribution				
Variable definition	Mean	S.D.	Min	Max
Number of household (HH) members	2.49	1.36	1.00	9.00
How long has the household lived here (years)	12.93	13.68	1.00	89.00
Annual household income* (with imputed values)	6.14	2.68	1.00	10.00
Number of vehicles in total household owns or has access to	0.92	0.92	0.00	10.00
Fraction of males in HH	0.48	0.32	0.00	1.00
Fraction of driving license holders in HH	0.61	0.38	0.00	1.00
Fraction of free-travel-pass holders in HH	0.27	0.41	0.00	1.00
Fraction of reduced-travel-pass holders in HH	0.13	0.27	0.00	1.00
Fraction of oyster-ticket holders in HH	0.43	0.42	0.00	1.00
Dummy for presence of student in HH	0.13	0.33	0.00	1.00
Fraction of employed HH members	0.53	0.39	0.00	1.00
Total number of trips	5.63	4.09	1.00	37.00
Dependent variables characteristics distribution				
Natural-log of Borough estimated price of a 2-bedroom	12.81	0.42	12.21	13.88
Natural-log of Borough population density (per squared-km)	15.90	15.16	12.74	16.56
Borough crime incidents density (per squared-km)	50.39	27.85	10.51	183.67
Borough PTAL average value	12.73	12.19	2.75	90.80

Notes: *The household income categories are defined as follows: 1: less than £5,000; 2: £5,000 - £9,999; 3: £10,000 - £14,999; 4: £15,000 - £19,999; 5: £20,000 - £24,999; 6: £25,000 - £34,999; 7: £35,000 - £49,999; 8: £50,000 - £74,999; 9: £75,000 - £99,999; 10: £100,000 or more

Table 3.2 provides the frequency distribution of the 31 boroughs of London as available in the data. These boroughs constitute the upper-level aggregate choice decision.

TABLE 3.2 Borough choice frequency distribution (n=4,866)

Borough Name	Freq.	%	Borough Name	Freq.	%
Barking and Dagenham	26	0.53	Islington	291	5.98
Barnet	163	3.35	Kensington and Chelsea	285	5.86
Bexley	198	4.07	Kingston upon Thames	310	6.37
Brent	116	2.38	Lambeth	236	4.85
Camden	108	2.22	Lewisham	120	2.47
City of London	53	1.09	Merton	80	1.64
Croydon	216	4.44	Newham	154	3.16
Ealing	141	2.90	Redbridge	166	3.41
Enfield	168	3.45	Richmond upon Thames	152	3.12
Greenwich	48	0.99	Southwark	139	2.86
Hackney	70	1.44	Sutton	176	3.62
Hammersmith and Fulham	49	1.01	Tower Hamlets	141	2.90
Harrow	163	3.35	Waltham Forest	165	3.39
Havering	312	6.41	Wandsworth	141	2.90
Hillingdon	178	3.66	Westminster	142	2.92
Hounslow	159	3.27	Total	4866	100.00

3.3 METHODOLOGICAL FRAMEWORK

First we set up the choice modelling problem as a classification problem. In order to predict the probabilistic consideration set, we employ probabilistic decision trees as described in the next subsection. This is followed by introduction to decision tree methodology, leading to Random Forests structure utilised to develop probabilistic outcomes. This non-compensatory method inspired by decision trees allow us to winnow the universal choice set to obtain a consideration set. In order to estimate the final observed choice decision, we employ a discrete choice model (Multinomial Logit and Mixed Logit). This sequential framework allows us to blend a non-compensatory method with a compensatory model to get the best of both worlds.

3.3.1 Problem set up

Let the dependent (endogenous) variable of interest be Y and the explanatory (exogenous) variables be denoted by X . We created N number of $X \times Y$ labelled examples as training data with k-fold cross-validation, where N is the number of observations in the data. Since we are interested in probabilistic predictions, our goal is to not just predict a label $y \in Y$, but to output the distribution $p(y|x)$, i.e. the conditional distribution of the label $\mathbf{y}$ given the features $\mathbf{x}$. We focus explicitly on *multi-class classification* which is another term for discrete choice model with more than two alternatives, where $Y \in 1, 2, \dots, K$, where K denotes the number of *classes in the classification problem* or *alternatives in choice decision*.

3.3.2 Decision Trees

A **decision tree** can be assumed to be hierarchical, axis-aligned, binary partitioning of X and a rule for predicting the label of test points given training data. The structure of the decision tree is a finite, rooted, strictly binary tree T , that is, a finite set of **nodes** such that: 1. every node j has exactly one **parent** node, except for a distinguished **root** node ϵ which has no parent, and 2) every node j is the parent of exactly zero or two **children** nodes, called the left child $left(j)$ and the right child $right(j)$. Denote the leaves of T (those nodes without children) by $leaves(T)$. Each node of the tree $j \in T$ is associated with a **block** $B_j \in \mathbb{R}^D$, where $D \in (X, Y)$. At the root, we have $B_\epsilon = \mathbb{R}^D$ while each **internal** node $j \in T$ not including $leaves(T)$ with two children represents a **split** of its parent's block into two halves, with $\delta_j \in \{1, \dots, D\}$ denoting the dimension of the split, and ξ_j denoting the location of the split. In particular,

$$B_{left\{j\}}\{x \in B_j : x_{\delta_j} \leq \xi_j\} \quad (3.1)$$

$$B_{right\{j\}}\{x \in B_j : x_{\delta_j} > \xi_j\} \quad (3.2)$$

We call the tuple $\tau = (T, \delta, \xi)$ a decision tree. Note that the blocks associated with the leaves so the tree form a partition of $\mathbb{R}^D$. We may write $B_j = (l_{j1}, u_{j1}] \times \dots \times (l_{jD}, u_{jD}]$, where l_{jd} and u_{jd} denote the *lower* and *upper* bounds, respectively, of the rectangular block B_j along dimension d . Let $l_j = \{l_{j1}, l_{j2}, \dots, l_{jD}\}$ and $u_j = \{u_{j1}, u_{j2}, \dots, u_{jD}\}$. Let $parent(j)$ denote the parent of node j . Let N_j denote the indices of training data points at node j , i.e., $N_j = \{n \in \{1, \dots, N\} : \mathbf{x}_n \in B_j\}$. Note that both B_j and N_j depend on the tuple τ , although we have chosen to elide this dependence for notation simplicity. Let $F_{N_j} = \{\mathbf{X}_{N_j}, Y_{N_j}\}$ denote the features and labels of training data points at node j . Let l_{jd}^x and u_{jd}^x denote the *lower* and *upper* bounds of training data points (hence the superscript x) respectively in node j along dimension d . Additionally, let $e_{jd}^x = (l_{jd}^x, u_{jd}^x]$ denote the extent of the training data in node j along dimension d .

For brevity, we'll also use the following shortcut notation to label the nodes of the decision tree: label the root node as the empty string ϵ and label $left(j) = j0$ and $right(j) = j1$. If each parent-child path is labelled 0 or 1 depending on the outcome of the binary decision, this labelling scheme ensures that the label of each node is the concatenation of the labels along the path from the root till that node. Once we have a decision tree structure, we also need a rule for predicting the label of a test points given training data. To this end, we will associate each leaf node j with a parameter θ_j that parametrizes the conditional distribution $p(y|x \in B_j)$. For instance, θ_j would parametrize the K -dimensional discrete distribution for classification problem (or discrete choice problem).

Learning a decision tree from training data involves two steps namely, learning the tree structure T and estimating the leaf node parameters θ . Popular decision tree induction algorithms include CART (Breiman et al., 1984) and C4.5 (Quinlan, 1993). While it is possible to learn a deep decision tree until there is a unique data point at each leaf node, it is common to limit the complexity of the decision tree by specifying a hyper-parameter that decides when to stop splitting a node. The most popular strategy is to require a minimum

number of samples (`min_samples_split`) at a node before it can be split. (A variant of this strategy is to require that a split leads to a minimum number of samples `min_samples_leaf` at each leaf node.) Alternative strategies include not splitting a node if all the class labels are identical (for classification problems) or limiting the maximum depth of the tree; however specifying maximum depth is relatively harder to specify in a dataset-agnostic fashion (since deeper and/or unbalanced trees might be preferable for some datasets). Due to its simplicity and robustness, we prefer `min_samples_split`.

We describe a typical decision tree induction algorithm in Algorithms 2.1 and 2.2. The procedure starts with the root node ϵ and recurses down the tree. At node j , $CandidateSplits_j$ denotes the set of candidate pair of valid split dimensions and locations, where a valid split is one where both children are non-empty. In practice, the set of valid split candidates is obtained by sorting the training data independently along each dimension; since the training data takes on only along a finite number of unique values, it is sufficient to consider a single split location (usually the midpoint) for each of these intervals (as any split location along this midpoint has the same accuracy on the training data). We greedily choose the best split dimension and split location from $CandidateSplits_j$ by optimizing an appropriate criterion, e.g. *information gain* or *Gini index* for classification and reduction in *MSE* for regression.

Algorithm 1: BuildDecisionTree ($F_{1:n}, \text{min_samples_split}$)

1. Initialize empty tree $T = \emptyset$, $\text{leaves}(T) = \emptyset$, $\delta = \emptyset, \xi = \emptyset$
 2. Set $N_\epsilon = \{1, 2, \dots, n\} \implies \text{entire dataset is used at root node}$
 3. $\text{ProcessBlock}(\epsilon, F_{N_\epsilon}, \text{min_samples_split}) \implies \text{Algorithm 2}$
-

Algorithm 2: ProcessBlock ($j, F_{N_\epsilon}, \text{min_samples_split}$)

1. Add j to T
2. if $|N_j| \geq \text{min_samples_split}$ **then**
3. Set $CandidateSplits_j$ to the set of all valid pairs of split dimensions and locations
4. Choose best split dimension δ_j and split location ξ_j amongst $CandidateSplits_j$ by optimizing appropriate criterion $\implies \text{greedy optimization}$
5. Set $N_{\text{left}(j)} = \{n \in N_j : \mathbf{X}_{n,\delta_j} \leq \xi_j\}$ and $N_{\text{right}(j)} = \{n \in N_j : \mathbf{X}_{n,\delta_j} > \xi_j\}$
6. $\text{ProcessBlock}(\text{left}(j), F_{N_{\text{left}(j)}}, \text{min_samples_split})$
7. $\text{ProcessBlock}(\text{right}(j), F_{N_{\text{right}(j)}}, \text{min_samples_split})$
8. **else** $\implies j \text{ is a leaf node}$

9. Add j to $\text{leaves}(\mathbf{T})$
10. Estimate $\boldsymbol{\theta}_j$ using Y_{N_j}

For leaf nodes, we estimate the parameters $\boldsymbol{\theta}_j$ using F_{N_j} . In the simplest case, $\boldsymbol{\theta}_j$ is estimated just using Y_{N_j} , independent of $\mathbf{X}_{N_j}$, however, we restrict our attention to the former since it is computationally fast. For classification (or discrete choice) problems, let c_{jk} denote the number of data points in node j with label k , i.e., $c_{jk} = \sum_{n \in N_j} 1[y_n = k]$; in this case $\boldsymbol{\theta}_j$ is estimated as:

$$\boldsymbol{\theta}_j = \frac{c_{jk} + \alpha}{|N_j| + K\alpha} \quad (3.3)$$

where we add a small constant α to smooth the empirical histogram of labels in node j and $|N_j|$ (the size of the set N_j) denotes the number of data points in node j .

3.3.3 Prediction with a decision tree

Recall that $\text{leaf}(x)$ denotes the unique leaf node $j \in \text{leaves}(\mathbf{T})$ such that $x \in B_j$. Prediction from a tree involves two steps: (i) traversing the decision tree starting from the root node to identify $\text{leaf}(x)$ and (ii) returning (a function of) the leaf node parameter $\boldsymbol{\theta}_{\text{leaf}(x)}$. For classification, one can return either the probabilistic prediction $\boldsymbol{\theta}_{\text{leaf}(x)}$ (which we are interested in) or most likely (with highest probability) class label (predicted alternative) $\arg \max_k \boldsymbol{\theta}_{\text{leaf}(x),k}$

3.3.4 Ensembles of decision trees: Random Forests

In *ensemble learning*, many ‘weak’ predictors are combined to obtain a ‘powerful’ predictor (Dietterich, 2000) that is more accurate than the household predictors. In the simplest case, the predictions from the ensemble are just a weighted additive combination of the predictions from the household predictors. While ensemble learning can be combined with any learning algorithm, ensembles of decision trees are very popular since decision trees are computationally fast to train and test. Ensembles of decision trees often achieve state-of-the-art performance in many supervised learning problems (Caruana and Niculescu-Mizil, 2006). Let $\{\tau_m, \boldsymbol{\theta}_m\}_{m=1}^M$ denote an ensemble of trees, where M denotes the number of trees in the ensemble. Let $g(\mathbf{x}; \tau_m, \boldsymbol{\theta}_m)$ denote the prediction from the m^{th} decision tree for a test data point $\mathbf{x}$. (We allow flexible notation for prediction to either be a point estimate or a probability distribution or

density.) The prediction from an ensemble can be written as:

$$g(\mathbf{x}; \{\tau_m, \boldsymbol{\theta}_m\}_{m=1}^M) = \sum_{m=1}^M \omega_m \cdot g(\mathbf{x}; \tau_m, \boldsymbol{\theta}_m) \quad (3.4)$$

where $\sum_{m=1}^M \omega_m = 1$. Ensemble of trees can be broadly classified into two families: *additive/boosted decision trees*, wherein each tree fits the residual not explained by the remainder of the trees, and *random forests*, wherein randomized independent decision trees are grown independently and predictions are averaged to reduce variance. For the purpose of this analysis, we will only focus on *Random Forests* (**RF**).

The classic method for decision tree induction involves selecting the best split dimension and location at each node based on a suitable quality criterion, such as information gain, in a greedy manner. Random forests differ from decision trees in that they use household trees that are randomized in order to decorrelate their predictions. This randomization can be achieved by:

- using bootstrap aggregation, also known as bagging, where each decision tree is trained on a slightly different training data set, or
- by randomly sub-sampling the set of candidate splits within each node.

The prediction from a random forest is usually an (unweighted) average of the predictions of household trees:

$$g(\mathbf{x}; \{\tau_m, \boldsymbol{\theta}_m\}_{m=1}^M) = \sum_{m=1}^M \frac{1}{M} \cdot g(\mathbf{x}; \tau_m, \boldsymbol{\theta}_m) \quad (3.5)$$

When using decision trees for classification, majority voting can be used if the trees output discrete class labels instead of probabilities. The weights used in voting can be optimized, such as through stacking. Random forests offer an advantage over decision trees by reducing variance through randomizing the household trees to decorrelate their predictions, even if this increases the bias slightly. Ensembles, such as random forests, can outperform single classifiers due to their ability to reduce the risk of choosing the wrong hypothesis, approximate the true function better in cases of local optima, and increase the effective hypothesis space through representation of weighted combinations of trees. Breiman-RF and Extremely randomized trees (ERT) are two popular random forest variants. Random forests are easy to implement

and parallelize, and have consistently ranked among the top-performing algorithms in machine learning. There are also extensions of random forests for density estimation, manifold learning, and semi-supervised learning (Criminisi et al., 2012).

3.3.5 Discrete Choice model

As mentioned before, the consideration probabilities across all alternatives obtained from the non-compensatory step 1 are then utilized to form consideration sets, conditional on which multinomial choice models are estimated. We will briefly review the Mixed Logit model (**MxL**).

Mixed Logit (MxL) model

We extend the MNL model by allowing for heterogeneity in preferences across households. Specifically, we assume that the coefficients of the explanatory variables are randomly drawn from a distribution, allowing for the estimation of household-specific coefficients. This is achieved by specifying a mixed logit model, which can be written as follows:

$$V_{ij} = \beta'_{1ij}x1_{ij} + \beta'_{2i}x2_i + \epsilon_{ij}$$

where V_{ij} is the utility of alternative j for household i , β_{1ij} is a vector of alternative-specific attributes' coefficients for household i and alternative j (which is allowed to be generalized as per equation below), $x1_{ij}$ is a vector of alternative-specific attributes for alternative j , and ϵ_{ij} is the error term. β_{2i} is a vector of socio-economic and demographic attributes' coefficients for household i (assumed homogenous across population), $x2_{ij}$ is a vector of these attributes for alternative j . The coefficients β_{1ij} are assumed to follow a normal distribution with mean β and variance-covariance matrix Σ , so that:

$$\beta_{1ij} \sim N(\beta, \Sigma)$$

To capture the effects of alternative-specific attributes, such as price, PTAL, population density, and crime density, we include these variables in the vector of explanatory variables $x1_{ij}$. The specific form of the vector $x1_{ij}$ will depend on the chosen set of attributes and their functional form. For example, if we have price (P_{ij}), PTAL ($PTAL_{ij}$), population density (D_{ij}), and

crime density (C_{ij}) as attributes, the vector $x1_{ij}$ can be defined as:

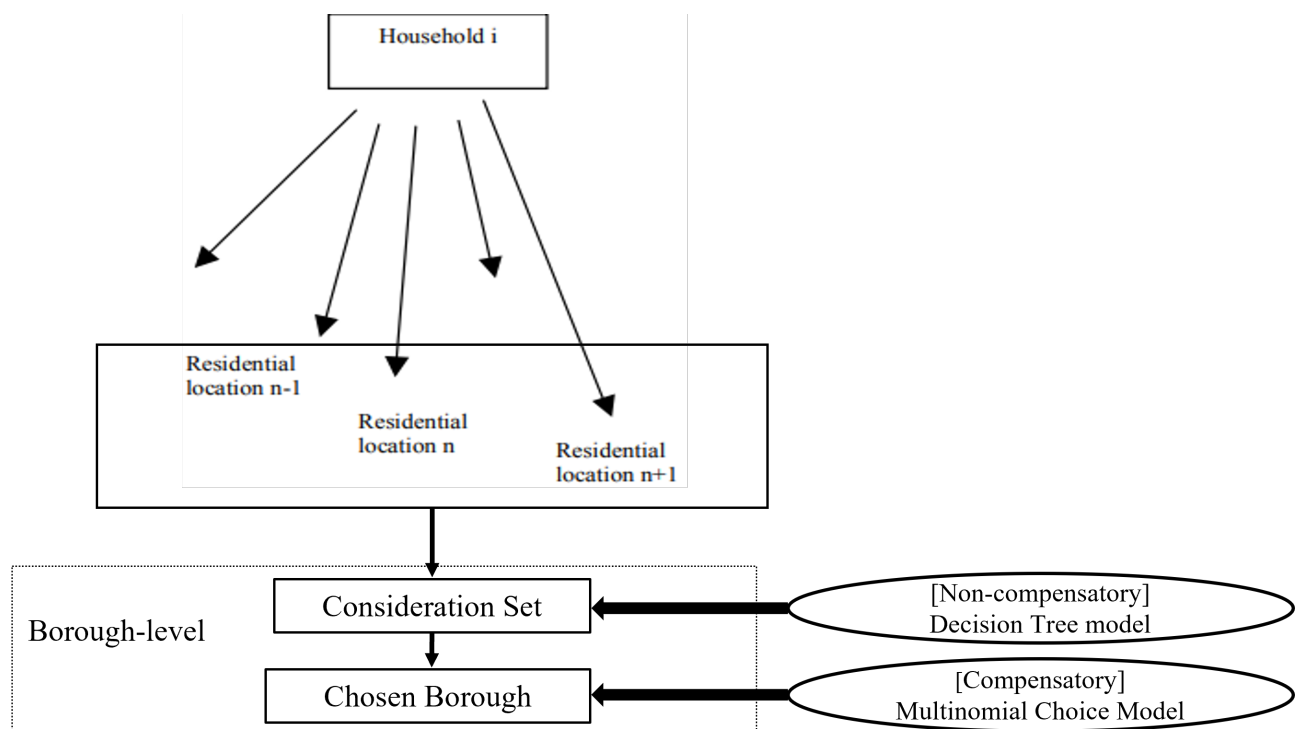
$$x1_{ij} = [P_{ij}, PTAL_{ij}, D_{ij}, C_{ij}]$$

Furthermore, to incorporate the correlations between alternatives belonging to different transport zones (1 to 6), we introduce zone correlation in the error term ϵ_{ij} . We can include six zone correlation terms, denoted as *zone_corr1* to *zone_corr6*, to capture the correlation between alternatives within each zone. The inclusion of these zone correlation terms allows us to estimate the correlations among alternatives within their respective zones. We estimate the parameters of the mixed logit model using maximum likelihood estimation, which involves integrating over the distribution of the coefficients for each household. We use the software package Apollo in R for this purpose. The mixed logit model allows us to relax the assumption of homogeneous preferences, and to estimate the variance of the coefficients across households. This is particularly useful in situations where we expect that preferences may vary systematically across households, or where we have limited information on the distribution of preferences in the population.

Finally, to get a sense of Greater London are with its 33 boroughs (out of which only 31 are in our empirical analysis), we have a look at Figure 3.1, providing a visual identity of London's boroughs within the greater London boundary. This is followed by Figure 3.2 providing a model implementation framework of this chapter. To summarise, we use a decision tree based model to winnow the universal choice set (London boroughs) to a smaller consideration set, conditional on which a multinomial choice model is estimated. The universal choice set consists of 31 boroughs in London, upon which the non-compensatory model is estimated. Once we get a consideration set of boroughs, the multinomial choice model is estimated to estimate the exact choice.



FIGURE 3.1 London Boroughs map

**FIGURE 3.2** Model implementation

3.4 ESTIMATION RESULTS

3.4.1 Description of model implementation and specification search

Model framework was estimated at the borough level, with consideration set formation and choice models estimated sequentially. The data set consisted of 4,866 households residing in the London metropolitan region. The dependent variable of interest in this analysis is the residential location choice at the aggregate level of the Borough, with 31 alternatives. We utilise the independent characteristics of the households and households from the top part of the Table 3.1 as independent variables for the consideration set formation; while, the alternative specific attributes are utilised as independent variables from the bottom part of the Table 3.1 as independent variables for the choice model. While the section 3.3 "methodological framework" presents the mathematical approach of the decision trees based consideration set formation and the mixed logit discrete choice models, the practical application of the same follows as below:

1. Step 1: At the first stage of the model, we employ the household socioeconomic and demographic variables, vehicle data, and trip data from the London Travel Demand Survey. These variables build a decision tree model that predicts the likelihood of consideration of different boroughs based on the values of these variables. We also test alternative-specific attributes to improve the accuracy of the model. This model is developed in R software using the libraries *RandomForest* and *Ranger* and comparatively evaluated for their complexity and checked for overfitting. *RandomForest* outcome was severely overfitted with an almost perfect probability of consideration assigned to the observed chosen alternative, therefore, the *Ranger* outcome was preferred for a more realistic distribution of probabilities to the alternatives. The model estimation is repeated to obtain the highest *test-train prediction* accuracy.
2. Step 2: The probabilistic consideration set is obtained in the form of probabilities for each of the alternatives for all observations in the dataset. For each observation, the sum of these probabilities is equal to one. However, for most of the observations, the top three alternatives' probabilities were much higher than the sum of the remaining 28 (out of total 31) and therefore, the consideration set size was set to three. The chosen alternatives are assigned a value of one and the others zero for each observation, thereby converting the consideration set into an *availability of alternatives* structure to be utilised

in the subsequent choice model. The model outcome is allowed to be heterogenous, hence, the consideration set was different for each observation in the dataset. After this step, this winnowed consideration set availability information is added to the data.

3. Step 3: At this stage of the model, we would use the availability of alternatives from the consideration set to estimate the preferences and choice behaviour of households at the borough level. Firstly, the preferred variables for estimation are the alternative-specific attributes (for empirical inferential analysis of estimating a location choice model), but the household socioeconomic and demographic information is also tested in model specifications to tease out any remainder variation in the dataset. As expected, much of socioeconomic and demographic information is statistically insignificant. The few relevant household attributes are incorporated in the model specification as interaction terms with alternative-specific attributes to infer their mediating impact on the outcome. This model is developed in R using *Apollo* library (Apollo, 2018).
4. Step 4: The mixed logit model fit is evaluated using log-likelihood and AIC/BIC values and parameter interpretability.

3.4.2 Consideration set formation results

The model to calculate consideration probabilities is employed on the full data set. The model provides a measure of feature importance, indicating the relative importance of different exogenous attributes in influencing residential location choices. We show a ranking of the most influential features in determining consideration probabilities in Figure 3.3 (followed by Table 3.3 matching variable names to their description). As discussed later, all exogenous variables were allowed in both the consideration model and the choice model (discussed later), but the variables of higher *importance* as shown by the plot are not statistically significant in the choice model. Therefore, it can be assumed that the majority of the effect of these variables is captured in the consideration set formation model.

While the random forest framework doesn't allow us to infer the parameter effect of variables, it does offer us an insight about which parameters do play a role. For example, the households structures 1 to 5 may not be the most significant here, but their presence resonates with earlier results by Bhat (2015) who found that the composition of a household significantly

affects their residential choices. They found that single individuals tend to prefer high-density urban areas, while nuclear families with children have a strong preference for single-family homes in low-density urban areas due to perceived benefits for children's security and well-being. Other types of households, like single parents or those with multiple adults, show a preference for rented single-story apartments in high-density urban areas. Their results also linked household income to greater commute distances, indicating that higher income households may often prefer locations far from the city-centre with longer commute times. This is often driven by the desire to access better job opportunities and markets, while maintaining a location preference which offers bigger household units. On the other hand, those with lower wages tend to prefer locations close to high-density employment destinations, perhaps to commute less or availability of affordable smaller units closer to city-centre.

Zolfaghari et al., (2013) employed a hazard-based consideration set formation model, and suggest that larger households seek less expensive areas to accommodate their space needs within a limited budget, households with children prefer safer and higher-quality neighbourhoods, households with higher income and number of vehicles correlate positively with living in more expensive suburban areas and longer commutes, presence of workers tend to prioritize proximity to their workplaces.

Similarly, the probability estimates generated by the consideration model are utilised to create observation-specific consideration sets. The model output sum of probabilities for the top 3 alternatives turns out to be greater than 0.5 with the remaining 28 alternatives with an average consideration probability of approximately 0.015, which means that for any specific observation the top three consideration probabilities acquire enough of the probability space to let go of the remaining 28 out of 31 alternatives. However, this resulting consideration set of 3 alternatives varies for each observation.

3.4.3 Mixed logit model results

On the consideration set we obtained from the decision tree formulation, we estimated a Mixed Logit (MxL) choice model. In order to make the estimation faster, we used traditional multinomial logit (MNL) model results as starting values for coefficients (as is common in literature involving estimation of complex choice models, Hensher, 1986, Gu et al., 2013). For comparative analysis, MNL was estimated based on a deterministic consideration-set

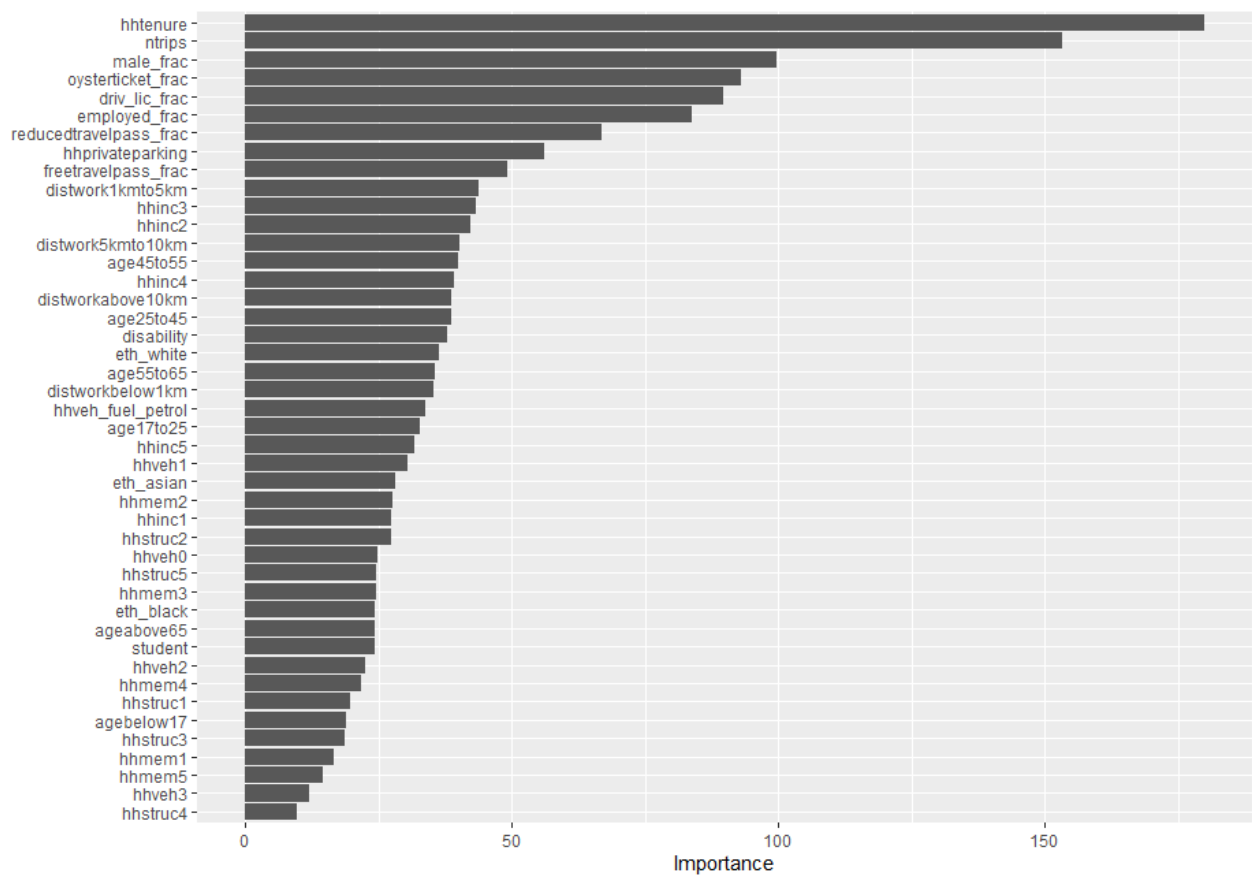


FIGURE 3.3 Variable Importance Plot for consideration of alternatives

TABLE 3.3 Matching variable names to their description

Variable name	Description
hhtenure	How long has the household lived here
ntrips	Total number of trips
male_frac	Fraction of males in HH
employed_frac	Fraction of employed HH members
driv_lic_frac	Fraction of driving licence holders in HH
freetravelpass_frac	Fraction of free travel pass holders in HH
reducedtravelpass_frac	Fraction of reduced travel pass holders in HH
oysterticket_frac	Fraction of oyster ticket holders in HH
hhprivateparking	Dummy for HH access to private, off-street parking?
distworkbelow1km	Presence of HH members with distance to work ≥ 0 and < 1 km
distwork1kmt05km	Presence of HH members with distance to work ≥ 1 and < 5 km
distwork5kmt010km	Presence of HH members with distance to work ≥ 5 and < 10 km
distworkabove10km	Presence of HH members with distance to work ≥ 10 km
hhinc1	Annual household income (with imputed values) less than £9999
hhinc2	Annual household income (with imputed values) £10,000 - £24,999
hhinc3	Annual household income (with imputed values) £25,000 - £49,999
hhinc4	Annual household income (with imputed values) £50,000 - £99,999
hhinc5	Annual household income (with imputed values) £100,000 or more
agebelow17	Presence of HH members in age ≥ 0 and < 17
age17to25	Presence of HH members in age ≥ 17 and < 25
age25to45	Presence of HH members in age ≥ 25 and < 45
age45to55	Presence of HH members in age ≥ 45 and < 55
age55to65	Presence of HH members in age ≥ 55 and < 65
ageabove65	Presence of HH members in age ≥ 65
disability	Dummy for presence of disability in HH
eth_white	Dummy for Ethnicity white, compared to 'other' category
eth_asian	Dummy for Ethnicity asian, compared to 'other' category
eth_black	Dummy for Ethnicity black, compared to 'other' category
hhveh_fuel_petrol	Dummy for type of fuel - petrol, compared to 'diesel and other'
hhveh0	Number of vehicles in total household owns equals zero
hhveh1	Number of vehicles in total household owns equals one
hhveh2	Number of vehicles in total household owns equals two
hhveh3	Number of vehicles in total household owns equals three or more
hhmem1	Number of household members equals one
hhmem2	Number of household members equals two
hhmem3	Number of household members equals three
hhmem4	Number of household members equals four
hhmem5	Number of household members equals five or more
hhstruc1	One-person household: A household consisting of one adult with no dependent children.
hhstruc2	Couple household: A household consisting of two adults in a marriage or civil partnership with no dependent children.
hhstruc3	Single-parent household: A household consisting of one adult with one or more dependent children.
hhstruc4	Multi-generational household: A household consisting of three or more generations living together.
hhstruc5	Other multi-person household: A household consisting of two or more adults and/or dependent children who are not in a couple or single-parent household.
student	Dummy for presence of student in HH

generation method. This method randomly picks three alternatives from the respective transport zone (including the chosen alternative) to homogeneously form a 3-alternative consideration-set per observation.

The Mixed Logit model was developed in steps where first all parameters are assumed to be homogeneous (non-random), then coefficients for price, PTAL, population density and crime density variables are allowed to be random parameters. Finally, we tried two different correlation-structures to model dependency between alternative. First, we attempted to estimate correlations between all pairs of alternatives, resulting in a total of 465 correlation coefficients (31 choose 2). This approach allows for maximum flexibility in capturing the interdependence between alternatives, but dramatically increases the complexity of the model and the number of parameters to be estimated. Second, we attempted a more nuanced approach of looking at correlations between alternatives, where the alternatives belonging to a particular Transport Fare Zone (1 to 6) were allowed a specific fixed-correlation with one-another, while being independent of all other alternatives. This reduced the number of correlations to be modelled, while offering insight into the dependency between alternatives. The model results shown below reflect this second approach to modelling dependency between alternatives. The model performance is evaluated using Bayesian Information Criterion (BIC). For a given empirical model, $BIC = K \ln(Q) - 2 \ln(L)$ where K is the number of parameters, Q is the number of observations and $\ln(L)$ is the log-likelihood value at convergence. The model with the lowest value of BIC is preferred. The final log-likelihood (AIC/BIC) value of the MxL model is -5242.19 (10522.38/10645.62), while the corresponding values for the traditional MNL model are -5240.49(10496.98/10548.9). From a strict model improvement fit statistics, the traditional MNL performs better, however, as detailed below the in the table, the coefficients are counter-intuitive; providing evidence in favour of non-compensatory consideration of alternatives by decision makers.

TABLE 3.4 Residential location choice model results

Variable definitions	Mixed Logit		Multinomial Logit	
	Estimate	t-stat	Estimate	t-stat
<i>(Alternative-specific attributes)</i>				
Average estimated price of a 2-bedroom	-0.2585	-2.55	0.5536	7.21
Borough PTAL	0.3764	5.22	-0.3503	-6.63
Borough population density	-0.1260	-1.11	-0.0884	-1.24
Borough crime density	-0.1010	-1.06	0.1598	2.05
<i>(Interaction: Average estimated price of a 2-bedroom *)</i>				
<i>(Base: Household income < £50,000)</i>				
Household income > £50,000 and < £100,000	-0.1918	-1.44	---	---
Household income > £100,000	-0.3474	-2.02	---	---
<i>(Interaction: Average estimated price of a 2-bedroom *)</i>				
<i>(Base: Number of household members = 1)</i>				
Number of household members = 2	-0.0788	-1.67	---	---
Number of household members = 3	---	---	---	---
Number of household members = 4	-0.1932	-1.96	---	---
Number of household members = 5 or more	---	---	---	---
<i>(Interaction: Borough population density (per squared-km) *)</i>				
<i>(Base: Household income < £50,000)</i>				
Household income > £50,000 and < £100,000	0.1774	1.61	---	---
Household income > £100,000	0.3295	2.34	---	---
<i>(Interaction: Borough population density (per squared-km) *)</i>				
<i>(Base: Household vehicle = 0)</i>				
Household vehicle > 0	-0.1440	-3.09	0.0715	3.36
<i>(Interaction: Borough crime density (per squared-km) *)</i>				
<i>(Base: Ethnicity = white)</i>				
Ethnicity = non-white	0.3799	4.41	-0.2891	-4.85
Household private parking	---	---	1.2966	2.97
Ownership of Oyster ticket	---	---	0.97	2.60
Correlation Zone 1	---	---	---	---
Correlation Zone 2	---	---	---	---
Correlation Zone 3	-1.32884	-3.36085	---	---
Correlation Zone 4	2.04470	5.98111	---	---
Correlation Zone 5	1.86183	4.43433	---	---
Correlation Zone 6	---	---	---	---

Notes: "—" signifies statistically not significant

The results confirm our hypothesis that decision makers first use a non-compensatory process

to reduce the number of alternatives from the universal choice set, followed by a compensatory process to obtain a final choice. Moreover, the DT-MxL allows us to capture behaviourally richer contributing factors. Table 4 present estimation results depicting the influence of the exogenous variables on the residential location choice. This consideration set varies for each household and therefore has been accommodated in the form on *availability* of the alternative in the estimation procedure. Given the empirical context of residential location choice modelling, the alternative-specific constants have been suppressed. Also, the household socio-economic and demographic factors have been interacted with alternative-specific attributes to allow for behavioural interpretation.

Our estimated model consists of alternative-specific attributes, as well as their interaction terms with socio-economic and demographic variables. Therefore, their interpretation may not be straightforward. Borough average estimated price of a 2-bedroom apartment variable has an independent coefficient as well as interaction with levels of household income and an interaction with levels of household members. Based on coefficients from Table 1, we find that a higher average estimated price of a 2-bedroom apartment negatively impacts (effect size = -0.2585) the probability of choice of the borough, for single-person households with income less than £50,000. Also, for single-person households, as we move on to a higher income group between £50,000 and £100,000, we have an additional negative impact on probability of choice of borough (effect size = $(-0.2585) + (-0.1918) = -0.4503$), and with the highest income group of income greater than £100,000, we have an additional negative impact on probability of choice of borough (effect size = $(-0.2585) + (-0.3474) = -0.6059$). Similarly, we can calculate the effect sizes for added number of members to the household as household income increases, by further adding the additional effect of price-household size interaction term. Our results resonate with earlier findings (Bhat, 2015) that conclude that as the number of children or adults increases, there's a greater preference for larger living spaces, more bedrooms, and more bathrooms, leading to a preference for suburban locations. Their results show a recursive effect of housing cost on residential location (and tenure decisions), and presence of children seem to play a more significant role in decisions about lot size and the number of bathrooms, thereby motivating the preference towards locations that offer bigger units, usually in a non-dense location. Comparatively, the MNL model incorrectly estimates the effect of average estimated price of a 2-bedroom apartment on choice of borough as positive, with interaction terms as

statistically insignificant.

Borough Public Transport Accessibility Level (PTAL) average value has a positive impact on the choice of borough, with much higher borough accessibility levels making the borough an attractive choice for residence. Comparatively, the MNL model incorrectly estimates the effect as negative. Unexpectedly, in their analysis, accessibility to employment exhibited a negative influence, possibly due to the desirability of remote neighbourhoods, which perhaps could be attributed to the structure of consideration set formation in the model design.

Borough population density (per squared-km) has an independent effect coefficient and interaction terms with household income and presence of household vehicle. Based on coefficients from Table 1, for households with zero vehicle ownership and income less than £50,000, higher population density negatively impacts the choice of residential location (effect size = -0.1260). Also, for households with one or more vehicles and income less than £50,000, higher population density negatively impacts the choice of residential location (effect size = $(-0.1260) + (-0.1440) = -0.2700$). However, for households with zero vehicle ownership and income group between £50,000 and £100,000, higher population density positively impacts the choice of residential location (effect size = $(-0.1260) + 0.1774 = 0.0514$). Similarly, for households with zero vehicle ownership and income group with income greater than £100,000, higher population density positively impacts the choice of residential location (effect size = $(-0.1260) + 0.3295 = 0.2035$). For households with one or more vehicles and income group between £50,000 and £100,000, higher population density negatively impacts (marginally different from zero effect size) the choice of residential location (effect size = $(-0.1260) + (-0.1440) + 0.1774 = -0.0926$). For households with one or more vehicles and income group with income greater than £100,000, higher population density positively impacts (marginally different from zero effect size) the choice of residential location (effect size = $(-0.1260) + (-0.1440) + 0.3295 = 0.0595$). While past results (Zolfaghari et al., 2013) revealed that zonal area and population density have a positive impact, and that household size, housing price, and commute time have negative impact, the analysis is not as mediated by the role played by income-segmentation and mediated by other households socioeconomic variables. Comparatively, the MNL model estimates the effect of Borough population density (per squared-km) on choice of borough as -0.0884, with no interaction terms as statistically significant.

Borough crime density (per squared-km) has an independent effect and interaction term with ethnicity dummy (white vs. non-white). For white households, the effect of crime density (per squared-km) on choice of residential location is negative (effect size = -0.1010), while for non-white households, the effect of crime-density on choice of residential location is positive (effect size = $(-0.1010) + 0.3799 = 0.2789$). Comparatively, the MNL model estimates the effect of Borough crime density (per squared-km) on choice of borough as positive (0.1598), with no accounting of heterogeneous mediating-effect of race on crime-residence location relationship.

Finally, the effect of household private parking availability and household ownership of Oyster pass is insignificant in our framework, but captured by MNL model. We also observe the Zone correlation coefficients for Zone 1 to 6. While the coefficients for zones 1, 2 and 6 are statistically insignificant, the one for zone 3 is negative and those for zones 4 and 5 are positive. This gives us an understanding of the impact of unobserved factors in affecting the choice probability for an alternative in a specific zone, on other alternatives in the same zone. For example, the Zone correlation 3 being negative, signifies that unobserved factors which may lead to a higher probability for alternative 22 (Lambeth), will result in a lower probability for alternatives 23 (Lewisham), 25 (Newham) and 32 (Wandsworth).

3.4.4 Comparative evaluation of discrete choice predictions

The models estimated were now used for prediction exercises on the sample using the parameter estimates and the independent variables. We provide comparative evaluation of the proposed model performance with MNL in Figure 3.4, where the predicted number of observations for each alternative are compared. Firstly, the observed frequency of chosen alternatives is mapped in *blue* in the histogram. Then, the prediction from the base MNL model is added in *orange*, which is the outcome of the estimated model using estimated parameters and dataset variables to calculate choice probabilities of the alternatives, and the chosen alternative for each observation. Finally, the estimated parameters from the Mixed logit model (second step of the two-step model) are used to predict probabilities and calculation of the chosen alternative, and this prediction is added in *grey* to the frequency histogram. It is important to note here that the purpose of this prediction using only the second step of the model is to show that a model estimated using a limited heterogeneous consideration set is better able to capture the heterogeneity in data and performs better than

MNL model, which may often lead to homogenous (average) prediction of choice probabilities.

While this is not an accurate measure of correct prediction, it does give us an understanding of the probabilistic distribution of the alternatives in the probability space being captured by the proposed model, as compared to the MNL which averages out the prediction for the alternatives. The caveat here being the fact that the proposed model shows signs of over-prediction for higher frequency choice alternatives and under-prediction for lower-frequency choice alternatives. Quantifying the separate impacts of consideration set formation and choice model on prediction remains a task for future research.

Our results resonate with those of Pramono and Oppewal (2012) who provide evidence in favour of the effectiveness of a two-stage approach in modelling residential location choices through various scenarios. Their models reveal that the two-stage approach, particularly the Cross-Nested Logit (CMNL) model, outperforms the one-stage Multinomial Logit (MNL) model, under increased spatial restrictions. They indicate that the two-stage model better captures the elimination process in the first choice stage, with restricted accessibility. This effect is highlighted by the different choice outcomes between similarly attractive and closely located alternatives. For instance, in a particular scenario, the CMNL shows significant increases in choice probabilities for more spatially superior alternatives, demonstrating its ability to effectively model the impact of accessibility restrictions. The two-stage approach better represents the decision-making process by considering the initial elimination of alternatives based on spatial factors before assessing attractiveness attributes. This underscores the significance of stage-wise modelling in accurately representing the impact of spatial restrictions on residential location choices. Zolfaghari et al., (2013) provide application of different choice set approaches, including Universal Choice Set (UCS), Simple Random Sampling (SRS), and Deterministic Constraint (DC), thereby, demonstrating the effectiveness of the housing screening model in reducing the size of the choice set.

3.5 CONCLUSION

In this chapter, we set out to address the consideration of alternatives in a large alternative residential location choice model. For this task, we employed probabilistic decision trees to predict an household's consideration set. Our motivation for the use of decision trees is to allow a non-compensatory protocol for consideration of alternatives. This allowed us to generate a

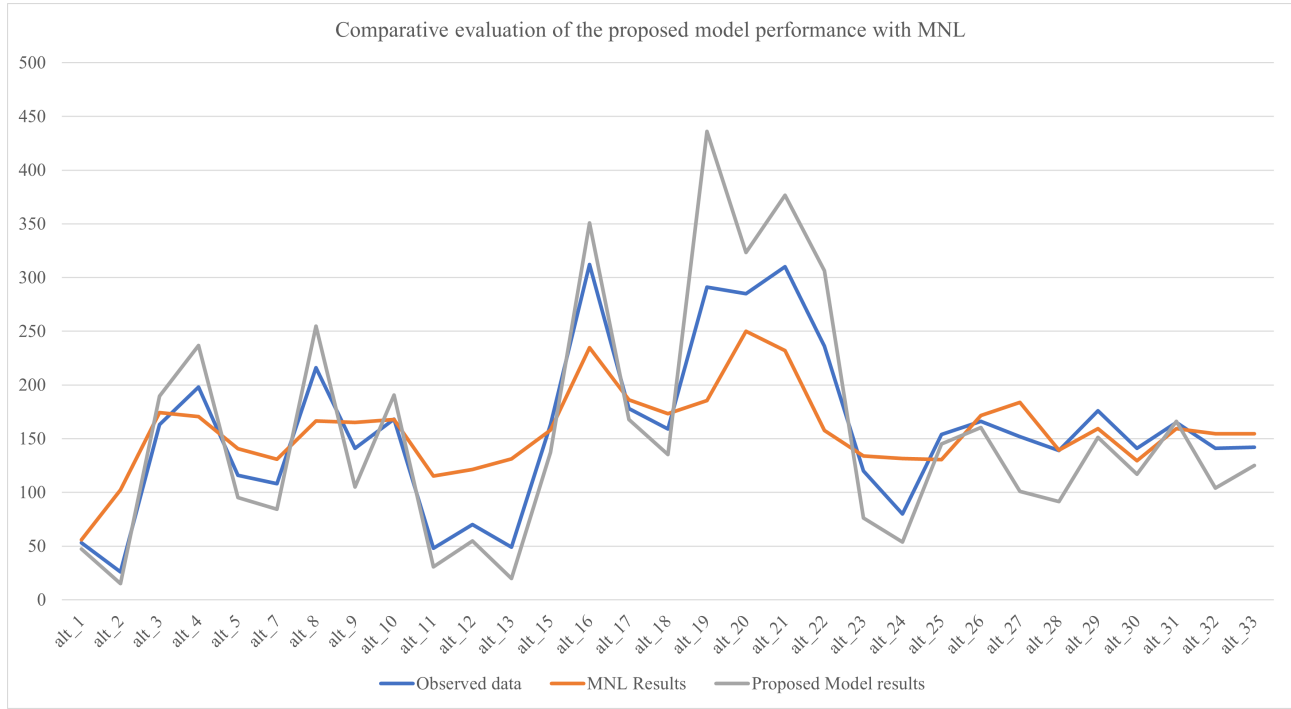


FIGURE 3.4 Comparative evaluation of discrete choice predictions

Notes: x-axis is the alternatives 1 (*alt_1*) to 33 (*alt_33*), except 6 and 14, representing the boroughs in the choice scenario; y-axis represents the frequency of choice of alternatives.

unique consideration set for each decision-making household in the 2018-2019 London Travel Demand Survey (LTDS) data set. We merged this data containing information on households, people, trips, and vehicles with add-on data for housing price, public transport accessibility and crime density. The derived consideration set is the utilised in a Mixed Logit model to estimate the compensatory choice decision.

Firstly, our results provide an understanding of the factors which contribute to the formation of consideration set formation. Although, decision tree model does not allow for estimation of the relative factorial coefficients, they do however allow for us to identify the factors' relative impact levels on the outcome of interest. In this case, Figure 4 provided us with a visual depiction of household socioeconomic and demographic factors which played a role in filtering or consideration of alternatives in a residential location decision. The results further bolster the use of such non-compensatory protocols by means of probability distribution among alternatives. The model output sum of probabilities for the top 3 alternatives turns out to be greater than 0.5 with the remaining 28 alternatives with an average consideration probability of approximately 0.015, which means that for any specific

observation (or decision-maker) it can be assumed that only a few alternatives are practically evaluated and their attributes then play a role in making a decision. While linking disjunction-of-conjunctions to decision trees in estimation has been performed earlier (Brathwaite et al., 2017; Gutiérrez-Vargas et al., 2023), the utilisation of this method to obtain probabilistic outcome is a novel contribution of our analysis. We also overcome the limited applicability of non-compensatory decision processes in choice modelling due to difficulties in estimation (Stüttgen et al., 2012; Kohli and Jedidi, 2017). Finally, while PIAL or the k-mix model (Swait, 2001, 2009) allow the use of conjunctive or disjunctive rules to determine the "dominance" or "rejection" state of alternatives, their limitation to two alternative scenarios can be overcome by our methodology.

Secondly, the Mixed Logit model estimated on this predicted consideration set allows for flexibility in capturing the interdependence between alternatives, where the alternatives belonging a particular Transport Fare Zone (1 to 6) were allowed a specific fixed-correlation with one-another, while being independent of all other alternatives. This reduced the number of correlations to be modelled, while offering insight into the dependency between alternatives. We observed a negative impact of Borough price variable (average price of a 2-bedroom apartment) on choice probability for lower-income single-person households, with an increasing negative impact as income goes higher. While intuitively one might assume that a higher-income household might have a higher tolerance (or lower susceptibility) towards income effects, our results show that even higher income groups are just as much (if not more) adversely affected by higher housing prices. The results of the comparative heuristic-MNL model provided a positive effect size for the income variable, which is intuitively incorrect, and thus fails to capture the true effect of the variable. Similarly, we observe a varying effect size of population density on borough choice, mediated by household vehicle ownership. However, the effect of crime when mediated by household's racial ethnicity produces significantly varied results. While the effect of crime-density on location choice probability for white households was observed to be negative as expected, the same effect for non-white households was positive. This ethnic variation of crime-density's effect on location choice gives us an insight into the racial and ethnic segregation evident in London housing. On the other hand, the results of the comparative heuristic-MNL model provided a positive effect size for the crime-density variable, which is intuitively incorrect and thus fails to capture the true

effect of the variable.

Lastly, the results confirm our hypothesis that decision makers may employ a non-compensatory process to reduce the number of alternatives from the universal choice set, followed by a compensatory process to obtain a final choice. Moreover, the *Decision tree - Mixed Logit* framework allows us to capture behaviourally richer contributing factors. This analysis is not without its limitations. Complex hybrid models such as the one presented here are data-hungry, which severely limits their replication for other empirical scenarios. We acknowledge that the merger of machine learning models (such as Decision Tress) and traditional econometric models is still in its nascent stages, and thus have to go a long way in addressing the economic concerns of this union. Moving forward, it may be useful to develop discrete choice model-based recursive partitioning frameworks which capture deterministic heterogeneity across households by estimating separate effects for different household classes or socioeconomic groups.

Chapter 4

Simultaneity of household trip frequencies and accessibility

Abstract

This chapter examines the endogenous relationship between residential level of accessibility and household trip frequencies to tease out the direct and indirect effects of observed behavioural differences. We estimate a multivariate ordered probit model system, which allows dependence in both observed and unobserved factors, using data from the 2016 Transportation Tomorrow Survey (TTS), a household travel survey in the Greater Golden Horseshoe Area (GGH) in Toronto. The modelling framework is used to analyse the influence of exogenous variables on eight outcome variables of accessibility levels and trip frequencies by four modes (auto, transit, bicycle and walk), and to explore the nature of the relationships between them. The results confirm our hypothesis that not only does a strong correlation exist between the residential level of accessibility and household trip frequency, but there are also direct effects to be observed. The complementarity effect between auto accessibility and transit trips, and the substitution effect observed between transit accessibility and auto trips, highlight the residential neighbourhood dissonance of transit riders. It shows that locations with better transit service are not necessarily locations where people who make more transit trips reside. Essentially, both jointness (due to error correlations) and directional effects observed between accessibility and trip frequencies of multiple modes offer strong support for the notion that accessibility and trip frequency by mode constitute a bundled choice and need

to be considered as such. The rest of the chapter presents an introduction to the simultaneity between accessibility and modal trip frequency, followed by the contributions of this analysis, description of the data, methodological framework (formulation and conceptual framework), estimation results (exogenous and endogenous variable effects), discussion and conclusions. The core findings of this chapter are available as:

Singh, A.C., A. Faghih-Imani, A. Sivakumar, Y. Xi and E. J. Miller. A joint analysis of accessibility and household trip frequencies by travel mode (accepted for presentation at World Conference on Transportation Research (WCTR), Montreal, Canada, July 2023; R&R with Transportation Research: Part A).

4.1 INTRODUCTION

4.1.1 Background and motivation

Travel behaviour exhibits a complex relationship with built-environment attributes, which has been extensively studied in the literature (Cervero and Kockelman, 1997; Frank and Engelke, 2001; Bagley and Mokhtarian, 2002; Cervero and Duncan, 2003; Krizek, 2003; Shay and Khattak, 2005; Lee and Moudon, 2006; Copperman and Bhat, 2007; Vance and Hedel, 2007; Brownstone and Golob, 2009; Cao et al., 2009; Cao, 2010; Zhang et al., 2012; He and Zhang, 2014; Liao et al., 2015; Næss et al., 2018). The neighbourhood-level accessibility, measuring the impact of the immediate built environment on travel behaviour, may be homogenous across different travel modes, but the influence becomes disproportionate beyond the immediate vicinity. This can be observed through variations in accessibility levels across different travel modes. While several earlier studies have focused on immediate neighbourhood characteristics such as the availability of facilities, land use density, and street network features (Handy, Cao, and Mokhtarian, 2005; Mokhtarian and Cao, 2008; Cao et al., 2009; Singh et al., 2018), there are relatively fewer studies that assess the impact of built environment beyond the immediate neighbourhood (Krizek, 2003; Ewing and Cervero, 2010). In fact, one can argue that focusing on network-level accessibility can lead to a wider understanding of the factors influencing travel behaviour.

The network-level accessibility is defined as the capacity to reach valued destinations where opportunities, activities, and individuals are located. It is typically quantified based on travel impedance and the number of reachable opportunities (Xi, Miller, and Saxe, 2018). Ewing and Cervero (2010) observed that accessibility to destinations strongly affects household travel behaviour. Accessibility to job opportunities within an isochrone is the most common measure of accessibility used in regional planning (Boisjoly and El-Geneidy, 2017) and can be found in the analysis of land value (Du and Mulley, 2012; Iacono et al., 2008), commuting distance (Levinson, 1998; Manaugh et al., 2010), social equity (Manaugh and El-Geneidy, 2012; Foth et al., 2013) and mode share (Foth et al., 2014; Owen and Levinson, 2015). Therefore, we hypothesise that the observed levels of network-level accessibility (represented by cumulative opportunities approach measured by the number of jobs reachable within the isochrones) by different travel modes have a substantial impact on household travel behaviour. This study aims to investigate the direct and indirect effects between travel behaviour and the level of accessibility (represented

by level of access to job opportunities) by empirically analysing the direction and extent of their mutual influence, focusing on multi-modal household trip frequencies and accessibility. In the rest of this chapter, we will use the term accessibility, instead of referring to the multi-faceted network-level accessibility, for ease in presentation.

4.1.2 Simultaneity bias in accessibility and travel behaviour

The lack of identification of- and solving for- endogeneity issues is evident in travel-behaviour studies. There are several studies (such as De Vos et al., 2019; Haque et al., 2019; Zarabi et al., 2019; Farinloye et al., 2019; Kamruzzaman et al., 2020) that mainly focus on establishing and exploring unidirectional effects, thereby ignoring any possibility of an immediate or delayed feedback from endogenous choices. It can be argued that the findings of studies such as Janke and Handy (2019), which only consider the changes in attitudes towards cycling caused by residential relocation and not whether the relocation was caused due to an inherent desire to move to a neighbourhood with more active transport choices, produce biased and inconsistent estimates. Evidence supporting the prevalence of this type of simultaneity bias has been documented (Guevara and Ben-Akiva, 2006), for instance leading to inconsistent estimation of price coefficients in housing price models (Waddell, 1992; Bhat and Guo, 2004). Simultaneity bias has been observed in various activity travel behaviour models, dating back to the early 1990s (Kitamura et al., 1992; Golob et al., 1994). Despite the recognition of the importance of considering the simultaneity of choices to avoid erroneous parameter estimates (Pinjari et al., 2011) most research still models transport related choices independently of each other and considers the other choices as exogenous.

According to De Witte et al. (2013), variables used to explain mode choice include the number of household cars, variables linked to the built environment, land use and network accessibility and travel cost, which is linked to the decision to own a transit pass. Usually, these variables are considered to be explanatory variables for mode choice, which makes it difficult to distinguish between a causal relationship and pure correlation. While one may hypothesise that household car ownership has a significant impact on mode choice, it is also possible that a household has several cars because all adults in the household commute by car. In the same manner, the characteristics of the available transport modes may be strongly linked to the residential location. For instance, some commuters may have poor public transport options

available because they live in a low-density zone. The car dependency may then be partly due to a residential self-selection effect. Pinjari et al. (2011) summarise the various interdependencies between this continuum of choices into four categories. First, long-term choices have a direct effect on short-term choices. Second, the residential self-selection effect is due to the selection of individuals in residential areas based on lifestyle preferences, which are also related to car or transit-pass ownership or to mode choice. Third, bicycle, car or transit-pass ownership is endogenous with respect to mode choice (such as individuals commuting by train buy an annual transit-pass). Fourth, some choices are associative (and not causal) and can be explained by common unobserved latent variables (that is a household whose member has a transit-pass will have fewer cars available, maybe because of positive perceptions towards public transport modes or environmental concerns). In statistical terms, ignoring the interdependence of the transport related choices, and estimating separate models when there is clear endogeneity, is misleading and may result in inefficient coefficients.

Another important consideration to model development is future forecasts and policy design based on the relationship between residential location and travel behaviour. Forecasts for policy design may be seriously misinformed if the model estimation is inconsistent. An inconsistent estimation approach may invalidate model analysis and has been noted to be unavoidable in models of residential location choice (Guevara and Ben-Akiva, 2006) because of endogeneity, unless it is addressed methodically. Given the substantial simultaneity endogeneity between accessibility and travel behaviour attributes, it is fitting to develop models which could jointly model these as a bundle of choices made simultaneously rather than sequentially.

4.1.3 Contributions of this chapter

This chapter presents a modelling framework that simultaneously models the level of accessibility and travel behaviour, by different transport modes. Specifically, we generate accessibility measures in the form of cumulative opportunities to jobs (that is, the total number of employment opportunities that can be reached from the household location) within an isochrone, by four modes of travel: auto, transit, bicycle, and walk. The travel behaviour is then examined by investigating the total number of trips made by the household by each of the four modes of travel. While the factors such as price and dwelling attributes may have a greater impact on choice of residential location, the residential location characterises the

impact of accessibility on travel choices, including number of trips by private vs. public transport, proportion of active transport trips, and total vehicle miles travelled. Among the types of accessibilities observed, the role of employment accessibility cannot be downplayed because of its significant impact on the socio-economic well-being of the households. While measures such as walkability index, transit connectivity and land use diversity have been analysed in the literature (Cao and Fan, 2012; Kim and Brownstone, 2013; Paleti et al., 2013; Bhat et al., 2016), modelling employment accessibility and its impacts provides a critical understanding of socio-economic welfare of the household. We hypothesise that using employment accessibility alone as the endogenous variable may allow us a clean identification of its direct effect on travel behaviour, as well as any remaining self-selection effects, in explaining household trips without problems of multi-collinearity with other accessibility characteristics.

This chapter uses a multivariate ordered-response model system for analysing the level of accessibility to jobs and the number of trips (trip frequencies) undertaken by a household, respectively, by each mode (auto, transit, bicycle and walk). Both household trip frequencies and accessibility measures are categorised in ordered levels of very low, low, medium, high, and very high. For an analysis like this, the household size plays a critical role, as we can expect the larger households to prefer suburban neighbourhoods and the younger-smaller households to self-select the city-centre neighbourhoods. This effect of household size can be accounted for by either creating a complex size variable comprising number of people in the household, age, and gender. However, to explicitly tease out the effect of household size and type in our analysis, we choose to circumvent the household size effect by comprehensive structuring of household types, including single adult with no kids, couple adult with no kids, single or couple with one or more kids, multi-adult with no kids, and multi-adult with one or more kids. In this model system, we allow dependence between the trip frequency and accessibility by different modes due to both observed and unobserved factors. The observed factors are accounted for by the exogenous variables available in the dataset. The inclusion of unobserved factors allows complementarity and substitution effects in trip counts (even after controlling for observed effects). For instance, individuals in the higher income bracket may have a higher propensity to driving but still live in transit-rich neighbourhoods. This would create a complementary relationship between frequency of driving trips and transit trips. Similarly, individuals with

a more active lifestyle may have a higher propensity to bicycle or walk, but not auto. This represents a substitution relationship. Overall, the extent of complementary and substitution relationships may be specific to the combinations of level of accessibility and trip frequencies, respectively, by each mode.

In this study, we demonstrate the applicability of the modelling framework using travel survey data from the 2016 Transportation Tomorrow Survey (TTS), a household travel survey in the Greater Golden Horseshoe Area (GGH) in Toronto (Data Management Group, 2016). The dataset is augmented with accessibility generated measures by different travel modes for the households' residential locations. The multivariate ordered probit model is structured in a way to tease out relationships among the endogenous variables (such as accessibility to opportunities by auto mode and auto trip frequency) while explicitly accounting for the influence of exogenous variables such as socio-demographic characteristics and other accessibility attributes. While the previously mentioned literature clearly motivates the need for methodical approaches to disentangle endogeneity, we attempt to fill in this gap by disentangling the bidirectional relationship between accessibility and trip frequency in an empirical context.

4.2 DATA DESCRIPTION

The data used for this study is a sample taken from the 2016 Transportation Tomorrow Survey (TTS). The TTS is a household travel survey collecting travel information from 5% of households in the GGH that has been conducted every five years since 1986 (Data Management Group, 2016). The TTS collects individual and household information in addition to data regarding the trips made in a day, such as travel mode, start time, etc. After removing the households with incomplete and inconsistent attributes, the remaining 42,718 households residing in the City of Toronto were employed from the dataset. These households made 196,221 trips, which are the basis of this analysis. The dependent variables of interest in this chapter are the household trip frequency by each mode and the respective modal accessibility measure calculated using home location. Trip records provided by the households were used to derive the total modal trip frequency estimates for each household. The modes chosen by the household were clustered into categories for auto, transit, bicycle and walk and aggregated at the household level. Modal accessibility measures were calculated for each

dissemination area (DA) in the City of Toronto and assigned to the households based on the DA of their home location. Dissemination areas are the smallest of Canadian census agglomerations, with populations between 400 and 700 people. We generated travel time isochrones by each mode and calculated the cumulative job opportunities captured within the isochrone. Different time cut-offs have been tested and at the end, 45-minutes is used for auto and transit modes and 30-minutes is used for bicycle and walk modes.

Auto and transit travel times were obtained from GTAModel V4.1 for the morning peak period (Travel Modelling Group, 2015) which accounts for road congestion and travelled (rather than posted) speeds. For bicycle and walk isochrones, a uniform speed of 15 km/h and 4 km/h were assumed. For bicycle accessibility measures, we also accounted for Level of Traffic Stress (LTS) on the road network. The LTS metric classifies street segments between LTS 1 (low stress) and LTS 4 (high stress), considering primary roadway attributes including: speed limit, number of vehicular lanes, bicycle lanes, parking and traffic signals (Furth et al., 2016). We defined the level of cycling stress for every link in Toronto's Road network. Using this street stress network, the 30-minute cycling isochrone and associated cumulative opportunities to jobs were generated using only the road links with low stress levels of 1 and 2 (for detailed discussion on cycle accessibility and level of traffic stress, see Faghih Imani et al., 2019). To accommodate the household size effect, the dataset comprising 42,718 households has been divided into smaller groups by comprehensive structuring of household types: single adult with no kids (13,613 households), couple adult with no kids (14,781 households), single or couple with one or more kids (6,526 households), multi-adult with no kids (5,660 households), and multi-adult with one or more kids (2,138 households).

A set of variables were extracted from the TTS survey, including the trip distance, household characteristics such as number of vehicles in the household, and household income. In addition to these, the distance from each household home location to the Toronto City Hall was computed as a measure to identify the impact of proximity to the central business district (CBD). The population density and job density at home location were generated at DA level and appended to the dataset. Table 4.1 provides an overview of the descriptive characteristics of the sample. Table 4.2 provides a distribution of the dependent variables as per the ordered scales of very low, low, medium, high and very high trip frequency. For auto trip frequency, zero trips per household is considered to be in the very low category, between one and two

trips is considered to be in the low category, between three and four trips are considered to be in the medium category, between five and seven trips are considered to be in the high category, and eight or more trips is considered to be in the very high category. For transit, bicycle and walk trip frequencies, zero trips per household is considered to be in the very low category, one trip is considered to be in the low category, two trips are considered to be in the medium category, between three and four trips are considered to be in the high category, and five or more trips are considered to be in the very high category. For auto, transit, bicycle and walk access, the continuous value distribution obtained from the accessibility calculation is divided into very low, low, medium, high, and very high categories by using the 20%, 40%, 60%, and 80% quantiles.

TABLE 4.1 Description of Sample (n=42,718)

Independent Variables Distribution				
Variable definition	Mean	S.D.	Min	Max
Fraction of female adults in the households	0.54	0.32	0.00	1.00
Fraction of male adults in the households	0.46	0.32	0.00	1.00
Distance to CBD (km)	10.18	6.05	0.00	26.77
Population Density (persons per sq. km)	14656.68	19416.81	0.00	240911.54
Employment Density (persons per sq. km)	6412.27	21079.68	0.00	263070.47
Population Density (Standardized value*)	0.00	1.00	-0.75	11.65
Employment Density (Standardized value*)	0.00	1.00	-0.30	12.18

Notes: *created by subtracting the mean from each value and dividing by the standard deviation (same as z scores).

TABLE 4.2 Trip frequency observations (%) by each mode (n=42,718)

Variable definition	Very Low	Low	Medium	High	Very High
Auto Trips	15588 (36.5)	9945 (23.3)	7771 (18.2)	5404 (12.7)	4010 (9.4)
Transit Trips	24588 (57.6)	1965 (4.6)	10139 (23.7)	4425 (10.4)	1601 (3.7)
Bike Trips	40568 (95.0)	110 (0.3)	1276 (3.0)	534 (1.3)	230 (0.5)
Walk Trips	34327 (80.4)	2097 (4.9)	4026 (9.4)	1722 (4.0)	546 (1.3)

To facilitate a broader understanding of the study area, and to provide empirical understanding of the Toronto region, the modal trip categories and accessibility categories are overlaid on the Toronto city map in Figures 4.1 and 4.2. Figure 4.1 shows a map of Greater Toronto region with modal accessibilities, clockwise (from top right in the figure) the levels of transit, walk, bicycle and auto accessibility (green for lowest category 1, up to red for highest

category 5). Similarly, Figure 4.2 shows a map of Greater Toronto region with modal trip frequencies, clockwise (from top right) the levels of transit, walk, bicycle and auto trips (green for lowest category 1, up to red for highest category 5). For transit mode, we can observe that while the south-central region provides highest levels of accessibility, the trip distribution is spread throughout the region, with no clear central transit trip demand. The east and the west sides of the region represent the lowest levels of transit accessibility, with non-zero levels of transit trips observed. For walk and bicycle modes, the accessibility representation shows multiple high-level zones, however, with very low levels of trips made by these modes. Lastly, for the auto mode, the zones with high levels of accessibility are evenly spread throughout the region, proving how the infrastructure is tilted towards auto mode choice. However, only south-central region in the map can be shown to have minimal number of auto-trips, as represented by more proportion of the green. To summarize, the auto-supportive infrastructure is available throughout the Toronto metropolitan region, providing higher levels of auto accessibility. On the other hand, the transit and bicycle networks primarily serve the city-center regions, with lower level of accessibility by transit and bicycle in the outer urban and suburban neighbourhoods.

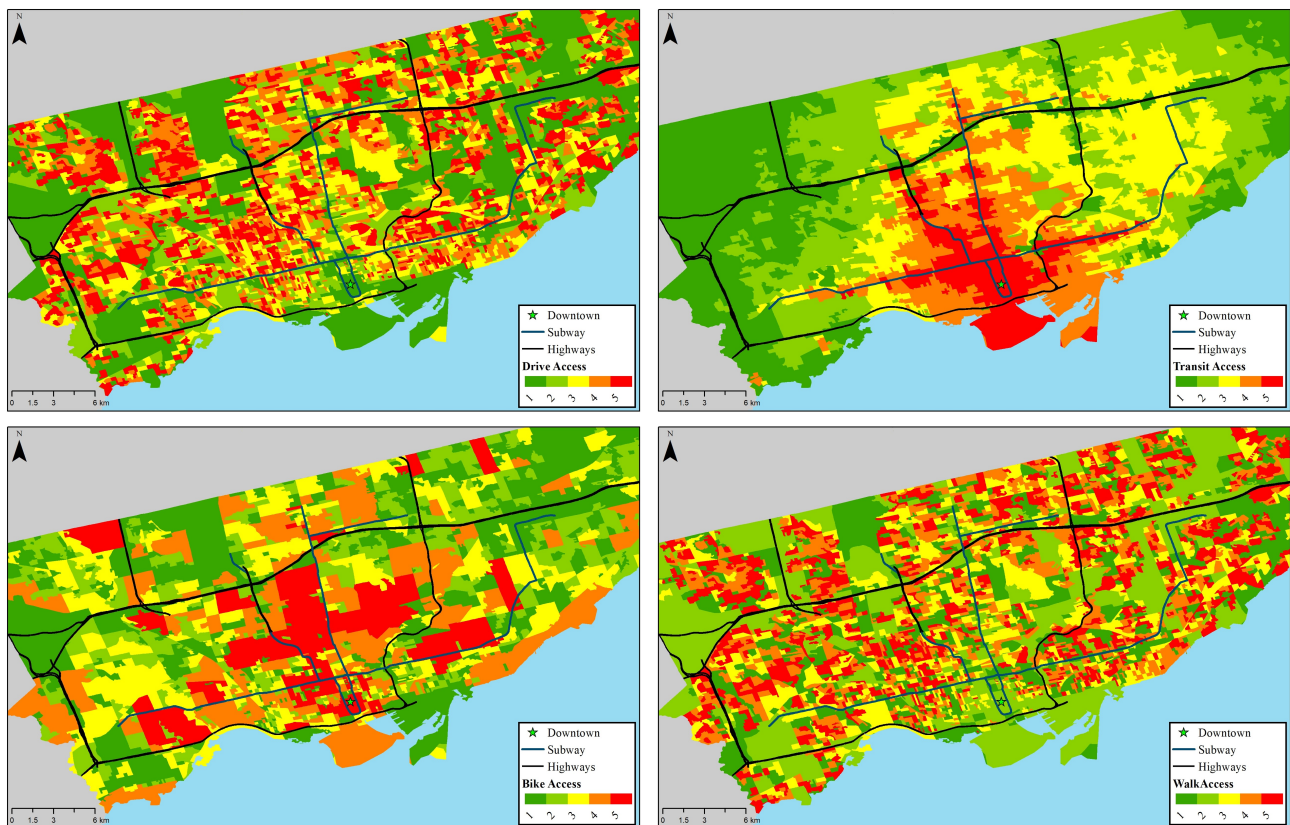


FIGURE 4.1 Map of Greater Toronto region with modal accessibilities, clockwise (from top right) the levels of transit, walk, bicycle and auto accessibility (green for lowest category 1, up to red for highest category 5)

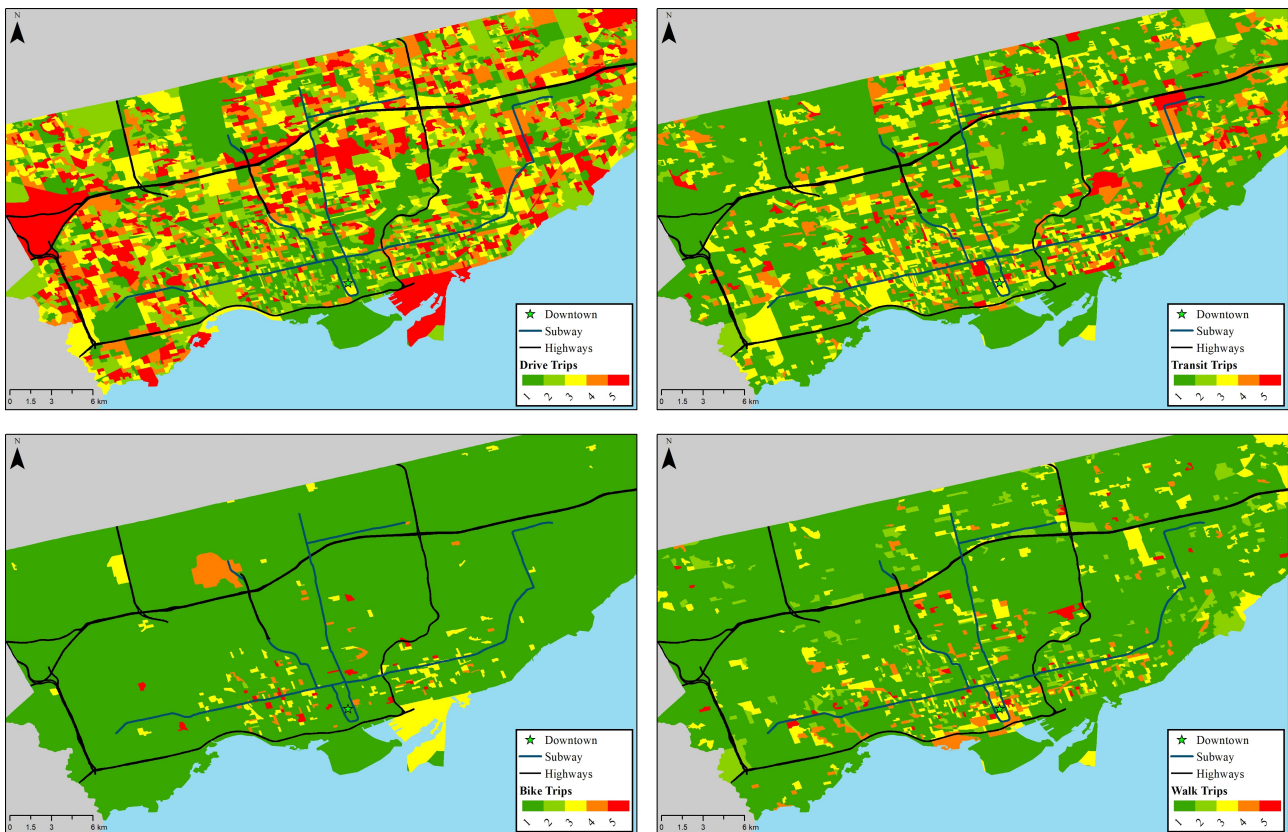


FIGURE 4.2 Map of Greater Toronto region with modal trip frequencies, clockwise (from top right) the levels of transit, walk, bicycle and auto trips (green for lowest category 1, up to red for highest category 5)

4.3 METHODOLOGICAL FRAMEWORK

For the household trip frequencies by each mode, the multivariate model system used in this chapter assumes an underlying set of multivariate continuous latent variables whose partitioning maps into the observed set of count outcomes. Based on the observed data, these count outcomes are translated into ordered categories of low, medium, and high trip counts. Examples of cases where the ordered-response system is used to model count outcomes include household car ownership levels (Kitamura, 1988; Bhat and Guo, 2007) and trip generation/stop-making (Meurs, 1989; Agyemang-Duah et al., 1995; Agyemang-Duah and Hall, 1997; Carrasco and Miller, 2009). For the accessibility by each mode outcome, the observed continuous outcomes are mapped into ordered categories as well¹. Thus, this joint model system can be estimated using a multivariate ordered-response framework. The ordered-response system allows the use of a general correlation matrix for the underlying latent variables, which translates to a flexible correlation pattern among the observed outcomes. Even though the ordered-response models are used for ordinal responses, the distinction between ordinal and cardinal responses is irrelevant for modelling purposes². This is valid in the cases where the count responses are finite and have a few discrete values, as in the current empirical application.

To avoid the convergence problems associated with estimating a joint multivariate ordered-response system with a large number of categories, we use the composite marginal likelihood technique (Varin and Vidoni, 2005; Varin, 2008). This chapter considers the eight-variate choice system as a bundled choice and explores the interrelationships among them, without imposing a structure a priori, wherein the accessibility is determined first and then exogenously affects the modal trip frequencies. This effort is motivated by the desire to seek answers to the questions described in section 1 where we aim to understand the

¹The authors have taken into account that aggregating integer data for accessibility to cumulative opportunities to jobs into categories can be limiting, but in this case the actual integer values for accessibility cannot be meaningfully interpreted for inferential analyses, and the sheer range of these values makes it computationally unreasonable to be used as ordered dependent categories. Therefore, we categorize modal accessibility into discrete groups which can be comparatively analysed and meaningfully interpreted.

²A perceived problem in the latter case may be that the ordered-response model entails the estimation of $N-1$ threshold values that horizontally partition the underlying continuous variable to map into the observed count values, where N is the largest possible count value. But, as has been demonstrated by Meyer (1990), there is little loss of efficiency due to the estimation of a large number of thresholds in the ordered-response model structure. As long as there are a few observations in each of the N categories under consideration, it is straightforward to estimate the ordered-response structure.

directionality, type, and extent of effect between the variables of interest. An exploration of the interrelationships between the accessibility by mode and trip frequencies by mode would help advance an understanding of how the availability (or the lack thereof) of certain modes influences trip making decisions, and whether higher accessibility by certain modes is a precise criterion for travellers' mode choice. This section presents the methodological approach used in this study. The model formulation (Multivariate Ordered Probit model) is presented first; this is followed by a description of the conceptual framework used to explore inter-outcome dependencies.

4.3.1 Model formulation

Let q be an index for households, i be the index for ordered response category (where I denote the total number of dependent categories for each individual) and number of ordered response values for category i be $K_i + 1$, such that:

$$q = 1, 2, \dots, Q \quad (4.1)$$

$$i = 1, 2, \dots, I \quad (4.2)$$

$$k = 1, 2, \dots, K_i \quad (4.3)$$

where $Q = 42,718$, in the current study, $I = 8$, and the discrete levels k belong in $0, 1, 2, \dots, K_i$ for category i .

This model found its origin in biostatistics and social sciences (Aitchison and Silvey, 1957; McKelvey and Zavoina, 1975) long before it was used for applied econometric studies (Lesaffre and Molenberghs, 1991; Kim, 1995). We can write the latent propensity (y_{qi}^*) for each ordered response as a function of its covariates, and this latent propensity can be translated into the observed outcome (y_{qi}) through threshold bounds:

$$\mathbf{y}_{qi}^* = \boldsymbol{\beta}_i' \mathbf{x}_{qi} + \boldsymbol{\epsilon}_{qi} \quad (4.4)$$

$$y_{qi} = k, \text{ if } \theta_i^k < y_{qi}^* < \theta_i^{k+1} \quad (4.5)$$

where x_{qi} is a $(L * 1)$ vector of exogenous variables (L does not include a constant), β_i is a corresponding $(L * 1)$ vector of coefficients to be estimated, ϵ_{qi} is a standard normal error term, and θ_i^k is the lower bound threshold for count level k of ordered response category i ($\theta_i^0 < \theta_i^1 < \theta_i^2 \dots < \theta_i^{K_i+1}$, $\theta_i^0 = -\infty$, $\theta_i^{K_i+1} = +\infty$, for each category i). The ϵ_{qi} terms are assumed independent and identical across individuals (for each and all i). For identification reasons, the variance of each ϵ_{qi} term is normalized to 1, to correctly identify and estimate the threshold values. However, we allow correlation in the ϵ_{qi} terms across ordered response categories i for each individual q .

Specifically, define $\epsilon_q = (\epsilon_{q1}, \epsilon_{q2}, \epsilon_{q3}, \dots, \epsilon_{qI})'$. Then, ϵ_q is multivariate normal distributed with a mean vector of zeros and a correlation matrix as follows:

$$\epsilon_q \simeq N \left[\begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho_{12} & \rho_{13} & \cdots & \rho_{1I} \\ \rho_{21} & 1 & \rho_{23} & \cdots & \rho_{2I} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ \rho_{I1} & \rho_{I2} & \rho_{I3} & \cdots & 1 \end{pmatrix} \right] \text{ or } \epsilon_q \simeq N[0, \Sigma] \quad (4.6)$$

The off-diagonal terms of Σ capture the error correlations across the underlying latent continuous variables of the different ordered response categories; that is, they capture the effect of common unobserved factors influencing the propensity of choice of count level for each ordered response. Thus, if ρ_{12} is positive, it implies that households with a higher-than-average propensity in their data sample towards the first ordered response variable are also likely to have a higher-than-average propensity towards the second ordered response variable. Of course, if all the correlation parameters (i.e., off-diagonal elements of Σ), are identically zero, the model system in Equations 3.4 and 3.5 collapses to independent ordered response probit models for each ordered response. Let the actual observed count level for household q and ordered response category i be m_{qi} . Then, the likelihood function for household q may be written as follows:

$$L_q(\delta) = Pr(y_{q1} = m_{q1}, y_{q2} = m_{q2}, \dots, y_{qI} = m_{qI}) \quad (4.7)$$

$$L_q(\delta) = \int_{\nu=\theta_1^{m_{q1}+1}-\beta'_1 x_{q1}}^{\theta_1^{m_{q1}+1}-\beta'_1 x_{q1}} \cdots \int_{\nu=\theta_I^{m_{qI}+1}-\beta'_I x_{qI}}^{\theta_I^{m_{qI}+1}-\beta'_I x_{qI}} \phi_I(\nu_1, \nu_2, \dots, \nu_I \mid \Sigma) d\nu_1 d\nu_2 \cdots d\nu_I \quad (4.8)$$

where $\phi_I(\nu_1, \nu_2, \dots, \nu_I \mid \Sigma)$ represents the multivariate normal density of dimension I with correlation matrix Σ evaluated at the abscissae $(\nu_1, \nu_2, \dots, \nu_I)$. In order to overcome the computational complexity involved in calculating a high-order I -dimensional rectangular integral, this study employs a Composite Marginal Likelihood (CML) computation approach, which involves approximating the higher-order integral through the computation of a series of bivariate marginal distributions; the CML estimator (theoretically speaking) loses some accuracy relative to traditional maximum likelihood estimation, though this efficiency loss has been showed to be negligible in practice (Varin and Vidoni, 2005). The model is estimated using R ‘mvord’ package (Hirk et.al., 2017), MASS package (Venables and Ripley, 2013) and Python software written by the authors.

4.3.2 Conceptual framework

The multivariate ordered probit modelling methodology is used to analyse the influence of exogenous variables on the eight outcome variables of accessibility levels and household trip frequencies and to explore the nature of the relationships among them. As noted earlier, the eight endogenous variables represent the four accessibility levels (one for each mode) and four trip-counts:

- Accessibility by auto, labelled as “Auto Acc”
- Accessibility by transit, labelled as “Tr Acc”
- Accessibility by bicycle, labelled as “Bicycle Acc”
- Accessibility by walk, labelled as “Walk Acc”
- Trip frequency by auto, labelled as “Auto Trips”
- Trip frequency by transit, labelled as “Tr Trips”
- Trip frequency by bicycle, labelled as “Bicycle Trips”

- Trip frequency by walk, labelled as “Walk Trips”

The presence of the eight endogenous variables gives rise to many possible specifications of relationships among them. With respect to the myriad of relationships among the endogenous variables, four model structures were considered. These are depicted in Figure 4.3 (as H1 through H4). They include:

- H1: Accessibility by mode affects corresponding Trip frequency
- H2: Trip frequency by mode affects corresponding Accessibility
- H3: Accessibility by mode affects corresponding Trip frequency with added interplay between Auto-Transit accessibility affecting each-other’s Trip frequency and same for Bicycle-Walk.
- H4: Accessibility by each mode affects all modal Trip frequencies

In all specifications, an error correlation structure is incorporated to accommodate the bundled nature of the many choices. Thus, the directionality of relationship of an observed endogenous variable to the propensity underlying another endogenous variable can only be in one direction. The reader will note that a recursive endogenous effect of one observed ordinal variable on the propensity of the other observed ordinal variable is possible in the equation system, although including both directions of observed effects lead to a logically inconsistent model structure because the probabilities of all the possible combinations of discrete observations will not sum to one (see Maddala, 1983 and 1986). Intuitively, the propensities are the precursors to the actual observed variables, and, when both the decisions are co-determined, it is impossible to have both observed variables structurally affect one another. In the current estimation effort, all models with every possible impact direction were estimated and the best fit version was chosen. However, it must be noted that regardless of the directionality between endogenous outcomes and its significance, the entire model system is a joint bundle of choices because of the correlation in the unobserved factors affecting underlying propensities. After estimating model specifications corresponding to H1 through H4, the analysis involved using goodness-of-fit measures, behavioural intuitiveness, and statistical significance of coefficients to arrive at a final model structure that was both intuitive and supported by the data. The final model structure is based on the hypothesis H4,

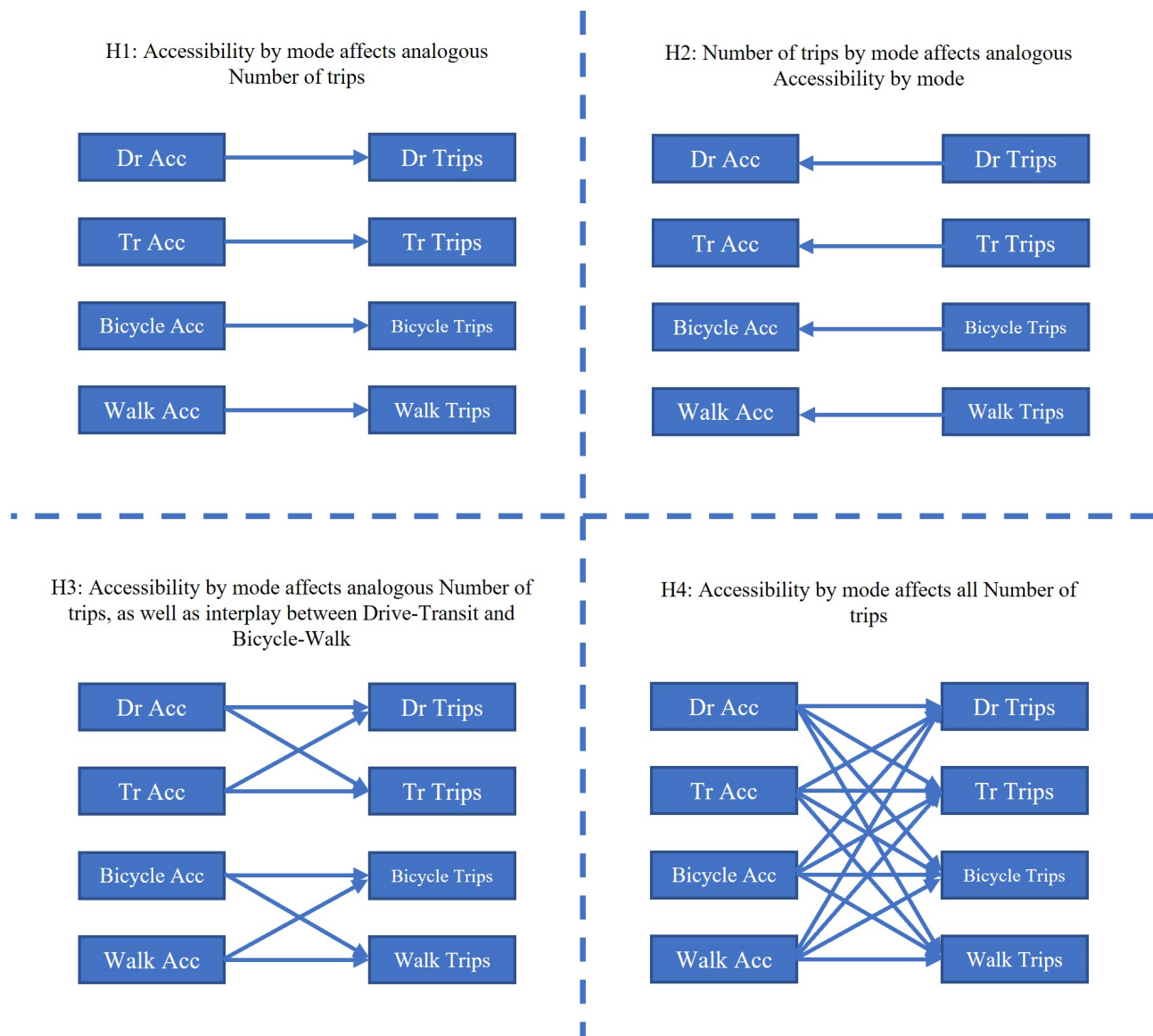


FIGURE 4.3 Hypotheses Regarding Relationships Among Endogenous Variables

which provides the highest goodness-of-fit and incorporates all cross-modal endogenous effects between accessibility and trip frequency. Models H1 and H3 were building blocks to the model H4, while H2 had no statistically significant endogenous effects. Thus, in addition to being contextually dependent, the results presented in this chapter are fundamentally empirical.

4.4 ESTIMATION RESULTS

The final model structure is based on the hypothesis H4. The results shown here are based on statistical significance, providing inferential information about the given empirical context. Additional evaluations of probability of correct predictions offered by the model are suppressed in the interest of brevity. Essentially, both jointness (due to error correlations) and directional effects need to be accommodated. Finally, in the remaining sections, we discuss the important results for each of the household structure groups and their interpretations. As mentioned before., the household structure models are estimated independently for single adult with no kids, couple adult with no kids, single or couple with one or more kids, multi-adult with no kids, and multi-adult with one or more kids. For brevity, this section focusses entirely on the results for “couple adult with no kids” household type for the hypothesis H4, which represents a household with two adults over the age of 16 years and zero child living together. We have chosen this household structure as it constitutes a greater proportion of total households in the region. While the results focus on this singular category, the findings of other household structure estimations are also similar and are available in the ‘Appendix’ chapter.

4.4.1 Exogenous variable effects

Tables 4.3, 4.4, 4.5 and 4.6 present estimation results depicting the influence of the exogenous variables on the various modal trips and accessibilities, respectively, for the multivariate models for each of the structures with best goodness-of-fit. While the modal trip frequencies are directly observed in the data set (in the form of household trips and primary mode for each trip), the accessibility values are calculated as detailed in the earlier sections. Dummy variables for presence of people less than 55 years of age are found to affect transit trip frequencies significantly and positively, as compared to those above 55 years of age. Transit trip frequencies are almost equally affected by people in all age-groups, with a slightly higher coefficient observed for households with young adults. This may be due to the tendency of young couples to prefer train and bus services, coupled with lower affordability of cars and self-selection into neighbourhoods closer to city centre. The propensity to engage in driving trips is negatively associated with presence of younger individuals in the household, while a positive effect is observed for presence of 56 years of age or older. Lower bicycle trip frequencies are associated with households with people in the age-group 66 and above, which can be attributed to the

fact that riding a bicycle can be risky and requires higher agility and fitness levels. Similarly, lower bicycle trip frequency is observed for individuals in the age group 16-25 which refers to high-school and university students. This could be attributed to either preference of driving to school or for university students the impact of living close to campus. This is further attributed by a higher walk trip coefficient for this age group, suggesting more walk trips, probably due to living on-campus or closer to central zones. We observe a negative association of walk trips with age group 66 years and above. Our results are similar to those observed by earlier research focusing on Canadian individuals signifying that younger individuals are more likely to walk, bike, or use transit (Newbold and Scott, 2017; Newbold and Scott, 2018), displaying a higher level of multimodality compared to older individuals (Sakaria and Stehfest, 2013). In contrast, older adults exhibit a decrease in the probability of using transit as they age (Moniruzzaman et al., 2013). It has been suggested that older seniors are resistant to adopting alternative transportation modes, although younger seniors show an increasing trend in the use of alternative forms of transport (Fordham et al. 2017). Reasons for this resistance may include unfamiliarity with the service, lack of habituation, and limited knowledge of the required technology, such as smartphone applications (Fordham et al. 2017).

Presence of students in the household has a positive association with auto, transit and walk trip frequencies, with the highest coefficient observed for transit trips which can be attributed to the fact that students often live with a limited expense budget and enjoy incentives to use public transport which induces a preference towards the mode. Presence of students in the household has a positive association with transit, bicycle and walk trip frequencies, with higher coefficients observed for transit and bicycle trips, which can be attributed to the fact that students often live with a limited expense budget and enjoy incentives to use public transport which induces a preference towards the mode. Our results are in line with Brown et al., (2016) who conclude that younger adults were more likely to ride transit than older adults due to life cycle factor such as being a student. Distance to CBD (measured in kilometres) is positively associated with auto and transit trip frequencies and negatively associated with bicycle and walk trip frequencies. The results presumably indicate the presence of more bicycle and walk friendly neighbourhoods close to downtown areas, while as we move away from the CBD, driving is preferred to bicycle and walk modes due to the increasing distance between destinations and lack of cycling and walking infrastructure. The

positive parameter for transit trips for distance to the CBD is also consistent with the observed levels of transit trips, as shown in Figure 2. Contrary to expectations, the higher levels of trips by transit mode can be observed in regions surrounding the city-centre, with few much farther from the CBD. The accessibility by each mode is negatively associated with the distance to CBD, denoting that the extent of employment opportunities reachable reduces as we move further away from the CBD. This trend can be seen in cities other than Toronto, as evident in results by Moniruzzaman et al., (2015) who found that distance to CBD negatively impacts walking trips but positively impacts vehicular trips in the city of Montreal.

Higher population density is associated with lower auto and bicycle trips but higher walk trips, which may be attributed to the fact that higher density neighbourhoods are often more congested and may incorporate lesser driving friendly infrastructures. On the other hand, higher population density is associated with higher auto and transit accessibilities, which suggests that the amount of employment opportunities by each of these modes is higher in regions where there is a greater density of settlement. A positive association of bicycle accessibility with employment density indicates that the cycling infrastructures are better placed for the locations with higher employment density. This is consistent with the finding of Imani et al. (2019) who found that due to lack of cycling infrastructure and high stress roads, the low-stress accessibility measures are frequently limited by space rather than the cut-off time (that is, there are often not enough available low-stress links for a 30-minute isochrone). Further, as expected, higher employment density is associated with higher transit accessibility and lower auto access. However, we observe a counterintuitive negative association with walk access.

Statistically significant income coefficients suggest that high-income households are more likely to engage in activities by driving, consistent with their greater vehicle ownership and ability to afford costs associated with driving. The coefficients increase consistently with the high income-levels in the group. Lower-income households are also observed to make fewer transit trips than their rich counterparts. Lower trip frequencies observed in lower income groups has been observed to be attributed to the cause of time-poverty which restricts the extent of non-mandatory activities that lower income group households can participate in and equity issue regarding the distribution of transport infrastructure in Toronto (Allen and Farber, 2020). Results from across the border in California (Boarnet et al., 2020) indicate that while high-income households do make a larger number of trips, there may be a

possibility of reduction in vehicular travel for residences very close to transit stations. However, they find no evidence of any income differences in changes in the propensity to use transit when living near versus far from a transit station. This conclusion is particularly fascinating as it bolsters our conclusion regarding neighbourhood dissonance of transit riders. Accessibility to opportunities by transit is observed to be negatively associated with all income groups, signalling towards a need for increasing overall transit accessibility. However, the higher income households have a slight advantage as compared to lower-income counterparts as they may reside in residential locations which have better public transport infrastructure and lower level of traffic stress for reaching a higher number of employment opportunities. Accessibility by bicycle increases with income level, as it can be estimated that higher income neighbourhoods provide better bicycling facilities and an overall better built-environment infrastructure. We observe a higher propensity to transit trips and a lower propensity to bicycle trips for households which have a higher fraction of female adults, consistent with previous findings (Mitra et al., 2016; Imani et al., 2019). Vehicle availability per licensed adult is found to significantly impact all modal trip frequencies. Greater than zero vehicle availability is associated with a higher number of auto trips, and a lower number of transit, bicycle and walk trips, as compared to a household with zero vehicle availability. This is consistent with the expectation that higher auto availability facilitates driving trips and inhibits trips by other modes of transport.

However, we observe a lower accessibility to opportunities by driving and transit for the households with vehicles. This could refer to a possibility that vehicle owners may reside in regions which are further away from locations of employment opportunities and may be willing to auto longer distances, or due to the fact that some low-income people without vehicles reside in suburban regions in Toronto with very high auto access. We also observe a lower transit accessibility for household with more than one vehicle available per licensed adult. Transit pass ownership significantly impacts driving negatively and transit trip frequencies positively, intuitively suggesting that a higher transit pass ownership suggests that a household makes higher transit trips and lower trips by other modes of transport. Households with workers show a significantly higher propensity to participate in trips as compared to household with no workers. Dual-worker households are more likely to participate in fewer activities than single-worker households, which could be due to both members participating in multiple mandatory

and non-mandatory trips. Finally, housing type significantly impacts auto, transit and bicycle trips. Apartment dwellers have a lower propensity to participate in auto and bicycle trips, and a higher propensity to make transit trips. This can be attributed to the fact that apartment dwellers usually have a lower vehicle ownership as compared to house or town house owners and usually reside in transit rich neighbourhoods. Apartment dwellers also have a lower auto and walk accessibility, which may refer to a distribution of non-egalitarian walk-friendly facilities. Empirically, the City of Toronto does have a huge number of high-rise apartment buildings in the suburban neighbourhoods that generally have relatively lower non-motorised accessibility, as compared to urban neighbourhoods closer to city centre.

TABLE 4.3 Exogenous Variable Effects on Modal trip frequencies

Variable definitions	Auto Trips		Transit Trips		Bicycle Trips		Walk Trips	
	Estimate	z value	Estimate	z value	Estimate	z value	Estimate	z value
<i>(Base: Age category 0-15 years)</i>								
Presence in age category 16-25 years	--	--	0.2743	6.85	-0.2890	-4.32	0.1069	2.24
Presence in age category 26-35 years	-0.1034	-3.86	0.2355	7.80	--	--	--	--
Presence in age category 36-45 years	-0.0531	-1.93	0.1625	5.43	--	--	--	--
Presence in age category 46-55 years	-0.0788	-3.14	0.1501	5.16	--	--	--	--
Presence in age category 56-65 years	0.1172	4.34	--	--	--	--	--	--
Presence in age category over 65 years	--	--	--	--	-0.4601	-7.62	-0.1022	-2.62
Presence of students in the household	--	--	0.2819	8.37	0.3018	5.58	0.1721	4.04
Distance to CBD (km)	0.0113	4.88	0.0118	4.07	-0.0931	-21.47	-0.0705	-22.97
Population Density (Standardized value)	-0.0291	-2.77	--	--	-0.0973	-4.49	0.0450	3.42
Employment Density (Standardized value)	-0.0208	-2.18	-0.0526	-5.06	-0.1252	-6.13	0.0683	6.23
<i>(Base: Income less than 15,000\$)</i>								
Income 15,000\$-40,000\$	—	—	0.0933	1.56	—	—	—	—
Income 40,000\$ to 60,000\$	0.1997	6.38	0.1084	1.76	—	—	—	—
Income 60,000\$ to 100,000\$	0.3342	11.29	0.2287	3.85	—	—	—	—
Income 100,000\$ to 125,000\$	0.4032	11.51	0.3286	5.18	—	—	—	—
Income greater than 125,000\$	0.4907	15.35	0.3162	5.08	—	—	—	—
<i>(Base: fraction of male adults in the household)</i>								
fraction of female adults in the household	—	—	0.1909	3.94	-0.2627	-2.94	—	—

TABLE 4.4 Exogenous Variable Effects on Modal trip frequencies (continued)

Variable definitions	Auto Trips		Transit Trips		Bicycle Trips		Walk Trips	
	Estimate	z value	Estimate	z value	Estimate	z value	Estimate	z value
<i>(Base case: Vehicle availability per licensed adult is zero)</i>								
Vehicle availability per licensed adult is between zero and one	1.3931	42.01	−0.5656	−19.43	−0.4609	−10.46	−0.2780	−8.47
Vehicle availability per licensed adults as one or more	1.6123	47.18	−1.0136	−33.66	−0.9081	−17.11	−0.5488	−15.76
<i>(Base case: Transit pass ownership less than one per adult in the household)</i>								
Transit pass ownership one or more per adult in the household	−0.3582	−12.79	0.7115	26.88	−0.2557	−5.86	—	—
<i>(Base case: No worker household)</i>								
Single worker household	—	—	0.4793	14.27	—	—	0.2339	5.44
Multi-worker household	0.1170	5.05	0.5985	18.12	0.2250	4.60	0.3034	6.90
<i>(base case: dwelling type house or town house)</i>								
Apartment type dwelling	−0.1020	−4.72	0.1426	5.83	−0.3216	−7.21	—	—

TABLE 4.5 Exogenous Variable Effects on Modal Accessibilities

Variable definitions	Auto Accessibility		Transit Accessibility		Bicycle Accessibility		Walk Accessibility		Estimation results
	Estimate	z value	Estimate	z value	Estimate	z value	Estimate	z value	
<i>(Base: Age category 0-15 years)</i>									
Presence in age category 16-25 years	—	—	—	—	—	—	—	—	
Presence in age category 26-35 years	—	—	—	—	−0.1074	−5.25	—	—	
Presence in age category 36-45 years	—	—	—	—	−0.0886	−3.71	—	—	
Presence in age category 46-55 years	—	—	—	—	—	—	—	—	
Presence in age category 56-65 years	—	—	—	—	—	—	—	—	
Presence in age category 66 and above years	—	—	—	—	—	—	—	—	
Presence of students in the household	—	—	—	—	—	—	—	—	
Distance to CBD (km)	−0.0133	−7.91	−0.1852	−80.81	−0.0838	−47.01	0.0207	13.34	
Population Density (Standardized value)	0.2365	25.66	0.2489	24.69	—	—	—	—	
Employment Density (Standardized value)	−2.1572	−112.71	0.4428	44.02	0.1059	13.24	−0.1416	−15.58	
<i>(Base: Income less than 15,000\$)</i>									
Income 15,000\$-40,000\$	0.1358	5.51	−0.1255	−2.31	—	—	0.0970	4.08	
Income 40,000\$ to 60,000\$	—	—	−0.1140	−2.07	0.1042	3.39	—	—	
Income 60,000\$ to 100,000\$	—	—	−0.1124	−2.11	0.1244	4.49	—	—	
Income 100,000\$ to 125,000\$	—	—	−0.1576	−2.81	0.1408	4.36	—	—	
Income greater than 125,000\$	—	—	−0.0951	−1.76	0.2282	8.09	—	—	
<i>(Base: fraction of male adults in the household)</i>									
fraction of female adults in the household	—	—	—	—	0.1029	2.33	—	—	

TABLE 4.6 Exogenous Variable Effects on Modal Accessibilities (continued)

Variable definitions	Auto Accessibility		Transit Accessibility		Bicycle Accessibility		Walk Accessibility	
	Estimate	z value	Estimate	z value	Estimate	z value	Estimate	z value
<i>(Base case: Vehicle availability per licensed adult is zero)</i>								
Vehicle availability per licensed adult is between zero and one	−0.0866	−3.57	−0.0716	−2.45	—	—	—	—
Vehicle availability per licensed adults as one or more	−0.0794	−3.35	−0.1735	−5.95	—	—	—	—
<i>(Base case: Transit pass ownership less than one per adult in the household)</i>								
Transit pass ownership one or more per adult in the household	—	—	—	—	—	—	—	—
<i>(Base case: No worker household)</i>								
Single worker household	—	—	0.0531	2.39	—	—	—	—
Multi-worker household	—	—	—	—	—	—	—	—
<i>(base case: dwelling type house or town house)</i>								
Apartment type dwelling	−0.6051	−28.92	—	—	−0.1321	−6.84	−0.4568	−24.0

4.4.2 Endogenous variable effects and correlation effects

A key objective of this study is to unravel the complex interplay of relationships among the eight endogenous variables with a view to better understand the complementarity and substitution effects that may be present. Table 4.7 offers estimates of the endogenous variable effects, while table 4.8 offers goodness of fit together with error correlations. A graphical depiction of the significant endogenous variable effects retained in the final model specification is shown in Figure 4.4 (specification H4 with only significant relationships). Note that these effects are direct effects because the jointness among the eight endogenous variables has been captured through the error correlations.

There is a positive effect of higher auto accessibility (significant only at 90% confidence interval, as compared to very low auto access) on transit trips. Indeed, a higher accessibility to opportunities by driving may be indicative of an outcome of the neighbourhood's central location within the metropolitan area. This is an interesting finding which hasn't been reported earlier for lack of multi-modal multivariable model estimations. However, higher transit accessibility has a negative effect on auto trip frequencies and a positive effect on transit trip frequencies. The complementarity between auto accessibility and transit trips and the substitution effect observed between transit accessibility and auto trips may hint towards the fact that providing greater transit facilities with better connectivity and employment reach can positively impact the desired reduction in driving trips. Further, it also highlights the residential neighbourhood dissonance of transit riders and shows that location with higher transit accessibility (that is. better transit service) are not necessarily locations where people who make more transit trips reside. The lack of good transit service may be the reason that some people have to make more transit trips to reach their destinations. Very high transit accessibility also positively impacts walk trips. This may as well be the case, as residences existing along transit rich neighbourhoods may also have accessibility to other commercial developments nearby, thereby causing a higher number of walk trips.

We observe a complementary relationship between bicycle accessibility on transit trips. Higher bicycle accessibility has a positive impact on transit trips. This could indicate towards the fact that a bicycle friendly neighbourhood may also be transit rich, thereby leading to more transit trips. This, however, depends on the fact that the level of stress associated with biking may vary from location to location and certain neighbourhoods with much lower level of

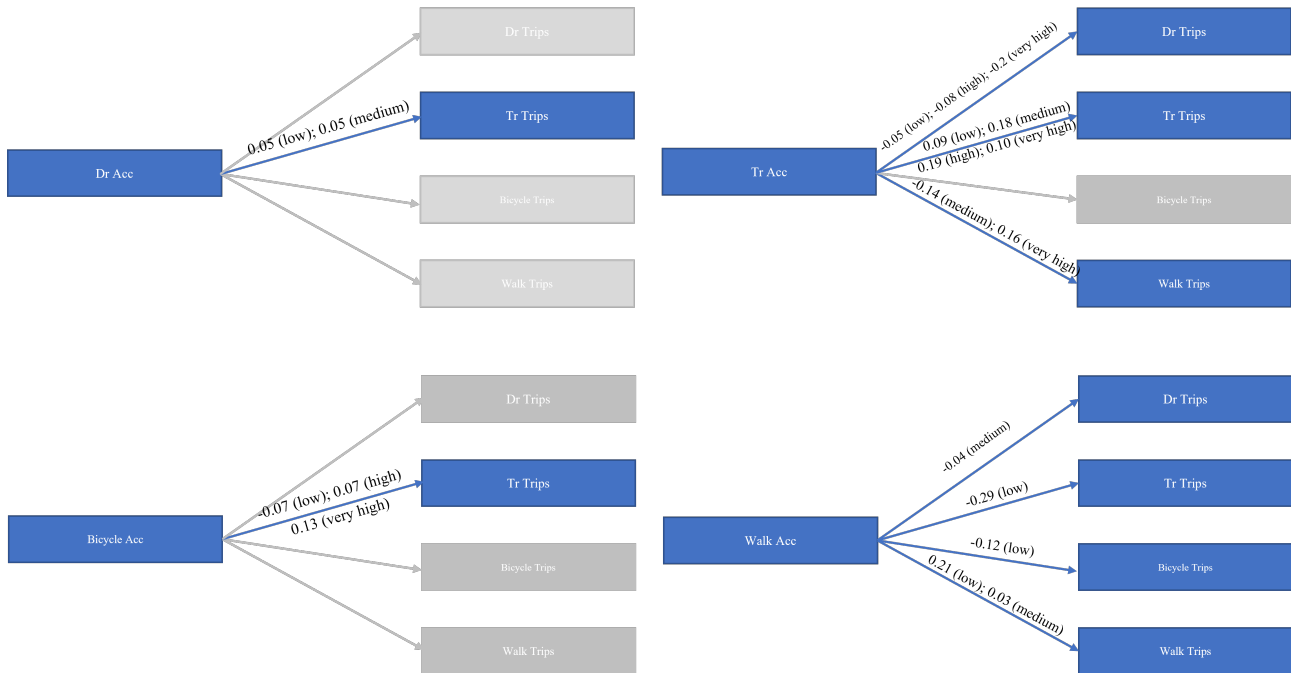


FIGURE 4.4 Endogenous Variables Effects (the highlighted blue are the significant effects, with values signifying the coefficients, and ‘low/medium/high/very high’ in parentheses signify the level of accessibility for which the effect is quantified)

stress from traffic are predominantly transit rich neighbourhoods. Similarly, we observe walk accessibility to have a negative impact on auto, transit, and bicycle trips. This is an interesting finding as it suggests that neighbourhoods with higher accessibility by walk, we can observe an increase in active trips. If the neighbourhood is only partially walk accessible (and by extension, non-motorized accessibility) then the traditional tendencies towards the use of vehicles cannot be curbed.

The error correlations in the bottom of the table show that several error covariances are statistically significant in this dataset. This is to be expected, as unobserved household socio-economic and demographic, as well as built environment attributes are likely to simultaneously affect trip frequencies and modal accessibilities. Mostly, all negative correlations between modal trip frequencies suggest that the unobserved attributes that motivate the use of any one mode choice, inhibit the use of all other mode types. The exception being the positive correlation between bicycle and walk trips, which shows that unobserved factors that support either of the active transport trips could also lead to similar other active modal preference. We observe a highly significant positive correlation between auto accessibility and walk access. However, a negative correlation between

transit-accessibility and walk-accessibility is observed, which could be the result of unobserved characteristics favouring one or the other. There are also a few significant correlations between access-trip combinations of endogenous variables. While the observed attributes resulted in a direct negative effect of transit accessibility on auto trips, the unobserved attributes lead to a positive correlation between the two. Also, the observed attributes resulted in a positive direct impact of auto accessibility on transit trips, the unobserved attributes lead to a negative correlation between the two. And the observed attributes resulted in a positive direct impact of transit accessibility on transit trips, the unobserved attributes lead to a negative correlation between the two. We observe a highly significantly negative correlation between auto and transit accessibilities, and a significantly positive correlation between auto and walk accessibilities. A positive correlation between transit and bicycle accessibilities could also be attributed to the fact that unobserved factors leading to better cycling facilities are predominantly transit rich neighbourhoods. However, a negative correlation between transit and walk accessibilities is observed, which could be the result of unobserved characteristics favouring one or the other. There are also a few significant correlations between access-trip combinations of endogenous variables. The observed attributes resulted in a positive direct impact of transit accessibility on transit trips, the unobserved attributes lead to a negative correlation between the two.

TABLE 4.7 Endogenous Variable Effects

Variable definitions	Auto Trips		Transit Trips		Bicycle Trips		Walk Trips	
	Estimate	z value	Estimate	z value	Estimate	z value	Estimate	z value
<i>(Base case: Drive Access Very Low)</i>								
Drive Access Low	—	—	0.0539	1.95	—	—	—	—
Drive Access Medium	—	—	0.0521	1.87	—	—	—	—
Drive Access High	—	—	—	—	—	—	—	—
Drive Access Very High	—	—	—	—	—	—	—	—
<i>(Base case: Transit Access Very Low)</i>								
Transit Access Low	−0.0510	−1.99	0.0967	2.60	—	—	—	—
Transit Access Medium	—	—	0.1826	4.91	—	—	−0.1469	−3.71
Transit Access High	−0.0832	−2.80	0.1913	4.63	—	—	—	—
Transit Access Very High	−0.2576	−7.14	0.1081	2.32	—	—	0.1610732	4.35
<i>(Base case: Bike Access Very Low)</i>								
Bike Access Low	—	—	−0.0682	−2.23	—	—	—	—
Bike Access Medium	—	—	—	—	—	—	—	—
Bike Access High	—	—	0.0742	2.56	—	—	—	—
Bike Access Very High	—	—	0.1351	4.56	—	—	—	—
<i>(Base case: Walk Access Very Low)</i>								
Walk Access Low	—	—	−0.1940	−6.79	−0.1271	−2.54	0.2120	6.24
Walk Access Medium	−0.0466	−1.99	—	—	—	—	0.0368	1.06
Walk Access High	—	—	—	—	—	—	—	—
Walk Access Very High	—	—	—	—	—	—	—	—

TABLE 4.8 Correlation Effects and Goodness-of-fit

Variable definitions	Auto Trips		Transit Trips		Bicycle Trips		Walk Trips	
	Estimate	z value	Estimate	z value	Estimate	z value	Estimate	z value
<i>Thresholds</i>								
Intercept 1 2	0.9848	19.53	1.0427	12.29	0.0686	0.85	0.4225	7.17
Intercept 2 3	1.7602	34.12	1.1854	13.97	0.1098	1.35	0.6582	11.10
Intercept 3 4	2.4563	46.93	2.1584	25.33	0.6315	7.59	1.3287	22.46
Intercept 4 5	3.1817	59.84	3.1570	35.97	1.1619	13.20	2.1287	33.62
<i>Correlation Terms</i>								
DriveTrips	1.0000		−0.2967	−28.19	−0.0884	−4.17	−0.1592	−11.89
TransitTrips			1.0000		−0.1775	−8.77	−0.0848	−6.56
BikeTrips					1.0000		0.0791	3.39
WalkTrips							1.0000	
Variable definitions	Auto Accessibility		Transit Accessibility		Bicycle Accessibility		Walk Accessibility	
	Estimate	z value	Estimate	z value	Estimate	z value	Estimate	z value
<i>Thresholds</i>								
Intercept 1 2	−1.2690	−41.13	−3.5503	−60.42	−1.7417	−42.45	−0.9365	−43.36
Intercept 2 3	−0.4760	−16.03	−2.6993	−46.74	−1.0938	−27.31	−0.2938	−13.24
Intercept 3 4	0.1679	5.79	−1.7775	−31.31	−0.4989	−12.55	0.2591	11.68
Intercept 4 5	0.8520	29.01	−0.6204	−11.25	0.1817	4.61	0.9063	39.07
<i>Correlation Terms</i>								
DriveTrips	−0.0166	−1.67	—	—	—	—	—	—
TransitTrips	—	—	−0.0545	−14.21	−0.0140	−2.45	—	—
BikeTrips	—	—	−0.0560	−2.62	—	—	—	—
WalkTrips	—	—	—	—	—	—	—	—
DriveAccess	1.0000		−0.0385	−3.62	—	—	0.7734	304.72
TransitAccess			1.0000		0.1463	14.87	−0.0841	−8.50
BikeAccess					1.0000		—	—
WalkAccess							1.0000	
Goodness-of-fit	observations		AIC		BIC		Likelihood (mean)	
	14,781		1774180.92		1781811.82		-886086.54	

4.5 CONCLUSIONS

The results of this chapter bolster the importance of a multivariate ordered probit modelling methodology, as there are significant error correlations that capture the effects of correlated unobserved attributes simultaneously affecting multiple trip frequency and accessibility outcomes. It is imperative that these error correlations be estimated so that consistent estimates of exogenous and endogenous variable effects can be obtained, and accurate forecasts of combinations of the eight endogenous variables may be made. The results confirm our hypothesis that not only does a strong association exist between employment accessibility and household trip frequency, but there are direct effects to be observed. Separate models were estimated for different household structures such as single adult with no kids, couple adult with no kids, single or couple with one or more kids, multi-adult with no kids, and multi-adult with one or more kids. The results section provided inferences about “couple adult with no kids” household type, which represents a household with two adults over the age of 16 years and zero child living together.

In addition to observed direct effects, there are significant unobserved effects (error correlations) that simultaneously affect multiple trip frequency and accessibility outcomes. Ignoring such effects would result in inconsistent estimates of exogenous and endogenous variable effects. The endogenous variable effects are of key interest in this study. The final model structure that provides the best and most intuitive results shows that there are both complementary and substitution effects at play. We observe a complementarity effect between auto accessibility and transit trips, and a substitution effect observed between transit accessibility and auto trips. Based on these results, policies oriented towards providing greater transit facilities with better connectivity and employment reach can positively impact the desired reduction in driving trips. Further, it also highlights the residential neighbourhood dissonance of transit riders and shows that locations with higher transit accessibility (that is, better transit service) are not necessarily locations where people who make more transit trips reside. People who are fairly transit dependent (e.g., low-income people and/or people without cars) need to take transit a lot, despite the lack of good transit access. For future transit-development, policymakers can utilize our results to identify the transit-poor neighbourhoods within Toronto and improve the transit facilities. Such policy measures can have a multifold advantage. A targeted transit development can provide a higher accessibility

to low-income neighbourhoods, but also might impact vehicular congestion on the streets of Toronto.

Further, the results demonstrate that on average, older households in the sample tend to live in suburban neighbourhoods that are auto-oriented, while younger households tend to live in more “urban” (aka downtown) walk/transit-oriented neighbourhoods. We also observe a preference to auto as one moves further away from the downtown. Suburbanization and its effects on variability of modal preferences is very well reflected in the results, such that driving is preferred to bicycle and walk modes due to the increasing distance between destinations and lack of cycling and walking infrastructure. The limiting bicycle infrastructure in Toronto is better placed for the locations with higher employment density, which does not serve the population at large and disincentivizes trip utility for higher population density neighbourhoods. We are also able to identify what are potentially time impoverished population groups, restricting the extent of activities that they can participate in and raising an equity issue regarding the distribution of transport infrastructure in Toronto.

Several policy recommendations can be inferred to address the complex dynamics of trip frequency, accessibility, and their implications for urban planning and transportation. These recommendations aim to optimize transportation systems, improve accessibility, and promote sustainable mobility choices while addressing the challenges posed by endogeneity:

1. **Integrated Transportation Planning:** Planners and policymakers should consider adopting integrated transportation planning approaches that account for the interrelationships between trip frequency, accessibility, and mode choice (among several other location choice and travel behaviour variables, as required). The multivariate ordered probit modelling methodology presented in this chapter provides a valuable framework to capture the simultaneous effects of various factors on travel behaviour. Developing joint models, similar to the one discussed, could enhance our understanding of these relationships and guide more effective policy interventions.
2. **Transit Development and Connectivity:** The study highlights the importance of enhancing transit facilities and connectivity, particularly in areas with limited transit access. Policies aimed at improving transit service quality, coverage, and connectivity can encourage a shift towards sustainable modes of transportation, reducing the reliance on private vehicles

and alleviating traffic congestion.

3. **Targeted Investment in Transit-Poor Neighbourhoods:** The findings indicate that transit-poor neighbourhoods exist within the study area. Policymakers can utilize the insights from the research to strategically target transit development and improvement initiatives in these underserved areas and marginalized communities.
4. **Promotion of Active Transportation:** Addressing the variability of modal preferences across different neighbourhoods requires a focus on promoting active transportation modes like cycling and walking. Investing in infrastructure such as dedicated bike lanes, pedestrian-friendly pathways, and safe crossings can encourage these sustainable modes, particularly in areas with higher population density and employment hubs.
5. **Equity Considerations:** The study highlights the presence of time-impoorished population groups and equity concerns related to the distribution of transport infrastructure. Planners should prioritize equitable access to transportation options for all segments of the population, ensuring that vulnerable and underserved communities have convenient and reliable transportation choices.
6. **Data-Informed Decision-Making:** Policymakers should leverage robust data sources and advanced modelling techniques to inform transportation-related policies. The methodologies used in this study demonstrate the potential for data-driven insights to guide evidence-based decision-making, resulting in more effective and targeted interventions.

Chapter 5

An instrumental variable model for addressing endogeneity in residential location choice

Abstract

Residential location choice models are vital for understanding location decisions and informing urban planning. This research develops a control function (CF) based discrete choice model, addressing endogeneity in residential location choice due to housing prices and travel behaviour. Our hypothesis highlights the prominent negative impact of housing prices by accounting for endogeneity, while also addressing the endogeneity with travel behaviour. Utilizing 2018-19 London Travel Demand Survey (LTDS) data, we comprehensively analyse household location choices and their relation to travel behaviour. The CF-MxL model shows superior fit statistics compared to the base MNL model, with no endogeneity correction. Findings underscore housing prices' negative influence on location choices, particularly affecting lower-income groups, as compared to highly subdued effect size of the incorrect MNL model. Additionally, vehicle miles travelled (VMT) significantly impacts location choices, with households having higher VMT showing aversion to postcodes with a higher population density. Furthermore, we observe a positive association of non-white households with locations infested with a higher crime density. This research offers insights into the interplay between residential location and travel behaviour, isolating and revealing the impact of

various postcode-attributes, socioeconomic and demographic characteristics. The rest of the chapter is presented with an introduction to the endogeneity in residential location choice models, particularly due to price and travel behaviour, followed by a brief overview of the methods employed to address these, description of the data, methodological framework, estimation results, discussion and conclusions.

5.1 INTRODUCTION

5.1.1 Background and motivation

Understanding the interactions between residential location and travel behaviour is essential for accurately predicting the spatial distribution and movement of population in cities (Ortuzar and Willumsen, 2011). Residential and commercial destinations serve as trip origins and destinations, requiring accessible transport services to support mobility. The relationship between residential choices and travel behaviour can be endogenous, where transport policies improving neighbourhood accessibility attract activity and lead to household relocations (Srinivasan and Ferreira, 2002; Tayyaran et al., 2003; Choocharukul et al., 2008; Tillema et al., 2010; Sierpiński, 2011; Hong and Goodchild, 2014). Conversely, the built environment's impact on travel behaviour is also widely studied (Singh et al., 2018). For instance, residents in suburban or rural neighbourhoods tend to make more car trips over longer distances, while urban dwellers utilize public transport, biking, and walking more frequently due to different built environment attributes and transport services.

The built environment of residential neighbourhoods influences travel attitudes, shaping travel behaviour (De Vos et al., 2018; Wang and Lin, 2019). Better transit connectivity in urban neighbourhoods may encourage public transport and non-motorized travel, while suburban areas can foster positive attitudes towards vehicle ownership. Residential location choices significantly affect travel satisfaction, with people biased towards their preferred transport mode (De Vos et al., 2016). Locations promoting preferred modes are perceived favourably, while those inhibiting them reduce satisfaction levels (Ginsberg and Churchman, 1984; Oh, 2003). Travel attitudes and satisfaction may change with residential environment changes. People's travel preferences impact residential choices, with vehicle owners opting for suburban areas with parking facilities and environmentally-conscious individuals choosing mixed-use neighbourhoods closer to city centres for better transit and active travel options.

The self-selection process in residential choices is frequently studied in travel behaviour research, as it mediates the built environment's effect on future travel behaviour (Næss, 2009; van Wee, 2009). However, explicit treatment is required to disentangle the influence of residential choices on travel behaviour and vice versa, enabling unbiased estimates and comprehensive understanding of the built environment's interplay with travel patterns. In other words, the endogenous nature of travel behaviour's impact on residential location must

be addressed for unbiased estimates in residential location choice models.

5.1.2 Methods to address endogeneity in residential location discrete choice models

Endogeneity in residential location choice models arises when any explanatory variable correlates with the error term due to omitted attributes, measurement or specification errors, or self-selection. Addressing endogeneity is important because it leads to inconsistent parameter estimates. These estimates can yield wrong policy implications, such as the effect of rental or purchase price can be grossly under-estimated (often with wrong sign) due to confounding effect of omitted attributes in a classic residential location choice model. Some travel behaviour attributes can be considered endogenous to residential location choice, such as transportation mode choice influenced by proximity to different modes of transportation. People who live close to a tube station may prefer the tube, while people who live farther away from a station may prefer to drive. Other attributes endogenous to residential location choice may include vehicle miles travelled, trip frequency, travel distance and travel time. The lack of solutions for endogeneity issues is evident in travel behaviour studies. There are several studies (such as Cheng et al., 2019; Farinloye et al., 2019; Haque et al., 2019; Thronicker and Klinger, 2019; Zarabi et al., 2019) that mainly focus on establishing and exploring unidirectional effects, while disregarding any possibility of an *immediate or delayed feedback* from endogenous variables in the model.

Dealing with endogeneity in discrete choice models (DCMs) differs from linear models, as it involves changing the scale of estimated models (Guevara and Ben-Akiva, 2012). Then another level of challenge arises in dealing with endogeneity in location choice DCMs (such as residential location choice models) due to the issues associated with spatial location choice, which introduce additional complexities and considerations. These models not only need to account for endogeneity in the explanatory variables, but they must also address spatial autocorrelation, heterogeneity in spatial preferences, and potential spatial spillover effects. Spatial econometric techniques and specialized instrumental variables become crucial in mitigating these challenges and ensuring that the estimated effects of variables, such as travel behaviour or alternative-specific attributes, accurately capture the underlying relationships within the context of location choice. The control function (CF) approach, commonly used to

address endogeneity in discrete choice models (DCM), is a two-stage procedure. Other methods to address endogeneity in logit-based discrete choice models include using Proxys (PR), instrumental variables (IV) methods, the Multiple Indicator Solution (MIS), and integrating latent variables (Guevara, 2015).

While there have been plenty of studies focussed on dealing with self-selection, no analysis explicitly corrects the model parameters in residential location discrete choice model for endogeneity due to travel behaviour. Additionally, in recent progress for cross-sectional travel behaviour analysis using logit-based discrete choice models, addressing the correlation between alternatives and endogeneity due to unobserved variables are simultaneous issues. However, these two issues are rarely tackled together, as the logit-based model structure can be complex. The mixed logit model has been shown to accommodate instrumental variable (IV) methods while describing correlations (Xu et al., 2017). On the other hand, a probit-based modelling approach (Bhat and Sidharthan, 2011; Bhat et al., 2014) can handle correlation between alternatives and simultaneous endogeneity, but estimating this model can be challenging.

5.1.3 Contributions of this chapter

Residential location choice models aim to quantify the influence of various factors, such as real estate and rent prices, dwelling attributes, job accessibility, education, healthcare services, and demographic attributes, on residential decisions (Waddell, 2009; Hurtubia et al., 2011; Farooq and Miller, 2012). This study contributes to the field by developing a control function (CF) based discrete choice model that addresses the endogeneity in residential location choice due to housing prices and travel behaviour simultaneously. Our hypothesis suggests that housing prices retain primary significance even after accounting for endogeneity and variation, but the impact of travel behaviour cannot be ignored. Addressing the endogeneity of both housing prices and travel behaviour is crucial for several reasons. Firstly, it enables an accurate capture of the bidirectional relationship between residential location choice and travel behaviour, providing a comprehensive understanding of their interplay. Secondly, simultaneous endogeneity correction yields unbiased estimates for both factors, ensuring reliable policy implications and decision-making processes. Thirdly, this comprehensive approach allows us to disentangle the unique and combined effects of housing prices and travel behaviour on residential location choice, shedding

light on their relative significance.

The proposed CF method utilizes instrumental variables correlated with the endogenous variables but not the error term, effectively handling endogeneity due to both price and vehicle miles travelled (VMT) in residential location choice. The data for this analysis is sampled from 2018-19 London Travel Demand Survey (LTDS). We model the residential location choice at the (disaggregate) outer postcode level. The endogeneity due to housing price at outer postcode level is addressed using instruments (*aggregate*) *housing price* and *public transport accessibility level (PTAL)* at the (aggregate) *borough level*. By using the price information at the more aggregate level, we ensure exogeneity and strength of the instrument. Similarly, the endogeneity due to household VMT is addressed using *average public transport accessibility level (PTAL) at the outer postcode level* for that household and the *household vehicle ownership*. PTAL is a measure of public transport accessibility that takes into account the frequency, reliability, and distance of public transport services, as well as the number of interchange opportunities. The results of models with and without correction for endogeneity are comparatively evaluated.

5.2 DATA DESCRIPTION

The data used for this study is a sample taken from the 2018-2019 London Travel Demand Survey (LTDS). The LTDS is a household travel survey collecting travel information from approximately 8,000 households in the London metropolitan region. The LTDS collects individual and household socio-economic and demographic information in addition to one day travel diary data (for each person above the age of five) regarding the trips made in a day, such as travel mode, start time, etc. After removing the households with incomplete and inconsistent attributes, the remaining 4,866 households residing in the London were employed from the data. These households reside in the 168 outer postcodes¹ of London. The dependent variable of interest in this analysis is the residential location choice at the outer postcode level. Data preparation, including the incorporation of additional housing price and PTAL information, was performed similarly to the process outlined in Chapter 3, as described in Section 3.3.

Table 5.1 provides an overview of the descriptive characteristics of the alternative-specific

¹After cleaning for missing information and data with fewer than 10 observations for any specific outer postcodes, we were left with 168 boroughs in London, out of a total of usual 272 boroughs which form the London metropolitan region.

attributes: Natural-log of estimated price of a 2-bedroom, Natural-log of population density (per squared-km), crime incidents density (per squared-km), and PTAL average values. The independent characteristics of the households and individuals follow the same description as described earlier in Table 3.1 of chapter 3. Table 5.2 provides the frequency distribution of the 168 outer postcodes of London as available in the data. These outer postcodes constitute the lower-level disaggregate choice decision. Figure 5.1 shows a visual map of London's outer postcodes within the greater London metropolitan region area.

TABLE 5.1 Description of Sample (n=4,866)

Dependent variables characteristics distribution				
Natural-log of Outer Postcode estimated price of a 2-bedroom	12.70	0.37	12.11	14.20
Natural-log of Outer Postcode population density (per km^2)	14.16	13.91	8.24	15.52
Outer Postcode crime incidents density (per km^2)	8.30	6.17	0.04	33.71
Outer Postcode PTAL average value	11.50	6.64	3.69	38.54

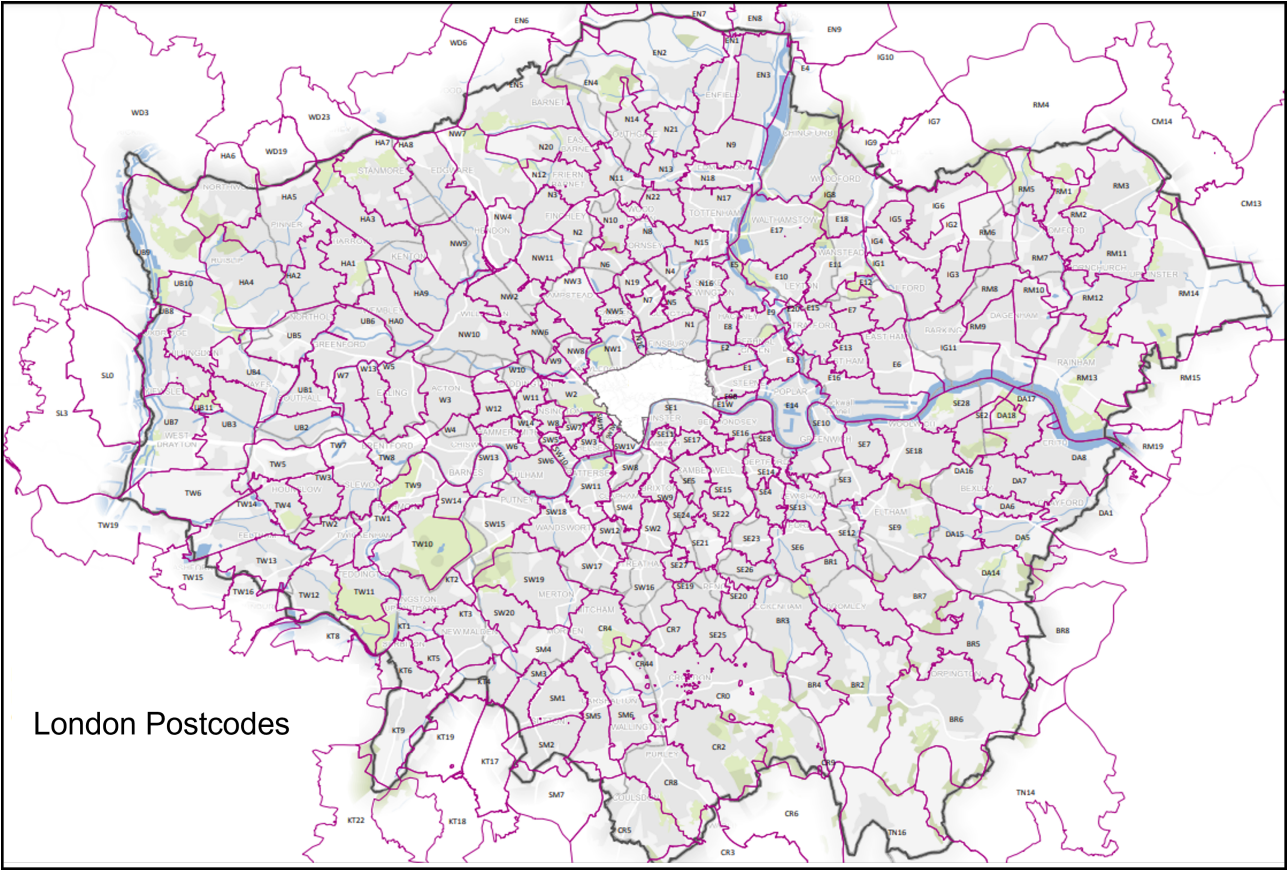


FIGURE 5.1 London Postcodes map

TABLE 5.2 Outer postcode choice frequency distribution (n=4,866)

Outer PC	Freq.	%	Outer PC	Freq.	%	Outer PC	Freq.	%	Outer PC	Freq.	%
BR1	35	0.72	HA0	29	0.60	NW7	19	0.39	SW1V	26	0.53
BR2	37	0.76	HA2	27	0.55	NW9	23	0.47	SW2	47	0.97
BR3	41	0.84	HA3	17	0.35	RM1	18	0.37	SW20	34	0.70
BR4	24	0.49	HA5	24	0.49	RM10	45	0.92	SW3	30	0.62
BR5	55	1.13	HA6	14	0.29	RM11	23	0.47	SW4	39	0.80
BR6	46	0.95	HA7	29	0.60	RM12	24	0.49	SW5	20	0.41
BR7	24	0.49	HA8	37	0.76	RM13	28	0.58	SW6	70	1.44
CR0	118	2.42	HA9	28	0.58	RM14	22	0.45	SW7	18	0.37
CR2	54	1.11	IG1	44	0.90	RM2	12	0.25	SW8	23	0.47
CR4	52	1.07	IG11	29	0.60	RM3	27	0.55	SW9	39	0.80
CR5	19	0.39	IG2	13	0.27	RM6	18	0.37	TW1	18	0.37
CR7	30	0.62	IG3	21	0.43	RM8	27	0.55	TW10	26	0.53
CR8	24	0.49	IG6	15	0.31	RM9	44	0.90	TW11	23	0.47
DA1	11	0.23	IG8	20	0.41	SE10	14	0.29	TW12	20	0.41
DA14	20	0.41	KT1	30	0.62	SE11	22	0.45	TW13	22	0.45
DA15	29	0.60	KT2	22	0.45	SE13	24	0.49	TW14	22	0.45
DA16	29	0.60	KT3	23	0.47	SE14	29	0.60	TW2	26	0.53
DA17	16	0.33	KT5	21	0.43	SE17	20	0.41	TW3	28	0.58
DA5	12	0.25	KT6	31	0.64	SE18	59	1.21	TW4	14	0.29
DA7	38	0.78	KT9	12	0.25	SE20	29	0.60	TW5	23	0.47
DA8	23	0.47	N1	20	0.41	SE22	28	0.58	TW7	28	0.58
E10	24	0.49	N11	12	0.25	SE23	29	0.60	TW8	15	0.31
E11	12	0.25	N12	23	0.47	SE24	10	0.21	TW9	21	0.43
E12	18	0.37	N13	23	0.47	SE25	27	0.55	UB1	30	0.62
E13	26	0.53	N15	35	0.72	SE27	14	0.29	UB10	33	0.68
E14	39	0.80	N16	40	0.82	SE3	24	0.49	UB2	16	0.33
E15	26	0.53	N17	36	0.74	SE4	20	0.41	UB3	31	0.64
E16	24	0.49	N18	29	0.60	SE5	11	0.23	UB4	29	0.60
E17	64	1.32	N2	23	0.47	SE6	39	0.80	UB5	40	0.82
E18	16	0.33	N20	19	0.39	SE7	23	0.47	UB6	50	1.03
E3	31	0.64	N21	26	0.53	SM1	31	0.64	UB7	22	0.45
E4	54	1.11	N3	25	0.51	SM2	18	0.37	UB8	37	0.76
E5	40	0.82	N4	14	0.29	SM3	26	0.53	W10	25	0.51
E6	44	0.90	N8	45	0.92	SM4	24	0.49	W12	56	1.15
E7	30	0.62	N9	42	0.86	SM5	38	0.78	W13	33	0.68
E8	24	0.49	NW1	34	0.70	SM6	28	0.58	W14	42	0.86
E9	25	0.51	NW10	61	1.25	SW11	49	1.01	W3	41	0.84
EN1	47	0.97	NW11	25	0.51	SW12	11	0.23	W4	13	0.27
EN2	30	0.62	NW2	22	0.45	SW13	16	0.33	W5	51	1.05
EN3	39	0.80	NW4	31	0.64	SW14	15	0.31	W6	30	0.62
EN4	25	0.51	NW5	19	0.39	SW16	13	0.27	W7	36	0.74
EN5	28	0.58	NW6	24	0.49	SW19	66	1.36	W8	15	0.31
									Total	4866	100.00

5.3 METHODOLOGICAL FRAMEWORK

In order to estimate the observed choice decision, we employ a discrete choice model (Multinomial Logit and Mixed Logit).

5.3.1 Discrete choice model accommodating Control Function (CF)

We will briefly review the Mixed Logit model (**MxL**). We extend the MNL model by allowing for heterogeneity in preferences across individuals. We accommodate the heterogeneity that the coefficients of the explanatory variables are randomly drawn from a distribution, allowing for the estimation of individual-specific coefficients, as well as allowing for correlations across alternatives. By applying the CF method, we handle the endogenous variable in residential location choice, namely, natural-log of outer postcode estimated price of a 2-bedroom and household vehicle miles travelled (VMT). The CF method in this study requires at least one instrumental variable that is correlated with the endogenous variable but not correlated with the error term of the residential location choice (Petrin and Train, 2003; 2010). The endogeneity due to housing price at outer postcode level is addressed using instruments *housing price and public transport accessibility level (PTAL) at the (aggregate) borough level*. By using the price information at the more aggregate level, we ensure exogeneity and strength of the instrument. Similarly, the endogeneity due to household VMT is addressed using *average public transport accessibility level (PTAL) at the outer postcode level and the household vehicle ownership*. Using household vehicle ownership as an instrument for vehicle miles travelled (VMT) could potentially be both appropriate and insightful, but it also comes with some caveats. On one hand, household vehicle ownership can be a relevant IV if it satisfies the two key criteria for a good instrument: relevance and exogeneity. If household vehicle ownership is strongly correlated with VMT (relevance) and is not directly correlated with the error term in the model (exogeneity), it could provide a valid instrument to help address endogeneity. On the other hand, there could be a potential endogeneity between household location and vehicle ownership. This is a concern because if the IV is correlated with unobservable factors that affect both VMT and location choice, it could violate the exogeneity assumption of instrumental variables, leading to biased estimates. Therefore, we check for validity of instruments using the tests described later in section 5.3.3.

Let's consider a residential location model with endogeneity due to travel behaviour

variable q . Assuming the true linear utility as:

$$U_{ij} = \beta_x x_{ij} + \beta_q q_{ij} + e_{ij} \quad (5.1)$$

where, β_q and β_x are parameters to be estimated, x_{ij} and q_{ij} are explanatory variables of the model with e_{ij} as the exogenous error term. The variable q_{ij} may be endogenous either due to omission or correlation with the error term. Now, let us consider that the k^{th} term of x_{ij} is correlated with q_{ij} as follows:

$$x_{kij} = \gamma_0 + \gamma_z z_{ij} + \gamma_q q_{ij} + \phi_{ij} \quad (5.2)$$

where ϕ_{ij} is the exogenous error term, z_{ij} is the exogenous attribute working as an instrument and independent from e_{ij} . For the model to be identifiable, there must be (at least) as many IV as endogenous variables in the model (Guevara and Ben-Akiva 2012). Following the assumption that variable q_{ij} may be endogenous, the specification needed to treat endogeneity would be:

$$x_{kij} = \gamma_0 + \gamma_z z_{ij} + \theta_{ij} \quad (5.3)$$

where the error term θ_{ij} contains both ϕ_{ij} and q_{ij} . If valid, such that z_{ij} is truly exogenous, then θ_{ij} will capture the entire part of x_{kij} that is endogenous. Therefore, residential location discrete choice model corrected using CF approach would have utility:

$$U_{ij} = \hat{\beta}_x x_{ij} + \beta_{\theta} \hat{\theta}_{ij} + \tilde{e}_{ij} \quad (5.4)$$

In practice, CF follows two stages:

1. Obtaining residuals $\hat{\theta}_{ij}$ by applying Ordinary Least Squares (OLS) regression to x_{kij} on z_{ij}
2. To estimate discrete choice model considering $\hat{\theta}_{ij}$ and x_{ij} attributes

We may obtain consistent estimators $\hat{\beta}_x$ for the β_x up to a scale (Guevara and Ben-Akiva 2012). There could also be a loss of efficiency due to two-stage estimation, unless the error terms are homoscedastic and non-autocorrelated (Rivers and Vuong, 1988). We may also face the

issue that the standard errors cannot be derived from the information matrix, hence requiring bootstrapping in the two-stage process. However, "CF method is robust to misspecifications of the distributional assumptions of the model" as discussed by Guevara (2015).

Empirically, we employ a mixed logit model, by rewriting x_{ij} into $x1_{ij}$ and $x2_{ij}$ from equation (9):

$$U_{ij} = \hat{\beta}_{x1}x1_{ij} + \hat{\beta}_{x2}x2_{ij} + \beta_{\theta}\hat{\theta}_{ij} + \tilde{e}_{ij} \quad (5.5)$$

where $\hat{\beta}_{x1}$ is a vector of alternative-specific attributes' coefficients for individual i and alternative j (which is allowed to be generalized as per equation below), $x1_{ij}$ is a vector of alternative-specific attributes for alternative j , and ϵ_{ij} is the error term. $\hat{\beta}_{x2}$ is a vector of socio-economic, demographic and travel behaviour attributes' coefficients for individual i (assumed homogeneous across population), $x2_{ij}$ is a vector of these attributes for alternative j . The coefficients $\hat{\beta} = (\hat{\beta}_{x1}, \hat{\beta}_{x2})$ are assumed to follow a normal distribution with mean β and variance-covariance matrix Σ , so that:

$$\hat{\beta} \sim N(\beta, \Sigma)$$

To capture the effects of alternative-specific attributes, such as price (variable with endogeneity), population density, and crime density, we include these variables in the vector of explanatory variables $x1_{ij}$. The specific form of the vector $x1_{ij}$ will depend on the chosen set of attributes and their functional form. For example, if we have price (P_{ij}), population density (D_{ij}), and crime density (C_{ij}) as attributes, the vector $x1_{ij}$ can be defined as:

$$x1_{ij} = [P_{ij}, D_{ij}, C_{ij}]$$

Furthermore, to incorporate the correlations between alternatives belonging to different transport zones (1 to 6), we introduce zone correlation in the error term ϵ_{ij} . We can include six zone correlation terms, denoted as *zone_corr1* to *zone_corr6*, to capture the correlation between alternatives within each zone. The inclusion of these zone correlation terms allows us to estimate the correlations among alternatives within their respective zones. We estimate the parameters of the mixed logit model using maximum likelihood estimation, which involves integrating over the distribution of the coefficients for each household (decision maker). We use the software package Apollo in R for this purpose. The mixed logit model allows us to

relax the assumption of homogeneous preferences, and to estimate the variance of the coefficients across individuals. This is particularly useful in situations where we expect that preferences may vary systematically across individuals, or where we have limited information on the distribution of preferences in the population.

5.3.2 Challenges

While the CF method provides a framework for addressing endogeneity, it also poses certain challenges that need to be considered:

1. Identification: Ensuring the identification of the control functions and instrumental variables is crucial. The control functions should capture the relationship between the endogenous variable and the instrumental variables without introducing biases or omitted variable problems.
2. Validity of instruments: The instrumental variables used in the CF method should be valid instruments. They should be correlated with the endogenous variable (*strong*) but not directly related to the error term of the model (*exogenous*). Testing the validity of instruments is an important step in the CF method.

5.3.3 Testing Validity of Instruments

To ensure the validity of instruments used in the CF method, two over-identification tests can be employed: the SREF (Refutability Test) and its variation, the SmREF (Modified Refutability Test) (Guevara, 2018).

Refutability Test (SREF)

This test follows in two stages:

1. Estimate the reduced form equation for x_{in} by OLS to obtain the residuals $\hat{\theta}$,
2. Estimate the discrete choice model considering $\hat{\theta}_{in}$ and the x_{in} attributes, but also the instrument z_{in} as an additional variable within the utility function to obtain the log-likelihood $l(\theta^{CF-Z})$:

$$U_{in} = \hat{\beta}_x x_{in} + \beta_{\hat{\theta}} \hat{\theta}_{in} + \beta_z z_{in} + \tilde{e}_{in} \quad (5.6)$$

In applications with more than one instruments, this process must be repeated for each instrument by adding it to the utility function and obtaining a log-likelihood. Also, in case of multiple instruments the Modified Refutability Test (S_{mREF}) can also be performed. The first stage of (S_{mREF}) is same as that of (S_{REF}), however in the second stage all the instruments are added at once within the utility function to obtain the log-likelihood $l(\theta^{CF-Z1-Z2})$. The statistics of the *Refutability Test* (S_{REF}) and its modified version (S_{mREF}) are the following:

$$S_{REF}^{Fixing-z} = -2(l(\theta^{CF}) - l(\theta^{CF-Z})) \sim \chi_r^2 \quad (5.7)$$

$$S_{mREF}^{Fixing-z} = -2(l(\theta^{CF}) - l(\theta^{CF-Z1-Z2})) \sim \chi_r^2 \quad (5.8)$$

where $l(\theta^{CF})$ is the log-likelihood of the corrected model in equation (5.4) and χ_r^2 is the value of the chi-squared distribution with degrees of freedom (r) equal to the over-identification of the model.

5.4 RESULTS AND DISCUSSION

The dependent variable of interest in this analysis is the residential location choice at the disaggregate level of Outer postcodes, with 168 alternatives. We utilise the socioeconomic and demographic characteristics of the households and the alternative specific attributes as independent variables for the choice model. In order to make the Mixed Logit (MxL) estimation faster, we used multinomial logit (MNL) model results as starting values for coefficients.

5.4.1 Description of model implementation and specification search

While the section 5.3 "methodological framework" presents the mathematical formulation of the mixed logit CF method, the practical application of the same follows as below:

1. Step 1: We develop the best case MNL model with all significant parameters, including interactions between alternative-specific attributes and household characteristics. This model, not corrected for endogeneity, will provide us as the base case to compare all further models against.
2. Step 2: We then develop an MNL model addressing endogeneity in residential location

choice due to *natural log of price of a 2-bedroom apartment at outer postcode level* using the instruments *natural log of price of a 2-bedroom apartment at borough level* and *public transport accessibility level (PTAL) at the borough level*. This model is validated using the (modified) refutability tests. This is done in two-steps, first a regression model is estimated with outcome variable *natural log of price of a 2-bedroom apartment at outer postcode level* and regressors as instruments and all other exogenous variables available in the data. The resulting residual of this regression model is included as an exogenous variable in the choice model as a *control function*, along with the price variable and all other exogenous variables from the data.

3. Step 3: We then develop an MNL model addressing endogeneity in residential location choice due to *natural log of vehicle miles travelled* using the instruments *natural log of public transport accessibility level (PTAL) at the outer postcode level* and *natural log of household vehicle ownership*. This model is validated using the (modified) refutability tests. This is done in two-steps, first a regression model is estimated with outcome variable *natural log of vehicle miles travelled* and regressors as instruments and all other exogenous variables available in the data. The resulting residual of this regression model is included as an exogenous variable in the choice model as a *control function*, along with the price variable and all other exogenous variables from the data.
4. Step 4: Finally, we develop a mixed logit (CF-MxL) model addressing endogeneity due to both *natural log of price of a 2-bedroom apartment at outer postcode level* and *natural log of vehicle miles travelled*, using the previously defined instruments. This is done in two-steps, first two regression models are estimated with outcome variables *natural log of price of a 2-bedroom apartment at outer postcode level* and *natural log of vehicle miles travelled*, respectively, and respective instruments as regressors along with all other exogenous variables available in the data. The resulting residuals of this regression model are included as exogenous variables in the choice model as *control functions*, along with the endogenous variables and other exogenous variables from the data. The Mixed Logit model was developed in steps where first all parameters are assumed to be homogeneous (non-random), then coefficients for price, PTAL, population density and crime density variables are allowed to be random parameters. We attempted a more

nuanced approach of looking at correlations between alternatives, where the alternatives belonging to a particular Transport Fare Zone (1 to 6) were allowed a specific fixed-correlation with one-another, while being independent of all other alternatives. This reduced the number of correlations to be modelled, while offering insight into the dependency between alternatives.

The model results shown below reflect this approach to modelling dependency between alternatives. The model performance is evaluated using Bayesian Information Criterion (BIC). For a given empirical model, $BIC = K \ln(Q) - 2 \ln(L)$ where K is the number of parameters, Q is the number of observations and $\ln(L)$ is the log-likelihood value at convergence. The model with the lowest value of BIC is preferred.

5.4.2 Model fit and validity of instruments

The final log-likelihood (AIC/BIC) value of the corrected CF-MxL model is -10821 (21682.01/21811.81), while the corresponding values for the incorrect MNL model are -10874.16 (21772.31/21850.19). From a strict model improvement fit statistics, the corrected CF-MxL performs better, and the coefficients are intuitive. The results confirm our hypothesis concerning endogeneity in residential location choice models due to alternative-attribute price and household travel behaviour. Moreover, the CF-MxL allows us to capture behaviourally richer contributing factors. The validity of the instruments was verified using the over identification tests for the exogeneity of the instruments in DCM proposed by Guevara (2018). In this case, $l(\theta^{CF})$ is the log-likelihood of the corrected model in equation 5.4 ($= -10821$) and χ_r^2 is the value of the chi-squared distribution with degrees of freedom (r) equal to the over-identification of the model. As we have four instruments and two endogenous variables, $r = 2$, and $\chi_r^2 = 5.99$.

$$S_{mREF}^{Fixing-zall} = -2(-10821 + 10818.67) = 4.66 < \chi_r^2 = 5.99 \quad (5.9)$$

Therefore, we can conclude that all our instruments are valid. To the best of our knowledge, these instruments had not been suggested before to correct for endogeneity due to price and travel behaviour in modelling residential location choice.

TABLE 5.3 First stage housing price regression model

Variable definitions	Estimate	t-stat
(Intercept)	5.790937	20.022
log(postcode population density)	0.638474	47.971
log(postcode crime density)	-0.578153	-40.169
log(number of GPs)	-0.06407	-8.218
Household income >£50,000 and <£100,000	—	—
Household income >£100,000	0.048583	3.732
Ethnicity = non-white	0.008784	0.337
log(PTAL at Borough level)	0.04093	4.052
log(price of a 2-bedroom apartment at Borough level)	-0.054943	-2.919
<i>(Interaction: log(crime density (per km²)) *)</i>		
<i>(Base: Ethnicity = white)</i>		
Ethnicity = non-white	-0.037058	-2.971

Notes: "—" signifies statistically not significant; log of variables denote the natural logarithm to the base 'e' = 2.7182.

TABLE 5.4 First stage vehicle miles travelled regression model

Variable definitions	Estimate	t-stat
(Intercept)	2.37088	3.102
log(postcode population density)	-0.13515	-1.969
log(postcode crime density)	—	—
log(number of GPs)	—	—
Household income >£50,000 and <£100,000	0.19284	4.035
Household income >£100,000	0.11123	1.893
Ethnicity = non-white	—	—
log(PTAL at outer postcode level)	-0.23112	-4.44
log(HH Vehicle ownership + 1)	1.84026	39.83
<i>(Interaction: log(crime density (per km²)) *)</i>		
<i>(Base: Ethnicity = white)</i>		
Ethnicity = non-white	—	—

Notes: "—" signifies statistically not significant; log of variables denote the natural logarithm to the base 'e' = 2.7182.

5.4.3 Comparative analysis of corrected and endogenous parameters outcomes

Here we describe and compare the models from Step 1 (base MNL) and Step 4 (CF-MxL) from the subsection 5.4.1 Description of model implementation and specification search. For CF-MxL model results, tables 5.3 and 5.4 present the first-stage regression models for the two endogenous variables, resulting in estimated residuals which were employed in the discrete choice model as control functions. Table 5.5 provides the estimation results depicting the influence of the exogenous variables on the residential location choice, and the right half of the table also presents

TABLE 5.5 Control function choice model result compared with simple MNL model

Variable definitions	CF Mixed logit Estimate	t-stat	Multinomial Logit Estimate	t-stat
<i>(Alternative-specific attributes)</i>				
log(price of a 2-bedroom apartment)	-0.1627	-2.96	-0.1523	-2.94
log(number of GPs)	0.1491	5.11	0.1695	5.98
log(population density)	0.3458	4.93	0.3323	4.89
log(crime density)	-0.0124	-0.17	0.0082	0.11
log(vehicle miles travelled)	4.8106	11.15	4.2170	11.37
<i>(Interaction: log(price of a 2-bedroom apartment) *)</i> <i>(Base: Household income < £50,000)</i>				
Household income > £50,000 and < £100,000	0.0731	1.05	0.0427	0.69
Household income > £100,000	0.2097	2.83	0.1389	2.04
<i>(Interaction: log(population density (per km²)) *)</i> <i>(Base: Household income < £50,000)</i>				
Household income > £50,000 and < £100,000	0.0074	0.16	-0.0040	-0.09
Household income > £100,000	-0.1609	-2.90	-0.1706	-3.16
<i>(Interaction: log(crime density (per km²)) *)</i> <i>(Base: Ethnicity = white)</i>				
Ethnicity = non-white	0.4301	9.26	0.3927	8.75
<i>(Interaction: log(vehicle miles travelled) *)</i>				
log(population density (per km ²))	-0.3077	-10.80	-0.2981	-10.74
log(crime density (per km ²))	0.1807	5.73	0.1697	5.51
Residual - price	-4.6056	-8.31	—	—
Residual - VMT	-0.8250	-2.44	—	—
Correlation Zone 1	-0.8538	-6.51	—	—
Correlation Zone 2	0.6713	4.63	—	—
Correlation Zone 3	—	—	—	—
Correlation Zone 4	—	—	—	—
Correlation Zone 5	-0.4421	-2.23	—	—
Correlation Zone 6	—	—	—	—

Notes: "—" signifies statistically not significant; log of variables denote the natural logarithm to the base 'e' = 2.7182.

comparative base MNL model results. Given the empirical context of residential location choice modelling, the alternative-specific constants have been suppressed. Also, the household socio-economic and demographic factors have been interacted with alternative-specific attributes to allow for their behavioural interpretation.

Our estimated model consists of alternative-specific attributes, as well as their interaction terms with socio-economic and demographic variables. Therefore, their

interpretation may not be straightforward. Average price of a 2-bedroom apartment variable has an independent coefficient (with endogeneity correction), residual coefficient and interaction with levels of household income. Based on coefficients from Table 5.3, we find that a higher natural log of average price of a 2-bedroom apartment negatively impacts (effect size = $(-0.1627) + (-4.6056) = -4.7683$) the probability of choice of the borough, for households with income less than £50,000. As we move on to a higher income group between £50,000 and £100,000, the impact of natural log of average price on probability of choice of borough (effect size = $(-0.1627) + (-4.6056) + (0.0731) = -4.6952$), and with the highest income group of income greater than £100,000, we have the impact of natural log of average price on probability of choice of borough (effect size = $(-0.1627) + (-4.6056) + (0.2097) = -4.5586$). This shows that while the impact of housing prices has a negative impact on probability of location choice, the higher income groups have marginally less aversion to higher prices. Comparatively, the MNL model (with no endogeneity correction) estimates the effect of natural log of the average price of a 2-bedroom apartment on choice of borough as negative, with effect size = -0.1523 for households with income less than £50,000, -0.1096 for income group between £50,000 and £100,000, and -0.0134. MNL model estimates that while lower income groups have a small negative association with price, it is approximately zero for the highest income group. Clearly, endogeneity due to price severely underestimates the impact of price on residential location choice, and the effect of unobserved factors such as sense of safety, comfort, housing dwelling attributes can be substituted by the instruments used in this estimation.

As the natural log of household vehicle miles travelled is not an alternative-specific attribute, we cannot infer its impact on location choice alternatives by itself. The effect size of the coefficient (with endogeneity correction) is 4.8106, and the residual effect size is -0.8250, which shows that while the corrected attribute has a positive utility for an alternative, the residual attribute has a negative association with the utility of an alternative. However, using the interaction of log of VMT with log of population density (alternative-specific attribute), we can infer the marginal impact of VMT mediated by population densities of the choice alternatives. The negative coefficient may be inferred as the negative association between households with higher VMT and location alternatives with a higher population density. While the literature documents the impact of a higher population density generally deterring

vehicle mode choice, our results also indicate that households with a higher VMT may be deterred by locations with a higher population density. The locations (postcodes) with a higher population density in London are usually closer to city centre, therefore, have fewer on-street and off-street parking facilities, and a higher public transport accessibility. These unobserved factors may contribute to a reduction in vehicle ownership, while those with a higher vehicle ownership may prefer suburban postcodes farther away from the centre of the city. The interaction of log of VMT with log of crime density (alternative-specific attribute) is positive, inferring that households with a higher VMT may prefer locations which are also attributed with relatively higher crime events, as compared to those with lower VMT. This result may indicate a mediating effect of unobserved factors on preference of locations with higher crime density, and thus requires further investigation in future.

The natural log of population density has an independent effect coefficient and interaction term with household income and vehicle miles travelled. Based on coefficients, for households with income less than £50,000, a higher population density positively impacts the choice of residential location (effect size = 0.3458). For households with income between £50,000 and £100,000, a higher population density positively impacts the choice of residential location (effect size = $0.3458 + 0.0074 = 0.3532$), which shows that the preference for areas with higher population density is slightly more pronounced. For households with income greater than £100,000, a higher population density positively impacts the choice of residential location (effect size = $0.3458 + (-0.1609) = 0.1849$), which shows that the preference for areas with higher population density is weaker among the highest income group households as compared to the lower income groups.

The natural log of crime density has an independent effect coefficient and an interaction term with ethnicity dummy (white vs. non-white). For white households, the effect of log of crime density on choice of residential location is negative (effect size = -0.0124), while for non-white households, the effect of crime-density on choice of residential location is positive (effect size = $(-0.0124) + 0.4301 = 0.4177$).

We also observe the Zone correlation coefficients for Zone 1 to 6. While the coefficients for zones 3, 4 and 6 are statistically insignificant, the ones for zones 1 and 5 are negative and the one for zone 2 is positive. This gives us an understanding of the impact of unobserved factors in affecting the choice probability for an alternative in a specific zone, on other alternatives in

the same zone. For example, the Zone correlation 1 and 5 being negative, signifies that within those zones, the unobserved factors which may lead to a higher probability for an alternative, will result in a lower probability for all other alternatives.

The model results for Steps 2 and 3 from the subsection 5.4.1 Description of model implementation and specification search, addressing each of the endogeneities due to price and VMT respectively, are provided in the 'Appendix' chapter.

5.5 CONCLUSION

In this research, we aimed to develop a control function (CF) based discrete choice model to address the endogeneity in residential location choice, considering the influence of both housing prices and travel behaviour simultaneously. Our analysis provides interesting insights into the field of residential location choice models by offering insights into the interplay between these factors and their impact on location decisions. Our hypothesis was confirmed, showing that housing prices remain a primary and substantial factor even after accounting for endogeneity and variation. However, we also found that the effect of travel behaviour cannot be ignored, and it must be addressed in conjunction with housing prices to obtain unbiased estimates and a comprehensive understanding of the residential location choice process. Addressing endogeneity for both housing prices and travel behaviour is vital for multiple reasons. Firstly, it allows for a more accurate depiction of the bidirectional relationship between residential location choice and travel behaviour, providing a holistic understanding of their interactions. Secondly, obtaining unbiased estimates for both factors ensures that policy implications and decision-making processes are grounded in reliable results. Thirdly, the comprehensive approach enables us to disentangle the unique and combined effects of housing prices and travel behaviour on location choices, shedding light on their relative significance.

We employed the CF method, utilizing instrumental variables correlated with endogenous variables but not the error term, effectively handling endogeneity due to both housing prices and vehicle miles travelled (VMT) in residential location choice. Our data were sampled from the 2018-19 London Travel Demand Survey (LTDS), enabling detailed analysis of household residential location choices and their relationship to travel behaviour. The results of our models revealed that housing prices have a negative impact on residential location choice, and this effect is more pronounced for lower-income groups. Simultaneously, travel behaviour, represented by

VMT, plays a role in location choices, with households having higher VMT being deterred by areas with higher population density. Moreover, non-white households show a preference for locations with higher crime density, while white households exhibit the opposite tendency. By utilizing the CF-MxL model, we achieved better model fit statistics compared to the MNL model (without endogeneity correction), highlighting the importance of accounting for endogeneity in residential location choice models. Our findings emphasize the significance of both housing prices and travel behaviour in shaping location choices, and this understanding can inform urban planning and policy decisions.

While our research provides valuable insights, it also presents potential avenues for further investigation, particularly regarding the mediating effects of unobserved factors on location choices related to crime density and travel behaviour. Overall, this study enriches the knowledge of residential location choice models and offers practical implications for urban policymakers and researchers.

Chapter 6

Conclusion

In this chapter, I summarize the contributions of the research in Section 6.1. Subsequently, I discuss some implications of consideration of alternatives and addressing endogeneity in residential location discrete choice modeling. Finally, based on the insights derived from this study, I conclude this dissertation by discussing future research directions.

6.1 SUMMARY

The primary objective of this dissertation is to propose novel solutions to address consideration of alternatives in choice contexts with a large number of alternatives, namely disaggregate residential location decision, and addressing endogeneity between residential location and travel behaviour.

The first objective was to develop a probabilistic consideration set formation model of residential location choices. Chapter 3 fulfils this objective by utilising probabilistic decision trees to predict individuals' consideration sets in a large alternative residential location choice model. The decision trees allow for non-compensatory consideration of alternatives, generating unique consideration sets for households in the London Travel Demand Survey data. The study then applies a Mixed Logit model to estimate the final choice decision, capturing interdependence between alternatives within specific Transport Fare Zones. The results reveal factors influencing consideration set formation, highlighting the impact of variables such as household income, borough prices, population density, and crime density on location choice probability. Ethnic variation in the effect of crime density on location choice indicates racial and ethnic segregation in London housing. The research confirms the use of a

non-compensatory process followed by a compensatory one in decision-making. However, the model has data limitations and challenges in merging machine learning and econometric approaches. Future work may involve developing model-based Mixed Logit structures to capture deterministic heterogeneity across households and socioeconomic groups.

The second objective was to develop a comprehensive framework for simultaneous determination of employment accessibility and household trips, by specific modes of drive, public transport, bicycle and walk. Chapter 4 presents a joint modelling framework that simultaneously analyses employment accessibility and household trip frequency across different transport modes in the City of Toronto. Accessibility measures for auto, transit, bicycle, and walk modes were created, along with household trip frequencies for each mode, categorized into low, medium, and high levels. The study uses a multivariate ordered probit model system, considering both observed and unobserved factors' dependence on trip frequencies and accessibility. The results reveal strong associations and direct effects between employment accessibility and household trip frequency. Complementary effects are observed between auto accessibility and transit trips, while substitution effects exist between transit accessibility and auto trips. Policymakers can use these findings to promote transit development in transit-poor neighbourhoods and reduce vehicular congestion in Toronto. The study also highlights the preference for auto use in suburban neighbourhoods, while younger households tend to reside in more walk/transit-oriented urban areas. The limitations of limited cycling and walking infrastructure in Toronto impact trip utility for high population density neighbourhoods, leading to inequities in transport infrastructure distribution. Moreover, certain population groups are time-impaired, limiting their activity participation and raising equity concerns regarding transport infrastructure distribution in the city.

The third objective is to address endogeneity between residential location choice and travel behaviour of a household. Chapter 5 developed a control function (CF)-based discrete choice model to address endogeneity in residential location choice, considering the joint influence of housing prices and travel behaviour. The analysis reveals significant insights into the interplay between these factors and their impact on location decisions. The study confirmed that housing prices remain a substantial factor even after accounting for endogeneity, while travel behaviour also plays a crucial role. Combining these factors is vital for a comprehensive understanding of residential location choice, ensuring reliable policy implications. The CF method, using

instrumental variables, effectively handled endogeneity due to housing prices and vehicle miles travelled (VMT). Data from the 2018-19 London Travel Demand Survey were used to show that housing prices negatively affect location choice, more so for lower-income groups. Higher VMT deters households from areas with high population density. Non-white households prefer higher crime density areas, while white households exhibit the opposite preference. The CF-MxL model outperformed the MNL model in fit statistics, underscoring the importance of considering endogeneity. The findings highlight the significance of housing prices and travel behaviour in location choices, offering insights for urban planning and policies. Further research could explore the mediating effects of unobserved factors on crime density and travel behavior related location choices. Overall, the study enriches residential location choice models and provides practical implications for policymakers and researchers.

6.2 DISCUSSION OF RESULTS

This research set out to support urban authorities make informed decisions for understanding the factors that lead to gentrification and marginalisation in communities of London metropolitan region, from the view point of residential location choices of different types of households. Our results provide an understanding of the factors which contribute to the consideration of alternatives, or in other words, what factors lead to rejection or retention of borough-alternatives in choice process of a household. We saw that a higher price to rent or buy a home adversely impacts all income groups with equal tolerance (or a lack thereof). However, our results teased out the disproportionate social safety offered to different ethnic groups. The effect of crime when mediated by household's racial ethnicity produces significantly varied results. While the effect of crime-density on location choice probability for white households was observed to be negative as expected, the same effect for non-white households was positive. This ethnic variation of crime-density's effect on location choice gives us an insight into the racial and ethnic segregation evident in London housing. These concrete results with statistical significance aim to motivate further analysis into factors that lead to marginalisation of communities by ethnicity, income and social-order

In the joint modelling framework that simultaneously models the employment accessibility and household trip frequency by different transport modes using data for the City of Toronto, we observed a complementarity effect between auto accessibility and transit

trips, and a substitution effect observed between transit accessibility and auto trips. From the perspective of a policymaker or urban authorities, these results form the basis of offering greater transit facilities with better connectivity and employment reach, and their positive impact on the desired reduction in driving trips. We highlight the residential neighbourhood dissonance of transit riders and show that locations with higher transit accessibility (that is, better transit service) are not necessarily locations where people who make more transit trips reside. This lack of accessibility of services to transit dependent users (such as low-income people and people without cars) results in excess time consumed in travelling and hence, leads to time-impoorished classes of society, thereby adversely affecting quality of life. For future transit-development policymakers can utilize our results to identify the transit-poor neighbourhoods within Toronto and improve the transit facilities. Such policy measures can have a multi-fold advantage. A targeted transit development can provide a higher accessibility to low-income neighbourhoods, but also might impact vehicular congestion on the streets of Toronto.

From the perspective of causal inference analysis, it is required to empirically observe the factors causing endogeneity. However, it is not possible to have all factors that contribute to self-selection, or attributes lead to common underlying utilities of various residential location and travel behaviour choices. The instrumental variable models explicitly correct for endogeneity bias without observing all required factors. Hence, the proposed methods require less data in addressing endogeneity, unlike current state-of-the-art causal inference techniques. The empirical applications in this research utilise typical travel demand surveys, such as TTS (Toronto) and LTDS (London). It is important for research to be implemented beyond artificially generated data, to allow for current and future applications accommodating data availability constraints. However, given the importance of uncovering cause and effect between residential location and travel behaviour choices, it would be wise to appreciate the value of causal inference as an immediate next step for future analysis.

In order to provide sustainable and resilient urban development, our research touches upon factors of social inequality, enabling inclusive and equitable development plans, and consider human security concerns for safer, healthier, and more resilient urban environments. For example, highlighting the relationship of accessibility with travel freedom in marginalized communities will help prevent scenarios of *urban food deserts*. Our research aligns with the

objective of identifying even more effective strategies for achieving UN's Sustainable Development Goal (SDG) 11: Make cities inclusive, safe, resilient and sustainable. Further, this research paves the ground for future analysis into causal ILUT microsimulation frameworks for disaggregated data. Advances in machine learning, such as Causal AI methods and high-fidelity modelling, in addition to vast computational capacity and big data availability, provide an unprecedented opportunity to develop novel solutions, with integrated energy and environmental sub-models. These efforts could help curb the environmental impacts of urbanization such as heat island effect, while deterring segregation and gentrification.

6.2.1 Accessibility measurement using this research

A more forward-looking approach to accessibility measurement should enable us to explore the philosophical questions in accurately measuring the different perspectives on accessibility. We have expressed this conceptualization in the form of a diagrammatic representation (Figure 6.1) of relationships between the 'modified' components of accessibility and incorporating the multidimensional interpretation of accessibility. This representation illustrates the fact that a revised framework for accessibility measurement must include all modes of access and incorporates several new components.

From the perspective of a policymaker, the land-use and transport components derive the location of opportunities and the cost of travel to reach them. This can be utilized to design accessibility. On the other end of the spectrum lies the individual characteristics along with their socioeconomic and demographic differences. These observed inter-individual characteristics can be utilized to measure the exhibited accessibility. Between the designed and exhibited interpretations of accessibility lies the perceived accessibility which can be difficult to measure but connects the elements of what the policymaker set out to design and how it was translated to the observed choices of the individual. Examples of this from the data may include perception of safety on the footpath or the width of sidewalk in a neighborhood which may affect the act of accessing the desired activities. In other words, developing on prior work, it is possible now to disentangle the various elements and dimensions of accessibility using the novel data sources, and concretely measure each of these to understand and appreciate the differences between the exhibited and designed accessibility measurement.

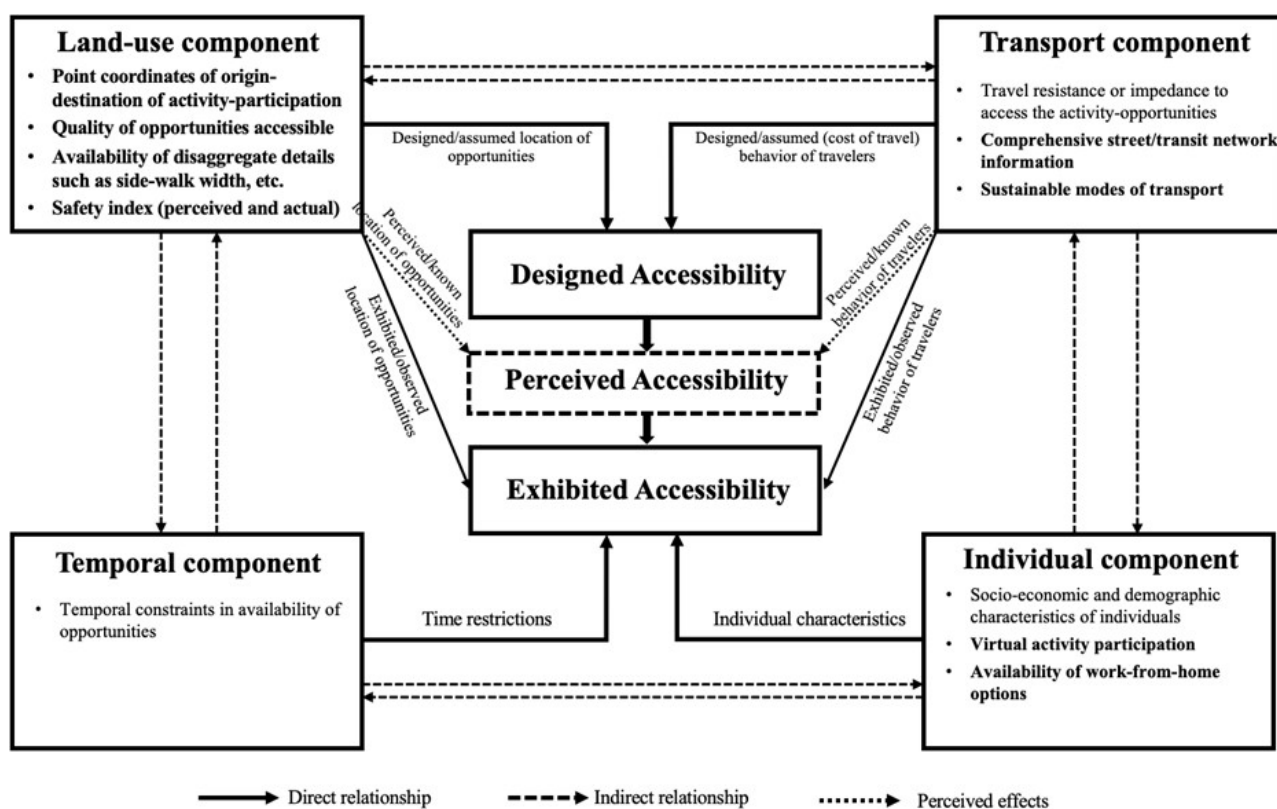


FIGURE 6.1 Diagrammatic illustration of enhanced multidimensional interpretation of accessibility

6.3 LIMITATIONS AND FUTURE RESEARCH DIRECTIONS

Chapter 3 demonstrates the applicability of non-compensatory protocols and offers insights into how socioeconomic and demographic factors filter alternatives in residential location decisions. However, our study has limitations; the complex hybrid model is data-intensive, limiting its applicability to other scenarios, and the merger of machine learning and econometric models is an evolving field that requires further development. Future research directions could involve developing discrete choice model-based recursive partitioning frameworks to account for deterministic heterogeneity across households or socioeconomic groups, addressing the limitations and enhancing the applicability of these methodologies for various contexts.

Chapter 4 presents a joint modelling framework that simultaneously models the employment accessibility and household trip frequency by different transport modes using data for the City of Toronto. While the analysis utilising multivariate ordered probit model system confirm our hypothesis that not only does a strong association exist between employment accessibility and household trip frequency, but there are direct effects to be observed, it does have certain limitations. More comprehensive and multidimensional accessibility indicators may provide a more realistic proxy for the residential location choice, since residential location or relocation is often caused by attractive jobs, education and healthcare opportunities and better lifestyle. Similarly, accessibility measures for walk accessibility may consider the “quality” of the walking experience. Also, trip frequencies might be usefully categorized by purpose. The TTS data has also been observed to under-report walk and bicycle trips and short distance trips (Harding et al., 2018).

Chapter 5 successfully develops a control function (CF) based discrete choice model to address endogeneity in residential location choice, certain limitations and potential future research directions should be acknowledged. Firstly, our analysis focuses on the endogeneity due to only housing prices and travel behaviour, leaving room for the consideration of other location-choice attributes and dwelling-level factors that could contribute to residential location decisions. To enhance the generalizability of our findings, future studies could replicate the analysis using diverse datasets from different geographic regions and time periods. The methodology could be refined or extended to address potential complexities arising from different types of endogeneity or unobserved heterogeneity. Exploring the

robustness of the CF approach across various contexts and scenarios would contribute to a more nuanced understanding of its applicability and limitations.

Chapter 7

Bibliography

Aditjandra, P.T., Cao, X. and Mulley, C., 2016. Exploring changes in public transport use and walking following residential relocation: a British case study. *Journal of Transport and Land Use*, 9(3), pp.77-95

Agyemang-Duah, K., Hall, F.L., 1997. Spatial transferability of an ordered response model of trip generation. *Transportation Research Part A* 31 (5), 389–402.

Agyemang-Duah, K., Hall, F.L., Anderson, W.P., 1995. Trip generation for shopping travel. *Transportation Research Record* 1493, 12–20.

Aitchison, J. and Silvey, S.D., 1957. The generalization of probit analysis to the case of multiple responses. *Biometrika*, 44(1/2), pp.131-140.

Allen, J., and Farber, S. (2020). Planning transport for social inclusion: An accessibility-activity participation approach. *Transportation Research Part D: Transport and Environment*, 78, 102212.

Alonso, W. 1964. *Location and Land Use*. Cambridge: Harvard University Press.

Andrews, R.L. and Srinivasan, T.C., 1995. Studying consideration effects in empirical choice models using scanner panel data. *Journal of Marketing Research*, 32(1), pp.30-41.

Araña, J.E. and León, C.J., 2009. Understanding the use of non-compensatory decision rules in discrete choice experiments: the role of emotions. *Ecological Economics*, 68(8-9), pp.2316-2326.

Arentze, T., Timmermans, H. and Hofman, F., 2007. Creating synthetic household populations: Problems and approach. *Transportation Research Record*, 2014(1), pp.85-91.

Arentze, T.A. and Timmermans, H.J., 2004. A learning-based transportation oriented simulation system. *Transportation Research Part B: Methodological*, 38(7), pp.613-633.

Bagley, M.N. and Mokhtarian, P.L., 2002. The impact of residential neighbourhood type on travel behaviour: A structural equations modelling approach. *The Annals of regional science*, 36(2), pp.279-297.

Bagley, M.N., and P.L. Mokhtarian. The Impact of Residential Neighborhood Type on Travel Behavior: A Structural Equations Modeling Approach. *Annals of Regional Science*, Vol. 36, No. 2, 2002, pp. 279–297. <https://doi.org/10.1007/s001680200083>.

Bayer, P., McMillan, R. and Rueben, K., 2005. Residential segregation in general equilibrium (No. w11095). National Bureau of Economic Research.

Ben-Akiva, M. and Boccara, B., 1995. Discrete choice models with latent choice sets. *International journal of Research in Marketing*, 12(1), pp.9-24.

Ben-Akiva, Moshe, McFadden, Daniel, Gopinah, Dinesh, Garling, Tommy, Bolduc,

Denis, Borsch Supan, Axel, Delquie, Philippe, Larichev, Oleg, Morikawa, Taka, Polydoropoulou, Amalia and Rao, Vithala. "Extended Framework for Modeling Choice Behavior." *Marketing Letters*, August 1999, 10 (3), pp. 187-203

Ben-Akiva, Moshe. "The Structure of Travel Demand Models." Ph.D. dissertation, Massachusetts Institute of Technology, 1972.

Berry, Steven, James Levinsohn & Ariel Pakes (1995) "Automobile Prices in Market Equilibrium," *Econometrica*, 63(4): 841-890 (July 1995).

Bhat, C.R. and Guo, J., 2004. A mixed spatially correlated logit model: formulation and application to residential choice modeling. *Transportation Research Part B: Methodological*, 38(2), pp.147-168.

Bhat, C.R. and Sidharthan, R., 2011. A simulation evaluation of the maximum approximate composite marginal likelihood (MACML) estimator for mixed multinomial probit models. *Transportation Research Part B: Methodological*, 45(7), pp.940-953.

Bhat, C.R., 2015. A comprehensive dwelling unit choice model accommodating psychological constructs within a search strategy for consideration set formation. *Transportation Research Part B: Methodological*, 79, pp.161-188.

Bhat, C.R., Born, K., Sidharthan, R. and Bhat, P.C., 2014. A count data model with endogenous covariates: formulation and application to roadway crash frequency at intersections. *Analytic Methods in Accident Research*, 1, pp.53-71.

Bhat, C.R., Guo, J.Y., 2007. A comprehensive analysis of built environment characteristics on household residential choice and auto ownership levels. *Transportation Research Part B* 41 (5), 506–526.

Bhat, C.R., S. Astroza, A.C. Bhat, and K. Nagel (2016) Incorporating a Multiple Discrete-Continuous Outcome in the Generalized Heterogeneous Data Model: Application to Residential Self-Selection Effects Analysis in an Activity Time-use Behaviour Model, *Transportation Research Part B*, 91, 52-76.

Bierlaire, M., Hurtubia, R. and Flötteröd, G., 2010. Analysis of implicit choice set generation using a constrained multinomial logit model. *Transportation research record*, 2175(1), pp.92-97.

Boarnet, M.G., Bostic, R.W., Rodnyansky, S., Burinskiy, E., Eisenlohr, A., Jamme, H.T. and Santiago-Bartolomei, R., 2020. Do high income households reduce driving more when living near rail transit? *Transportation research part D: transport and environment*, 80, p.102244.

Boisjoly, G. and El-Geneidy, A.M., 2017. How to get there? A critical assessment of accessibility objectives and indicators in metropolitan transportation plans. *Transport Policy*, 55, pp.38-50.

Boisjoly, G. and El-Geneidy, A.M., 2017. How to get there? A critical assessment of accessibility objectives and indicators in metropolitan transportation plans. *Transport Policy*, 55, pp.38-50.

Bovy, P.H., 2009. On modelling route choice sets in transportation networks: a synthesis. *Transport reviews*, 29(1), pp.43-68.

Brathwaite, T., Vij, A. and Walker, J.L., 2017. Machine learning meets microeconomics: The case of decision trees and discrete choice. *arXiv preprint arXiv:1711.04826*.

Breiman, L., 2001. Random forests. *Machine learning*, 45, pp.5-32.

Breiman, L., Friedman, J., Stone, C.J. and Olshen, R.A., 1984. *Classification and regression trees*. CRC press.

Brown, A.E., Blumenberg, E., Taylor, B.D., Ralph, K. and Voulgaris, C.T., 2016. A taste

for transit? Analyzing public transit use trends among youth. *Journal of Public Transportation*, 19(1), pp.49-67.

Brownstone, D. and Golob, T.F., 2009. The impact of residential density on vehicle usage and energy consumption. *Journal of urban Economics*, 65(1), pp.91-98.

Brownstone, D. and Golob, T.F., 2009. The impact of residential density on vehicle usage and energy consumption. *Journal of urban Economics*, 65(1), pp.91-98.

Buehler, R. *Transport Policies, Travel Behavior, and Sustainability: A Comparison of Germany and the U.S.* PhD dissertation. Rutgers University, New Brunswick, N.J., 2008. <http://dx.doi.org/doi:10.7282/T30R9PR7>

Bühlmann, P. and Hothorn, T., 2007. Boosting algorithms: Regularization, prediction and model fitting.

Cantillo, V. and de Dios Ortúzar, J., 2005. A semi-compensatory discrete choice model with explicit attribute thresholds of perception. *Transportation Research Part B: Methodological*, 39(7), pp.641-657.

Cantillo, V., Heydecker, B. and de Dios Ortúzar, J., 2006. A discrete choice model incorporating thresholds for perception in attribute values. *Transportation Research Part B: Methodological*, 40(9), pp.807-825.

Cao, X. (2010) Exploring Causal Effects of Neighbourhood Type on Walking Behaviour Using Stratification on the Propensity Score. *Environment and Planning A*, 42(2), 487-504.

Cao, X., 2014. Residential self-selection in the relationships between the built environment and travel behavior: Introduction to the special issue. *Journal of Transport and Land Use*, 7(3), pp.1-3.

Cao, X., and Y. Fan (2012) Exploring the Influences of Density on Travel Behaviour Using Propensity Score Matching. *Environment and Planning B*, 39(3), 459-470.

Cao, X., Mokhtarian, P. L., & Handy, S. (2009). Examining the impacts of residential self-selection on travel behaviour: A focus on empirical findings. *Transport Reviews*, 29, 359–395.

Cao, X., Mokhtarian, P. L., and Handy, S. (2009). Examining the impacts of residential self-selection on travel behaviour: A focus on empirical findings. *Transport Reviews*, 29, 359–395.

Cao, X., Mokhtarian, P. L., and Handy, S. (2009). Examining the impacts of residential self-selection on travel behaviour: A focus on empirical findings. *Transport Reviews*, 29, 359–395.

Caplin, A., Dean, M. and Leahy, J., 2019. Rational inattention, optimal consideration sets, and stochastic choice. *The Review of Economic Studies*, 86(3), pp.1061-1094.

Capurso, M., Hess, S. and Dekker, T., 2019. Modelling the role of consideration of alternatives in mode choice: An application on the Rome-Milan corridor. *Transportation Research Part A: Policy and Practice*, 129, pp.170-184.

Carrasco, J.-A., Miller, E.J., 2009. The social dimension in action: a multilevel, personal networks model of social activity frequency between individuals. *Transportation Research Part A* 43 (1), 90–104.

Caruana, R. and Niculescu-Mizil, A., 2006, June. An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd international conference on Machine learning* (pp. 161-168).

Cervero, R. and Duncan, M., 2003. Walking, bicycling, and urban landscapes: evidence from the San Francisco Bay Area. *American journal of public health*, 93(9), pp.1478-1483.

Cervero, R. and Duncan, M., 2003. Walking, bicycling, and urban landscapes: evidence

from the San Francisco Bay Area. *American journal of public health*, 93(9), pp.1478-1483.

Cervero, R., & Kockelman, K. (1997). Travel demand and the 3Ds: Density, diversity, and design. *Transportation Research Part D: Transport and Environment*, 2(3), 199-219.

Cervero, R., and Kockelman, K. (1997). Travel demand and the 3Ds: Density, diversity, and design. *Transportation Research Part D: Transport and Environment*, 2(3), 199-219.

Chan, K.Y. and Loh, W.Y., 2004. LOTUS: An algorithm for building accurate and comprehensible logistic regression trees. *Journal of Computational and Graphical Statistics*, 13(4), pp.826-852.

Cheng, L., Chen, X., De Vos, J., Lai, X. and Witlox, F., 2019. Applying a random forest method approach to model travel mode choice behavior. *Travel behaviour and society*, 14, pp.1-10.

Chipman, H.A., George, E.I. and McCulloch, R.E., 1998. Bayesian CART model search. *Journal of the American Statistical Association*, 93(443), pp.935-948.

Choocharukul, K., Van, H.T. and Fujii, S., 2008. Psychological effects of travel behavior on preference of residential location choice. *Transportation Research Part A: Policy and Practice*, 42(1), pp.116-124.

Cockx, K. and Canters, F., 2020. Determining heterogeneity of residential location preferences of households in Belgium. *Applied Geography*, 124, p.102271.

Copperman, R., and C.R. Bhat (2007) An Analysis of the Determinants of Children's Weekend Physical Activity Participation. *Transportation*, 34(1), 67-87

Copperman, R.B. and Bhat, C.R., 2007. An analysis of the determinants of children's weekend physical activity participation. *Transportation*, 34, pp.67-87.

Criminisi, A., Shotton, J. and Konukoglu, E., 2012. Decision forests: A unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning. *Foundations and trends® in computer graphics and vision*, 7(2-3), pp.81-227.

Data Management Group. (2016). *Transportation Tomorrow Survey*. Retrieved from <http://dmg.utoronto.ca/>

De Vos, J., Derudder, B., Van Acker, V. and Witlox, F., 2012. Reducing car use: changing attitudes or relocating? The influence of residential dissonance on travel behavior. *Journal of Transport Geography*, 22, pp.1-9.

De Vos, J., Ettema, D. and Witlox, F., 2018. Changing travel behaviour and attitudes following a residential relocation. *Journal of transport geography*, 73, pp.131-147.

De Vos, J., Ettema, D. and Witlox, F., 2019. Effects of changing travel patterns on travel satisfaction: A focus on recently relocated residents. *Travel Behaviour and Society*, 16, pp.42-49.

De Vos, J., Mokhtarian, P.L., Schwanen, T., Van Acker, V. and Witlox, F., 2016. Travel mode choice and travel satisfaction: bridging the gap between decision utility and experienced utility. *Transportation*, 43, pp.771-796.

De Witte, A., Hollevoet, J., Dobruszkes, F., Hubert, M. and Macharis, C., 2013. Linking modal choice to motility: A comprehensive review. *Transportation Research Part A: Policy and Practice*, 49, pp.329-341.

Dietterich, T.G., 2000, June. Ensemble methods in machine learning. In *International workshop on multiple classifier systems* (pp. 1-15). Berlin, Heidelberg: Springer Berlin Heidelberg.

Du, H. and Mulley, C., 2012. Understanding spatial variations in the impact of accessibility on land value using geographically weighted regression. *Journal of Transport and*

Land Use, 5(2), pp.46-59.

Du, H. and Mulley, C., 2012. Understanding spatial variations in the impact of accessibility on land value using geographically weighted regression. *Journal of Transport and Land Use*, 5(2), pp.46-59.

Ettema, D. and Nieuwenhuis, R., 2017. Residential self-selection and travel behaviour: What are the effects of attitudes, reasons for location choice and the built environment?. *Journal of transport geography*, 59, pp.146-155.

Ewing, R. and Cervero, R. (2001) 'Travel and the Built Environment: A Synthesis', *Transportation Research Record*, 1780(1), pp. 87–114. doi: 10.3141/1780-10.

Ewing, R., and Cervero, R. (2010). Travel and the built environment: A meta-analysis. *Journal of the American planning association*, 76(3), 265- 294.

Farinloye, T., Mogaji, E., Aririguzoh, S. and Kieu, T.A., 2019. Qualitatively exploring the effect of change in the residential environment on travel behaviour. *Travel behaviour and society*, 17, pp.26-35.

Farinloye, T., Mogaji, E., Aririguzoh, S. and Kieu, T.A., 2019. Qualitatively exploring the effect of change in the residential environment on travel behaviour. *Travel behaviour and society*, 17, pp.26-35.

Farooq, B. and Miller, E. J. 2012. Towards integrated land use and transportation: A dynamic disequilibrium-based microsimulation framework for built space markets. *Transportation Research Part A: Policy and Practice*, 46, 1030-1053

Farooq, B., Salvini, P. A., & Miller, E. J. (2008). Development of an operational integrated urban model system. Volume X: ILUTE Software Documentation (Vol. X).

Fernández-Antolín, A., Guevara, C. A., de Lapparent, M., & Bierlaire, M. (2016). Correcting for endogeneity due to omitted attitudes: Empirical assessment of a modified MIS method using RP mode choice data. *Journal of Choice Modelling*, 20, 1-15.

Fordham, L.; Gris , E.; El-Geneidy, A. When I'm 64: Assessing generational differences in public transit use of seniors in Montreal, Quebec, Canada. *Transp. Res. Rec.* 2017, 2651, 31–41.

Foth, N., Manaugh, K. and El-Geneidy, A.M., 2013. Towards equitable transit: examining transit accessibility and social need in Toronto, Canada, 1996–2006. *Journal of transport geography*, 29, pp.1-10.

Foth, N., Manaugh, K. and El-Geneidy, A.M., 2013. Towards equitable transit: examining transit accessibility and social need in Toronto, Canada, 1996–2006. *Journal of transport geography*, 29, pp.1-10.

Foth, N., Manaugh, K. and El-Geneidy, A.M., 2014. Determinants of mode share over time: how changing transport system affects transit use in Toronto, Ontario, Canada. *Transportation Research Record*, 2417(1), pp.67-77.

Foth, N., Manaugh, K. and El-Geneidy, A.M., 2014. Determinants of mode share over time: how changing transport system affects transit use in Toronto, Ontario, Canada. *Transportation Research Record*, 2417(1), pp.67-77.

Frank, L.D. and Engelke, P.O., 2001. The built environment and human activity patterns: exploring the impacts of urban form on public health. *Journal of planning literature*, 16(2), pp.202-218.

Frank, L.D. and P.O. Engelke (2001) The Built Environment and Human Activity Patterns: Exploring the Impacts of Urban Form on Public Health. *Journal of Planning Literature*, 16(2), 202-218.

Furth, P.G., Mekuria, M.C. and Nixon, H., 2016. Network connectivity for low-stress bicycling. *Transportation research record*, 2587(1), pp.41-49.

Gilbride, T.J. and Allenby, G.M., 2004. A choice model with conjunctive, disjunctive, and compensatory screening rules. *Marketing Science*, 23(3), pp.391-406.

Ginsberg, Y. and Churchman, A., 1984. Housing satisfaction and intention to move: Their explanatory variables. *Socio-Economic Planning Sciences*, 18(6), pp.425-431.

Goetzke, F. and Weinberger, R., 2012. Separating contextual from endogenous effects in automobile ownership models. *Environment and Planning A*, 44(5), pp.1032-1046.

Golob, T.F., Kitamura, R. and Lula, C., 1994, January. Modeling the effects of commuting time on activity duration and non-work travel. In *73rd Annual Meeting of the Transportation Research Board*, Washington, DC.

Green et al., 1988 Green, P.E. and Krieger, A.M., 1988. Choice rules and sensitivity analysis in conjoint simulators. *Journal of the Academy of Marketing Science*, 16, pp.114-127.

Griliches, Zvi. "Specification Bias in Estimates of Production Functions." *Journal of Farm Economics*, February 1957, 39(1), pp. 8-20. . "Hedonic Price Indices for Automobiles," in *The price statistics of the federal government*. Washington, DC: National Bureau of Economic Research, 1961, pp. 173-96

Guevara, C. (2010), "Endogeneity and Sampling of Alternatives in Spatial Choice Models," Ph.D. Thesis, Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, MA.

Guevara, C.A. and Ben-Akiva, M., 2006. Endogeneity in residential location choice models. *Transportation research record*, 1977(1), pp.60-66.

Guevara, C.A. and Ben-Akiva, M.E., 2012. Change of scale and forecasting with the control-function method in logit models. *Transportation Science*, 46(3), pp.425-437.

Guevara, C.A., 2015. Critical assessment of five methods to correct for endogeneity in discrete-choice models. *Transportation Research Part A: Policy and Practice*, 82, pp.240-254.

Guevara, C.A., 2018. Overidentification tests for the exogeneity of instruments in discrete choice models. *Transportation Research Part B: Methodological*, 114, pp.241-253.

Guevara, C.A., and M. Ben-Akiva. Endogeneity in Residential Location Choice Models. *Transportation Research Record: Journal of the Transportation Research Board*, No. 1977, 2006, pp. 60–66. <https://doi.org/10.3141/1977-10>.

Guevara, C.A., Tang, Y. and Gao, S., 2017. The initial condition problem with complete history dependency in learning models for travel choices. *Transportation research procedia*, 23, pp.758-771.

Guo, J.Y. and Bhat, C.R., 2007. Operationalizing the concept of neighborhood: Application to residential location choice analysis. *Journal of transport geography*, 15(1), pp.31-45.

Gutiérrez-Vargas, Á.A., Meulders, M. and Vandebroek, M., 2023. Modeling preference heterogeneity using model-based decision trees. *Journal of choice modelling*, 46, p.100393.

Haab, T.C. and Hicks, R.L., 1997. Accounting for choice set endogeneity in random utility models of recreation demand. *Journal of Environmental Economics and Management*, 34(2), pp.127-147.

Handy, S., 2020. Is accessibility an idea whose time has finally come? *Transportation Research Part D: Transport and Environment*, 83, p.102319.

Handy, S., Cao, X., & Mokhtarian, P. L. (2005). Correlation or causality between the built environment and travel behavior? Evidence from Northern California. *Transportation*

Research Part D: Transport and Environment, 10, 427–444.

Handy, S., Cao, X., and Mokhtarian, P. L. (2005). Correlation or causality between the built environment and travel behaviour? Evidence from Northern California. *Transportation Research Part D: Transport and Environment*, 10, 427–444.

Haque, M.B., Choudhury, C., Hess, S. and dit Sourd, R.C., 2019. Modelling residential mobility decision and its impact on car ownership and travel mode. *Travel behaviour and society*, 17, pp.104-119.

Haque, M.B., Choudhury, C., Hess, S. and dit Sourd, R.C., 2019. Modelling residential mobility decision and its impact on car ownership and travel mode. *Travel behaviour and society*, 17, pp.104-119.

Haque, M.B., Choudhury, C., Hess, S. and dit Sourd, R.C., 2019. Modelling residential mobility decision and its impact on car ownership and travel mode. *Travel behaviour and society*, 17, pp.104-119.

Harding, Chris, Monika Nasterska, Leila Dianat, and Eric J. Miller. "Effect of land use and survey design on trip underreporting in Montreal and Toronto's regional surveys." *European Journal of Transport and Infrastructure Research* 18, no. 1 (2018).

Hauser, J.R., Ding, M. and Gaskin, S.P., 2009, May. Non-compensatory (and compensatory) models of consideration-set decisions. In *Proceedings of the Sawtooth Software Conference* (Vol. 14, pp. 207-232). Provo, UT: Sawtooth Software.

Hawkins, J. and Nurul Habib, K., 2019. Integrated models of land use and transportation for the autonomous vehicle revolution. *Transport reviews*, 39(1), pp.66-83.

Helveston, J. P. (2016). *Development and Adoption of Plug-in Electric Vehicles in China: Markets, Policy, and Innovation*. PhD Thesis, Carnegie Mellon University.

Hess, S. and Daly, A. eds., 2014. *Handbook of choice modelling*. Edward Elgar Publishing.

Hillel, T., Bierlaire, M., Elshafie, M.Z. and Jin, Y., 2021. A systematic review of machine learning classification methodologies for modelling passenger mode choice. *Journal of choice modelling*, 38, p.100221.

Hirk, R., Hornik, K. and Vana, L., 2017. mvord: an R package for fitting multivariate ordinal regression models. *R package vignette*, 2018a. URL https://cran.r-project.org/web/packages/mvord/vignettes/vignette_mvord.pdf.

Hong, J. and Goodchild, A., 2014. Land use policies and transport emissions: Modeling the impact of trip speed, vehicle characteristics and residential location. *Transportation Research Part D: Transport and Environment*, 26, pp.47-51.

Hong, J., Shen, Q. and Zhang, L., 2014. How do built-environment factors affect travel behavior? A spatial analysis at different geographic scales. *Transportation*, 41, pp.419-440.

Hothorn, T., Hornik, K. and Zeileis, A., 2006. Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical statistics*, 15(3), pp.651-674.

Huber, J. and Klein, N.M., 1991. Adapting cutoffs to the choice environment: the effects of attribute correlation and reliability. *Journal of Consumer Research*, 18(3), pp.346-357.

Hurtubia, R., Bierlaire, M. and Martínez, F. Dynamic microsimulation of location choices with a quasi-equilibrium auction approach. *Swiss Transport Research Conference*, 2011 Monte Verità, Switzerland.

Iacono, M., Levinson, D., & El-Geneidy, A. (2008). Models of transportation and land use change: A guide to the territory. *Journal of Planning Literature*, 22(4), 323–340. doi:10.1177/0885412207314010

Iacono, M., Levinson, D., and El-Geneidy, A. (2008). Models of transportation and land use change: A guide to the territory. *Journal of Planning Literature*, 22(4), 323–340. doi:10.1177/ 0885412207314010

Imani, A.F., Miller, E.J. and Saxe, S., 2019. Cycle accessibility and level of traffic stress: A case study of Toronto. *Journal of transport geography*, 80, p.102496.

Janke, J. and Handy, S., 2019. How life course events trigger changes in bicycling attitudes and behaviour: Insights into causality. *Travel behaviour and society*, 16, pp.31-41.

Jedidi, K. and Kohli, R., 2005. Probabilistic subset-conjunctive models for heterogeneous consumers. *Journal of Marketing Research*, 42(4), pp.483-494.

Kahneman, D., 1997. New challenges to the rationality assumption. *Legal Theory*, 3(2), pp.105-124.

Kahneman, Daniel. "Economists Have Preferences, Psychologists Have Attitudes." Invited presentation to NSF Symposium on Preference Elicitation, Berkeley, CA, August 1997.

Kamruzzaman, M., De Vos, J., Currie, G., Giles-Corti, B. and Turrell, G., 2020. Spatial biases in residential mobility: Implications for travel behaviour research. *Travel behaviour and society*, 18, pp.15-28.

Kaplan, S., Bekhor, S. and Shiftan, Y., 2009. Two-stage model for jointly revealing determinants of noncompensatory conjunctive choice set formation and compensatory choice. *Transportation research record*, 2134(1), pp.153-163.

Kaplan, S., Shiftan, Y. and Bekhor, S., 2012. Development and estimation of a semi-compensatory model with a flexible error structure. *Transportation Research Part B: Methodological*, 46(2), pp.291-304.

Karlaftis, M.G., 2004. A DEA approach for evaluating the efficiency and effectiveness of urban transit systems. *European Journal of Operational Research*, 152(2), pp.354-364.

Kim, J., and D. Brownstone (2013) The Impact of Residential Density on Vehicle Usage and Fuel Consumption: Evidence from national samples. *Energy Economics*, 40, 196-206.

Kim, J.-H. (1994) Residential mobility intention by tenure type in Korea, *Journal of Social Science*, 4(2), pp. 1–22.

Kim, J.H., Pagliara, F. and Preston, J., 2005. The intention to move and residential location choice behaviour. *Urban studies*, 42(9), pp.1621-1636.

Kim, K., 1995. A bivariate cumulative probit regression model for ordered categorical data. *Statistics in Medicine*, 14(12), pp.1341-1352.

Kitamura, R., 1988. A dynamic model system of household car ownership, trip generation, and modal split: Model development and simulation experiment. In: *Proceedings of the 14th Australian Road Research Board Conference, Part 3*, pp. 96–111.

Kitamura, R., Mokhtarian, P.L., Laidet, L., 1997. A micro-analysis of land use and travel in five neighborhoods in the San Francisco Bay Area. *Transportation* 24, 125–158. <https://doi.org/10.1023/A:1017959825565>

Kitamura, R., Robinson, J., Golob, T., Bradley, M., Leonard, J. and van der Hoorn, T., 1992. A comparative analysis of time use data in the Netherlands and California.

Knox, P. and Pinch, S. (2000) *Urban Social Geography: An Introduction*. Englewood Cliffs, NJ: Prentice-Hall.

Kohli R, Jedidi K (2017) Relations between EBA and nested logit models. In: *Operations Research Forthcoming*

Krizek, K.J., 2003. Residential relocation and changes in urban travel: Does neighborhood-scale urban form matter?. *Journal of the American Planning Association*,

69(3), pp.265-281.

Krizek, K.J., 2003. Residential relocation and changes in urban travel: Does neighbourhood-scale urban form matter? *Journal of the American Planning Association*, 69(3), pp.265-281.

Landwehr, N., Hall, M. and Frank, E., 2005. Logistic model trees. *Machine learning*, 59, pp.161-205.

Lee, C. and A.V. Moudon (2006) The 3Ds+ R: Quantifying Land Use and Urban Form Correlates of Walking. *Transportation Research Part D*, 11(3), 204-215.

Lee, C. and Moudon, A.V., 2006. Correlates of walking for transportation or recreation purposes. *Journal of Physical Activity and health*, 3(s1), pp.S77-S98.

Lemon, S.C., Roy, J., Clark, M.A., Friedmann, P.D. and Rakowski, W., 2003. Classification and regression tree analysis in public health: methodological review and comparison with logistic regression. *Annals of behavioral medicine*, 26, pp.172-181.

Lerman, S.R., 1976. Location, housing, automobile ownership, and mode to work: a joint choice model. *Transportation Research Record*, 610, pp.6-11.

Lesaffre, E. and Molenberghs, G., 1991. Multivariate probit analysis: a neglected procedure in medical statistics. *Statistics in Medicine*, 10(9), pp.1391-1403.

Levinson, D.M., 1998. Accessibility and the journey to work.

Levinson, D.M., 1998. Accessibility and the journey to work. *Journal of transport geography*, 6(1), pp.11-21.

Liao, F.H., Farber, S. and Ewing, R., 2015. Compact development and preference heterogeneity in residential location choice behaviour: A latent class analysis. *Urban studies*, 52(2), pp.314-337.

Liao, F.H., Farber, S. and Ewing, R., 2015. Compact development and preference heterogeneity in residential location choice behaviour: A latent class analysis. *Urban studies*, 52(2), pp.314-337.

Liao, F.H., Farber, S. and Ewing, R., 2015. Compact development and preference heterogeneity in residential location choice behaviour: A latent class analysis. *Urban Studies*, 52(2), pp.314-337.

Lin, T., Wang, D. and Guan, X., 2017. The built environment, travel attitude, and travel behavior: Residential self-selection or residential determination?. *Journal of transport geography*, 65, pp.111-122.

Loh, W.Y. and Shih, Y.S., 1997. Split selection methods for classification trees. *Statistica sinica*, pp.815-840.

Loh, W.Y. and Vanichsetakul, N., 1988. Tree-structured classification via generalized discriminant analysis. *Journal of the American Statistical Association*, 83(403), pp.715-725.

Loh, W.Y., 2002. Regression tress with unbiased variable selection and interaction detection. *Statistica sinica*, pp.361-386.

Loh, W.Y., 2011. Classification and regression trees. *Wiley interdisciplinary reviews: data mining and knowledge discovery*, 1(1), pp.14-23.

Lowry, I. S. (1964). A model of metropolis. Retrieved from https://www.rand.org/content/dam/rand/pubs/research_memoranda/2006/RM4035.pdf

Lowry, I. S. 1964. A model of metropolis. Technical report, RAND Europe.

Luce, R. Duncan. Individual choice behavior. New York: Wiley, 1959.

Lurkin, V., Garrow, L. A., Higgins, M. J., Newman, J. P., & Schyns, M. (2017). Accounting for price endogeneity in airline itinerary choice models: An application to

- Continental US markets. *Transportation Research Part A: Policy and Practice*, 100, 228-246
- Maddala, G.S., 1983. *Qualitative and limited dependent variable models in econometrics*. Cambridge: Cambridge, 498.
- Maddala, G.S., 1986. *Limited-dependent and qualitative variables in econometrics* (No. 3). Cambridge university press.
- Manaugh, K. and El-Geneidy, A., 2012. 12. Who Benefits from New Transportation Infrastructure? Using Accessibility Measures to Evaluate Social Equity in Public Transport Provision. *Accessibility analysis and transport planning: Challenges for Europe and North America*, 211.
- Manaugh, K. and El-Geneidy, A., 2012. Who benefits from new transportation infrastructure? Using accessibility measures to evaluate social equity in public transport provision. In *Accessibility Analysis and Transport Planning*. Edward Elgar Publishing.
- Manaugh, K., Miranda-Moreno, L.F. and El-Geneidy, A.M., 2010. The effect of neighbourhood characteristics, accessibility, home-work location, and demographics on commuting distances. *Transportation*, 37(4), pp.627-646.
- Manaugh, K., Miranda-Moreno, L.F. and El-Geneidy, A.M., 2010. The effect of neighbourhood characteristics, accessibility, home-work location, and demographics on commuting distances. *Transportation*, 37, pp.627-646.
- Manski, C.F., 1977. The structure of random utility models. *Theory and decision*, 8(3), p.229.
- Marsala, C. and Petturiti, D., 2015. Rank discrimination measures for enforcing monotonicity in decision tree induction. *Information Sciences*, 291, pp.143-171.
- Marschak, J., 1960. Theory of an efficient several-person firm. *The American Economic Review*, 50(2), pp.541-548.
- Marschak, Jacob. "Binary Choice Constraints on Random Utility Indicators," in Kenneth Arrow, ed., *Stanford symposium on mathematical methods in the social sciences*. Stanford, CA: Stanford University Press, 1960, pp. 312-29.
- Martínez, F., Aguila, F. and Hurtubia, R., 2009. The constrained multinomial logit: A semi-compensatory choice model. *Transportation Research Part B: Methodological*, 43(3), pp.365-377.
- McFadden, D., 1977. Modelling the choice of residential location.
- McKelvey, R.D. and Zavoina, W., 1975. A statistical model for the analysis of ordinal level dependent variables. *Journal of mathematical sociology*, 4(1), pp.103-120.
- Meurs, H., 1989. Dynamic analysis of trip generation. In: *Presented at the International Conference on Dynamic Behaviour Analysis*, Kyoto, Japan, July 21–23, 1989
- Meyer, B.D., 1990. Unemployment insurance and unemployment spells. *Econometrica* 58 (4), 757–782.
- Miller, E. (2006). *Integrated Urban Models: Theoretical Prospects*. Proceedings of the 11th International Conference on Travel Behaviour Research. Kyoto, Japan.
- Miller, E. J. (2018). Accessibility: measurement and application in transportation planning. In *Transport Reviews* (Vol. 38, Issue 5, pp. 551–555). Routledge. <https://doi.org/10.1080/01441647.2018.1492778>
- Miller, E.J., 2018. Accessibility: measurement and application in transportation planning. *Transport Reviews*, 38(5), pp.551-555.
- Minka, Thomas P.. "Bayesian model averaging is not model combination." (2002).
- Mitra, R., Smith Lea, N., Cantello, I., and Hanson, G. (2016). Cycling behaviour and

potential in the Greater Toronto and Hamilton Area. Toronto, ON: Ryerson University.

Moeckel, R., 2017. Working from home: Modeling the impact of telework on transportation and land use. *Transportation Research Procedia*, 26, pp.207-214.

Mokhtarian, P.L., and X. Cao. Examining the Impacts of Residential Self-Selection on Travel Behavior: A Focus on Methodologies. *Transportation Research Part B: Methodological*, Vol. 42, No. 3, 2008, pp. 204–228. <https://doi.org/10.1016/j.trb.2007.07.006>.

Mokhtarian, P.L., and X. Cao. Examining the Impacts of Residential Self-Selection on Travel Behaviour: A Focus on Methodologies. *Transportation Research Part B: Methodological*, Vol. 42, No. 3, 2008, pp. 204–228. <https://doi.org/10.1016/j.trb.2007.07.006>.

Moniruzzaman, M., Páez, A., Scott, D. and Morency, C., 2015. Trip generation of seniors and the geography of walking in Montreal. *Environment and Planning A: Economy and Space*, 47(4), pp.957-976.

Moniruzzaman, P.A.; Habib, K.N.; Morency, C. Mode use and trip length of seniors in Montreal. *J. Transp. Geogr.* 2013, 30, 89–99.

Moore, T. and Thorsnes, P., “The Transportation/Land Use Connection, American Planning Association,” Planning Advisory Service, 1994, Report Number 448/449, Chicago.

Næss, P., 2009. Residential self-selection and appropriate control variables in land use: Travel studies. *Transport Reviews*, 29(3), pp.293-324.

Næss, P., 2015. Built environment, causality and travel. *Transport reviews*, 35(3), pp.275-291.

Næss, P., Peters, S., Stefansdottir, H. and Strand, A., 2018. Causality, not just correlation: Residential location, transport rationales and travel behavior across metropolitan contexts. *Journal of transport geography*, 69, pp.181-195.

Næss, P., Peters, S., Stefansdottir, H. and Strand, A., 2018. Causality, not just correlation: Residential location, transport rationales and travel behaviour across metropolitan contexts. *Journal of Transport Geography*, 69, pp.181-195.

Newbold, K.B.; Scott, D.M. Driving over the life course: The automobility of Canada’s millennial, generation x, baby boomer and greatest generations. *Travel Behav. Soc.* 2017, 6, 57–63.

Newbold, K.B.; Scott, D.M. Insights into public transit use by Millennials: The Canadian experience. *Travel Behav. Soc.* 2018, 11, 62–68.

Newton, M.A. and Raftery, A.E., 1994. Approximate Bayesian inference with the weighted likelihood bootstrap. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 56(1), pp.3-26.

Oakil, A.T.M., Manting, D. and Nijland, H., 2016. Determinants of car ownership among young households in the Netherlands: The role of urbanisation and demographic and economic characteristics. *Journal of transport geography*, 51, pp.229-235.

Oh, J.H., 2003. Social bonds and the migration intentions of elderly urban residents: The mediating effect of residential satisfaction. *Population research and policy review*, 22, pp.127-146.

Ortúzar, J. and Willumsen, L.G., 2011. *Modelling transport*. John wiley & sons.

Ortúzar, J. and Willumsen, L.G., 2011. *Modelling transport*. John wiley & sons.

Owen, A. and Levinson, D.M., 2015. Modeling the commute mode share of transit using continuous accessibility to jobs. *Transportation research part A: policy and practice*, 74, pp.110-122.

Owen, A., and Levinson, D.M., 2015. Modeling the commute mode share of transit using

continuous accessibility to jobs. *Transportation Research Part A: Policy and Practice*, 74, pp.110-122.

Paleti, R., 2015. Implicit choice set generation in discrete choice models: Application to household auto ownership decisions. *Transportation Research Part B: Methodological*, 80, pp.132-149.

Paleti, R., C.R. Bhat, and R.M. Pendyala (2013) Integrated Model of Residential Location, Work Location, Vehicle Ownership, and Commute Tour Characteristics. *Transportation Research Record: Journal of the Transportation Research Board*, 2382, 162-172.

Pei, S. and Hu, Q., 2018. Partially monotonic decision trees. *Information Sciences*, 424, pp.104-117.

Petrin, A. and Train, K., 2003. Omitted product attributes in discrete choice models.

Petrin, A. and Train, K., 2010. A control function approach to endogeneity in consumer choice models. *Journal of marketing research*, 47(1), pp.3-13.

Pinjari, A.R., Pendyala, R.M., Bhat, C.R. and Waddell, P.A., 2011. Modeling the choice continuum: an integrated model of residential location, auto ownership, bicycle ownership, and commute tour mode choice decisions. *Transportation*, 38(6), p.933.

Potharst, R. and Feelders, A.J., 2002. Classification trees for problems with monotonicity constraints. *ACM SIGKDD Explorations Newsletter*, 4(1), pp.1-10.

Pramono, A. and Oppewal, H., 2012. Accessibility and the role of the consideration set in spatial choice modelling: A simulation study. *Journal of choice modelling*, 5(1), pp.46-63.

Prelec, Drazen. "Values and Principles: Some Limitations on Traditional Economic Analysis," in A. Etzioni and P. Lawrence, eds., *Perspectives on socioeconomics*. London: M. E. Sharpe, 1991, pp. 131-45.

Quinlan, J.R., 1993, June. Combining instance-based and model-based learning. In *Proceedings of the tenth international conference on machine learning* (pp. 236-243).

Rivers, D. and Vuong, Q.H., 1988. Limited information estimators and exogeneity tests for simultaneous probit models. *Journal of econometrics*, 39(3), pp.347-366.

Roberts, J.H. and Lattin, J.M., 1991. Development and testing of a model of consideration set composition. *Journal of Marketing Research*, 28(4), pp.429-440.

Rubin, D.B., 1981. The bayesian bootstrap. *The annals of statistics*, pp.130-134.

Rutz, O.J. and Watson, G.F., 2019. Endogeneity and marketing strategy research: An overview. *Journal of the Academy of Marketing Science*, 47, pp.479-498.

Sakaria, N.; Stehfest, N. *Millennials and Mobility: Understanding the Millennial Mindset and New Opportunities for Transit Providers* (No. Task 17, TCRP Project J-11); American Public Transportation Association: Washington, DC, USA, 2013.

Scheiner, J., Holz-Rau, C. Changes in travel mode use after residential relocation: a contribution to mobility biographies. *Transportation* 40, 431–458 (2013). <https://doi.org/10.1007/s11116-012-9417-6>

Schirmer, P.M., Van Eggermond, M.A. and Axhausen, K.W., 2014. The role of location in residential location choice models: a review of literature. *Journal of Transport and Land Use*, 7(2), pp.3-21.

Schwanen, T. and Mokhtarian, P.L., 2005. What if you live in the wrong neighborhood? The impact of residential neighborhood type dissonance on distance traveled. *Transportation Research Part D: Transport and Environment*, 10(2), pp.127-151.

Shay, E. and Khatkhat, A.J., 2005. Automobile ownership and use in neotraditional and

conventional neighborhoods. *Transportation research record*, 1902(1), pp.18-25.

Shay, E., and Khattak, A.J., 2005. Automobile ownership and use in neotraditional and conventional neighbourhoods. *Transportation research record*, 1902(1), pp.18-25.

Shocker, A.D., Ben-Akiva, M., Boccara, B. and Nedungadi, P., 1991. Consideration set influences on consumer decision-making and choice: Issues, models, and suggestions. *Marketing letters*, 2, pp.181-197.

Sierpiński, G., 2011. Travel behaviour and alternative modes of transportation. In *Modern Transport Telematics: 11th International Conference on Transport Systems Telematics, TST 2011, Katowice-Ustroń, Poland, October 19-22, 2011. Selected Papers 11* (pp. 86-93). Springer Berlin Heidelberg.

Simon, Herbert A. "Rationality as Process and as Product of Thought." *American Economic Review*, May 1978 (Papers and Proceedings), 68(2), pp. 1-16.

Singh, A.C., Astroza, S., Garikapati, V.M., Pendyala, R.M., Bhat, C.R. and Mokhtarian, P.L., 2018. Quantifying the relative contribution of factors to household vehicle miles of travel. *Transportation research part D: transport and environment*, 63, pp.23-36.

Singh, A.C., Astroza, S., Garikapati, V.M., Pendyala, R.M., Bhat, C.R. and Mokhtarian, P.L., 2018. Quantifying the relative contribution of factors to household vehicle miles of travel. *Transportation research part D: transport and environment*, 63, pp.23-36.

Singh, A.C., Astroza, S., Garikapati, V.M., Pendyala, R.M., Bhat, C.R. and Mokhtarian, P.L., 2018. Quantifying the relative contribution of factors to household vehicle miles of travel. *Transportation research part D: transport and environment*, 63, pp.23-36.

Singleton, R. A. Jr., & Straits, B. C. (2005). *Approaches to social research* (4th ed.). New York, NY: Oxford University Press

Sivakumar, A., and A. Pinjari. Recent Advances in Activity and Travel Pattern Modelling. *Transportation*, Vol. 39, No. 4, 2012, pp. 749–754. <https://doi.org/10.1007/s11116-012-9401-1>.

Srinivasan, S. and Ferreira, J., 2002. Travel behavior at the household level: understanding linkages with residential choice. *Transportation Research Part D: Transport and Environment*, 7(3), pp.225-242.

Strobl, C., Malley, J. and Tutz, G., 2009. An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological methods*, 14(4), p.323.

Stüttgen, P., Boatwright, P. and Monroe, R.T., 2012. A satisficing choice model. *Marketing Science*, 31(6), pp.878-899.

Swait, J., 2001. A non-compensatory choice model incorporating attribute cutoffs. *Transportation Research Part B: Methodological*, 35(10), pp.903-928.

Swait, J., 2001. A non-compensatory choice model incorporating attribute cutoffs. *Transportation Research Part B: Methodological*, 35(10), pp.903-928.

Swait, J., 2001. Choice set generation within the generalized extreme value family of discrete choice models. *Transportation Research Part B: Methodological*, 35(7), pp.643-666.

Swait, J., 2009. Choice models based on mixed discrete/continuous PDFs. *Transportation Research Part B: Methodological*, 43(7), pp.766-783.

Swait, J., Adamowicz, W., Hanemann, M., Diederich, A., Krosnick, J., Layton, D., Provencher, W., Schkade, D. and Tourangeau, R., 2002. Context dependence and aggregation in disaggregate choice analysis. *Marketing Letters*, 13, pp.195-205.

Swait, J.D., 1984. Probabilistic choice set generation in transportation demand models

(Doctoral dissertation, Massachusetts Institute of Technology).

Tang, L., Xiong, C. and Zhang, L., 2015. Decision tree method for modeling travel mode switching in a dynamic behavioral process. *Transportation Planning and Technology*, 38(8), pp.833-850.

Taussig, Frank. *Principles of economics*. New York: Macmillan, 1912.

Tayyaran, M.R., Khan, A.M. and Anderson, D.A., 2003. Impact of telecommuting and intelligent transportation systems on residential location choice. *Transportation Planning and Technology*, 26(2), pp.171-193.

Thronicker, I. and Klinger, T., 2019. Interest into travel-related interventions among urban movers and non-movers. *Travel Behaviour and Society*, 16, pp.88-98.

Thurstone L.L., 1927. Psycho physical analysis. *The American journal of psychology*, 38(3), pp.368-389.

Thurstone, L. L. "A Law of Comparative Judgment." *Psychological Review*, 1927, 34, pp. 273-86.

Tillema, T., Dijst, M. and Schwanen, T., 2010. Face-to-face and electronic communications in maintaining social networks: The influence of geographical and relational distance and of information content. *New media & society*, 12(6), pp.965-983.

Timmermans, H.J., 2003. The saga of integrated land use-transport modeling: how many more dreams before we wake up?. In *Proceedings of the International Association of Traveler Behavior Conference*.

Van Wee, B., 2009. Self-selection: a key to a better understanding of location choices, travel behaviour and transport externalities?. *Transport reviews*, 29(3), pp.279-292.

Vance, C. and Hedel, R., 2007. The impact of urban form on automobile travel: disentangling causation from correlation. *Transportation*, 34(5), pp.575-588.

Vance, C. and Hedel, R., 2007. The impact of urban form on automobile travel: disentangling causation from correlation. *Transportation*, 34, pp.575-588.

Varin, C. and Vidoni, P., 2005. A note on composite likelihood inference and model selection. *Biometrika*, 92(3), pp.519-528.

Varin, C., 2008. On composite marginal likelihoods. *AStA Advances in Statistical Analysis*, 92(1), p.1.

Velikova, M. and Daniels, H., 2004. Decision trees for monotone price models. *Computational Management Science*, 1, pp.231-244.

Venables, W.N. and Ripley, B.D., 2013. *Modern applied statistics with S-PLUS*. Springer Science & Business Media.

Vij, A., Carrel, A. and Walker, J.L., 2013. Incorporating the influence of latent modal preferences on travel mode choice behavior. *Transportation Research Part A: Policy and Practice*, 54, pp.164-178.

von Thunen, J. H. 1826. *Der isolierte Staat in beziehung auf Landwirtscha und Nationalokonomie*, Gustav Fisher, Stuttgart. English edition: *The isolated state* (trans: Wartenburg CM (1966), edited by Hall P). Oxford: Pergamon.

Waddell, P. 2009. *Parcel-Level Microsimulation of Land Use and Transportation: The Walking Scale of Urban Sustainability*. 12th International Conference on Travel Behavior Research. Jaipur, India.

Waddell, P., 1992. A multinomial logit model of race and urban structure. *Urban Geography*, 13(2), pp.127-141.

Waddell, P., Borning, A., Noth, M., Freier, N., Becke, M. and Ulfarsson, G., 2003.

Microsimulation of urban development and location choices: Design and implementation of UrbanSim. *Networks and spatial economics*, 3(1), pp.43-67.

Wang, D. and Lin, T., 2019. Built environment, travel behavior, and residential self-selection: A study based on panel data from Beijing, China. *Transportation*, 46(1), pp.51-74.

Wang, Y. and Wu, L., 2010, June. Integrated land-use and transportation models. In 2010 18th International Conference on Geoinformatics (pp. 1-4). IEEE.

Wehenkel, L. and Olaru, C., 2003. A complete fuzzy decision tree technique. *Fuzzy sets and systems*, 138(2), pp.221-254.

Wen, C. H., & Chen, P. H. (2017). Passenger booking timing for low-cost airlines: A continuous logit approach. *Journal of Air Transport Management*, 64, 91-99.

Williams, H.C.W.L. and Ortúzar, J.D.D., 1982. Behavioural theories of dispersion and the mis-specification of travel demand models. *Transportation Research Part B: Methodological*, 16(3), pp.167-219.

Wilson A. G., 1970, *Entropy in Urban and Regional Modelling* (Pion, London).

Wilson, A.G. 1970. *Entropy in Urban and Regional Modelling*. London, Pion.

Wilson, A.G. 1971. A Family of Spatial Interaction Models and Associated Developments *Environment and Planning* 3: 1-32.

Wolday, F., Cao, J. and Næss, P., 2018. Examining factors that keep residents with high transit preference away from transit-rich zones and associated behavior outcomes. *Journal of Transport Geography*, 66, pp.224-234.

Wu, F. (2004) Intraurban residential relocation in Shanghai: modes and stratification, *Environment and Planning A*, 36, pp. 7–25.

Wu, H. and Levinson, D., 2020. Unifying access. *Transportation Research Part D: Transport and Environment*, 83, p.102355.

Wu, H., and Levinson, D. (2020). Unifying access. *Transportation Research Part D: Transport and Environment*, 83, 102355. <https://doi.org/10.1016/j.trd.2020.102355>

Xi, Y. (Luna), Miller, E. J., and Saxe, S. (2018). Exploring the Impact of Different Cut-off Times on Isochrone Measurements of Accessibility. *Transportation Research Record: Journal of the Transportation Research Board*, 2672(49), 113–124. <https://doi.org/10.1177/0361198118783113>

Xi, Y., Miller, E.J. and Saxe, S., 2018. Exploring the impact of different cut-off times on isochrone measurements of accessibility. *Transportation Research Record*, 2672(49), pp.113-124.

Xu, M., Meng, Q., Liu, K. and Yamamoto, T., 2017. Joint charging mode and location choice model for battery electric vehicle users. *Transportation Research Part B: Methodological*, 103, pp.68-86.

Yang, L., Liang, Y., Zhu, Q. and Chu, X., 2021. Machine learning for inference: Using gradient boosting decision tree to assess non-linear effects of bus rapid transit on house prices. *Annals of GIS*, 27(3), pp.273-284.

Zarabi, Z., Manaugh, K. and Lord, S., 2019. The impacts of residential relocation on commute habits: A qualitative perspective on households' mobility behaviors and strategies. *Travel Behaviour and Society*, 16, pp.131-142.

Zarabi, Z., Manaugh, K. and Lord, S., 2019. The impacts of residential relocation on commute habits: A qualitative perspective on households' mobility behaviours and strategies. *Travel Behaviour and Society*, 16, pp.131-142.

Zegras, P. C., Li, M., Kilic, T., Lozano-Gracia, N., Ghorpade, A., Tiberti, M., & Zhao,

F. (2018). Assessing the representativeness of a smartphone-based household travel survey in Dar es Salaam, Tanzania. *Transportation*, 1-29.

Zeileis, A., Hothorn, T. and Hornik, K., 2008. Model-based recursive partitioning. *Journal of Computational and Graphical Statistics*, 17(2), pp.492-514.

Zhang, L., Hong, J., Nasri, A. and Shen, Q., 2012. How built environment affects travel behavior: A comparative analysis of the connections between land use and vehicle miles traveled in US cities. *Journal of transport and land use*, 5(3), pp.40-52.

Zhang, L., Hong, J., Nasri, A. and Shen, Q., 2012. How built environment affects travel behaviour: A comparative analysis of the connections between land use and vehicle miles traveled in US cities. *Journal of transport and land use*, 5(3), pp.40-52.

Zhu, W. and Timmermans, H., 2010. Cognitive process model of individual choice behaviour incorporating principles of bounded rationality and heterogeneous decision heuristics. *Environment and Planning B: Planning and Design*, 37(1), pp.59-74.

Zolfaghari, A., Polak, J. and Sivakumar, A., 2016. Choice set imputation in atomistic spatial choice models. *Transportation Research Record*, 2564(1), pp.138-146.

Appendix A

Appendix: Chapter 3

A.1 PROBABILISTIC PREDICTIONS

Individuals with the same values for their explanatory variables may make different choices, so models of decision-making need to be able to produce probabilistic predictions. Fortunately, decision trees can and often do make probabilistic predictions in their output nodes, with the probability of a given alternative being predicted to be the fraction of observations in that output node who chose the alternative in question. (Arentze and Timmermans, 2004; Strobl et al., 2009). There are two observations that economically motivate the move from deterministic outputs to probabilistic outputs. First, individuals may have probabilistic outputs in mind when using disjunctions-of-conjunctions. For instance, they may say "if any of these conjunctive conditions are met, then it is highly likely that I will choose a certain alternative." In this case, the estimated decision tree will estimate what 'highly likely' means for this population. Second, people violate their stated thresholds and attribute cut-offs when using non-compensatory protocols such as conjunctive and disjunctive rules (Green et al., 1988; Huber and Klein, 1991; Swait, 2001). This means that even if an individual consciously operates as if satisfaction of some set of conjunctive conditions will result in a deterministic outcome, there is still some probability that they may choose another alternative because they are violating their own conditions.

A.2 HETEROGENEITY IN A POPULATION'S NON-COMPENSATORY RULES

When describing human behaviour, it is often unreasonable to expect that all individuals in a population will use exactly the same non-compensatory rules. In the context of residential location choice, the presence of heterogeneity in the split point for an explanatory variable can be classified as local heterogeneity. This means that even within a homogeneous population, different individuals may have different preferences or decision-making thresholds, which results in different split points for the same explanatory variable. Local heterogeneity is fully addressed by 'soft' decision trees, which place a probability distribution over the splitting point of each continuous explanatory variable (Quinlan, 1990; Wehenkel, 2003). However, in some cases, the structure of the decision tree may also differ across individuals. This is called global heterogeneity, where the tree structure itself varies among different subpopulations, indicating the presence of more fundamental differences in decision-making processes. Decision tree ensembles, such as random forests (Breiman, 2001) or boosted trees (Bühlmann and Hothorn, 2007), cater to global heterogeneity concerns. These ensembles represent global

heterogeneity in much the same way that ensembles of discrete choice models represent heterogeneity amongst the compensatory decision protocols used by varying market segments in a population (Vij et al., 2013). The main feature of decision tree ensembles is that many trees are estimated, and predictions are made by averaging the predictions of each tree in the ensemble. Additionally, decision tree ensembles account for the ensemble’s asymptotic behaviour as the number of observations used to estimate the trees approaches infinity (Minka, 2002). Asymptotically, ensembles that implement model averaging reduce to the estimation of a single tree, while ensembles that implement model combination still estimate multiple, distinct decision trees. Decision tree ensembles, such as Bayesian decision trees and ‘bagging’ lead to the estimation of a single tree, which we interpret as catering for estimation uncertainty. Although latent class models and decision tree ensemble methods share similarities, they differ in implementation. Latent class models estimate a relatively small number of classes, while ensemble methods usually result in models with hundreds of decision trees. However, having many trees makes sense behaviourally, as the disjunctions-of-conjunctions used by individuals can differ in many ways, leading to many different decision trees used by different people in a population.

A.3 CONTEXT-DEPENDENT PREFERENCE HETEROGENEITY

The importance of the decision-making context as a determinant of outcomes has been recognized in the discrete choice literature and broader human decision-making literature (Swait et al., 2002). The concept of context-dependent preference heterogeneity refers to the idea that the choice context may affect an individual’s preferences or sensitivities to a given set of explanatory variables. For example, an individual’s preference for a particular neighbourhood might depend on whether they are looking for a place to rent or buy. If the individual is looking to buy, they may be more sensitive to the quality of local schools or the presence of nearby parks and recreational facilities, whereas if they are looking to rent, they may be more sensitive to factors such as the cost of rent or the availability of public transportation. Another example of context-dependent preference heterogeneity for residential location choice context is based on the individual’s life stage. A family with young children may prioritize access to good schools and child-friendly amenities, while a retiree may value quietness and proximity to healthcare facilities. Additionally, an individual’s work location and commute preferences can also influence their preferred residential location. Someone who values a short commute may prioritize living in a particular area closer to their workplace, while someone who prefers more space may be willing to live further away from their workplace. While a piecewise linear function may be sufficient for simple scenarios, more complex conditions may require accommodating methods that traditional discrete choice literature does not offer.

However, the literature on decision tree methods provides a solution to account for context-dependent preference heterogeneity. Model trees, a variant of decision trees, can address this phenomenon by using a statistical model as output at a given node (Chan and Loh, 2004; Landwehr et al., 2005; Zeilis et al., 2008). The decision tree is used to determine the output node for a given observation, and the output node’s statistical model provides the final outcome probabilities. In the case of discrete choice models used in output nodes, preference heterogeneity is represented by differing systematic utility functions in the models used in different nodes. For instance, a decision tree that is split on the price or cost variable may have different systematic utility functions for the low and high travel cost output nodes,

with different coefficients for the travel cost variable, representing the different preferences in the two contexts. Moreover, many decision tree and discrete choice methods can be viewed as special cases of model trees. For example, standard decision trees can be seen as model trees with discrete choice models such as the MNL used in each node, and each alternative's systematic utility comprising only an alternative specific constant (ASC). Other proposed models, such as the parametric-action decision tree, the hybrid CART-logit model, the tree-augmented logistic model, and the two-stage MNL model, estimate a decision tree and then incorporate a dummy variable for each output node into one's systematic utility functions. The semi-compensatory models used in discrete choice literature are also special cases of model trees, where parameters of the systematic utility function for available alternatives are constrained to be equal across output nodes, and output nodes that result in a given alternative not being available simply set the systematic utility for that alternative to negative infinity.

A.4 UNCERTAINTY IN ESTIMATION

In statistical applications, it is often important to quantify inferential uncertainty. For models with continuous parameters, uncertainty can be quantified using the sampling distribution of the estimator. However, decision trees are made up of discrete parameters such as the depth of the tree, the variables on which the tree is split, and the values of those variables. Therefore, in discrete settings, uncertainty is quantified by the probability of a given combination of parameters generating the data. However, since it is typically not feasible to enumerate all possible decision trees, approximations must be made to quantify this uncertainty (Chipman et al., 1998). One approach to estimate the uncertainty is to use ensemble methods, such as Bayesian decision trees and bagging. Bayesian methods estimate posterior probabilities of parameter values, providing the desired uncertainty quantification. Meanwhile, the bootstrap is equivalent to a Bayesian analysis with a specific prior, making it useful for estimating uncertainty using bagging (Rubin, 1981; Newton and Raftery, 1994). By taking the fraction of times a particular decision tree appears in the ensemble, we can estimate the probability that this decision tree is the correct one. However, it is important to note that these ensemble methods are approximate measures of estimation uncertainty since they may not include all possible decision trees (Chipman et al., 1998).

A.5 MONOTONICITY CONSTRAINTS

Decision models are often bound by economic constraints, such as the monotonicity constraint. This constraint means that, all else being equal, as the price of a normal good increases, the probability of choosing that good should decrease or remain the same. To comply with this constraint, linear-in-parameter systematic utility functions in discrete choice models use sign constraints on model coefficients. Noncompliance with economic theory is unlikely in discrete choice models. To ensure compliance, monotonic decision trees can be used, which have been developed by researchers (Potharst and Feelders, 2002; Velikova and Daniels, 2004; Marsala and Petturiti, 2015; Pei et al., 2016). Monotonic decision trees ensure that the desired monotonicity constraints are not violated during the estimation process, making them ideal for estimating the disjunctions-of-conjunctions used in a population. Monotonic decision trees guarantee compliance with economic theory, which will be beneficial for choice modellers who want to ensure that their decision trees are sensible.

Appendix B

Appendix: Chapter 4

The household structure models for the **Chapter 4: Simultaneity of household trip frequencies and accessibility** are estimated independently for the following household structure types:

1. single adult with no kids
2. couple adult with no kids
3. single or couple with one or more kids
4. multi-adult with no kids
5. multi-adult with one or more kids

TABLE B.1 Single adult and no kids: Exogenous Variable Effects on Modal trip frequencies

Variable definitions	Auto Trips		Transit Trips		Bicycle Trips		Walk Trips	
	Estimate	z value	Estimate	z value	Estimate	z value	Estimate	z value
<i>(Base: Age category 16-25 years)</i>								
Presence in age category 26-35 years	—	—	—	—	—	—	−0.3006	−4.01
Presence in age category 36-45 years	—	—	—	—	—	—	−0.4593	−5.81
Presence in age category 46-55 years	—	—	—	—	—	—	−0.4606	−5.79
Presence in age category 56-65 years	—	—	—	—	—	—	−0.4837	−6.14
Presence in age category 66 and above years	—	—	—	—	−0.1949	−4.20	−0.6933	−8.45
Presence of students in the household	—	—	—	—	—	—	0.1698	2.82
Distance to CBD (km)	—	—	—	—	0.0109	2.70	−0.0492	−11.98
Population Density (Standardized value)	—	—	—	—	—	—	—	—
Employment Density (Standardized value)	−0.0118	−1.41	—	—	—	—	0.0513	5.14
<i>(Base: Income less than 15,000\$)</i>								
Income 15,000\$-40,000\$	—	—	—	—	—	—	—	—
Income 40,000\$ to 60,000\$	—	—	—	—	—	—	0.1135	2.71
Income 60,000\$ to 100,000\$	—	—	—	—	0.0643	1.44	0.2314	5.68
Income 100,000\$ to 125,000\$	—	—	—	—	—	—	0.2702	4.59
Income greater than 125,000\$	—	—	—	—	—	—	0.3531	5.93
<i>(Base: trip day Friday)</i>								
Monday to Thursday as trip day	—	—	—	—	—	—	0.0574	1.77
<i>(Base: fraction of male adults in the household)</i>								
fraction of male adults in the household	—	—	—	—	—	—	—	—

TABLE B.2 Single adult and no kids: Exogenous Variable Effects on Modal Accessibilities

Variable definitions	Auto Accessibility		Transit Accessibility		Bicycle Accessibility		Walk Accessibility	
	Estimate	z value	Estimate	z value	Estimate	z value	Estimate	z value
<i>(Base: Age category 16-25 years)</i>								
Presence in age category 26-35 years	-0.1273	-4.72	-0.2613	-3.81	-0.1153	-4.53	—	—
Presence in age category 36-45 years	-0.1238	-4.35	-0.2709	-3.88	—	—	—	—
Presence in age category 46-55 years	—	—	-0.3034	-4.36	—	—	—	—
Presence in age category 56-65 years	—	—	-0.2561	-3.74	—	—	—	—
Presence in age category 66 and above years	—	—	-0.2199	-3.28	—	—	—	—
Presence of students in the household	0.1061	2.74	0.1361	2.66	—	—	—	—
Distance to CBD (km)	-0.0220	-11.76	-0.1990	-83.37	-0.0830	-42.98	0.0168	9.50
Population Density (Standardized value)	0.2561	27.20	0.2268	24.43	-0.0686	-8.11	0.0358	4.03
Employment Density (Standardized value)	-2.2480	-124.02	0.4861	48.85	0.0969	11.52	-0.1453	-16.89
<i>(Base: Income less than 15,000\$)</i>								
Income 15,000\$-40,000\$	—	—	—	—	0.1065	3.29	—	—
Income 40,000\$ to 60,000\$	—	—	—	—	0.1605	4.66	—	—
Income 60,000\$ to 100,000\$	-0.1588	-6.66	—	—	0.1739	5.11	-0.1415	-6.28
Income 100,000\$ to 125,000\$	-0.1857	-4.69	—	—	0.1463	3.25	-0.1936	-4.99
Income greater than 125,000\$	-0.2200	-5.09	0.1169	2.99	0.2092	4.63	-0.2194	-5.44
<i>(Base: Vehicle availability per licensed adult is zero)</i>								
Vehicle availability per licensed adult is one or more	-0.0913	-4.25	-0.1824	-8.65	-0.0629	-3.08	-0.0392	-1.96
<i>(Base: Transit pass ownership zero per adult in the household)</i>								
Transit pass ownership one or more per adult in the household	-0.0394	-1.89	—	—	—	—	—	—
<i>(Base: dwelling type house or townhouse)</i>								
Apartment type dwelling	-0.5602	-24.65	—	—	—	—	-0.4927	-22.68

TABLE B.3 Single adult and no kids: Endogenous Variable Effects

Variable definitions	Auto Trips		Transit Trips		Bicycle Trips		Walk Trips	
	Estimate	z value	Estimate	z value	Estimate	z value	Estimate	z value
<i>(Base case: Drive Access Very Low)</i>								
Drive Access Low	—	—	—	—	0.1537	3.37	—	—
Drive Access Medium	—	—	—	—	—	—	—	—
Drive Access High	—	—	—	—	—	—	—	—
Drive Access Very High	—	—	—	—	—	—	—	—
<i>(Base case: Transit Access Very Low)</i>								
Transit Access Low	—	—	—	—	0.1237	2.14	—	—
Transit Access Medium	—	—	—	—	—	—	—	—
Transit Access High	—	—	—	—	0.1268	2.56	0.1246	2.60
Transit Access Very High	—	—	—	—	—	—	0.3114576	5.66
<i>(Base case: Bike Access Very Low)</i>								
Bike Access Low	—	—	—	—	—	—	—	—
Bike Access Medium	—	—	—	—	—	—	—	—
Bike Access High	—	—	—	—	—	—	−0.0723	−1.92
Bike Access Very High	—	—	—	—	—	—	−0.0572	−1.56
<i>(Base case: Walk Access Very Low)</i>								
Walk Access Low	—	—	—	—	—	—	0.2028	6.34
Walk Access Medium	—	—	—	—	—	—	—	—
Walk Access High	—	—	—	—	—	—	—	—
Walk Access Very High	—	—	—	—	—	—	—	—

TABLE B.4 Single adult and no kids: Correlation Effects and Goodness-of-fit

Variable definitions	Auto Trips		Transit Trips		Bicycle Trips		Walk Trips	
	Estimate	z value	Estimate	z value	Estimate	z value	Estimate	z value
<i>Thresholds</i>								
Intercept 1 2	0.1603	14.77	0.5124	38.24	1.9825	35.99	0.4503	4.37
Intercept 2 3	0.8865	70.80	0.6314	46.27	1.9998	36.13	0.7594	7.37
Intercept 3 4	1.6155	90.37	1.6120	85.11	2.4181	39.99	1.6307	15.51
Intercept 4 5	2.4113	68.48	2.5996	61.98	2.9227	36.67	2.4940	21.65
<i>Correlation Terms</i>								
DriveTrips	1.0000		−0.6048	−74.13	−0.3794	−17.30	−0.0839	−5.77
TransitTrips			1.0000		−0.3648	−14.06	−0.1243	−8.30
BikeTrips					1.0000		−0.2215	−10.73
WalkTrips							1.0000	
Variable definitions	Auto Accessibility		Transit Accessibility		Bicycle Accessibility		Walk Accessibility	
	Estimate	z value	Estimate	z value	Estimate	z value	Estimate	z value
<i>Thresholds</i>								
Intercept 1 2	−1.5844	−45.59	−3.7638	−51.97	−1.6796	−49.12	−1.2224	−41.77
Intercept 2 3	−0.6136	−18.28	−2.9544	−41.15	−1.0763	−32.58	−0.4501	−15.20
Intercept 3 4	0.0666	2.02	−2.0870	−29.58	−0.4879	−14.99	0.1482	5.02
Intercept 4 5	0.7600	22.68	−0.8620	−12.68	0.1963	6.10	0.7979	26.19
<i>Correlation Terms</i>								
DriveTrips	−0.0439	−4.04	—	—	—	—	−0.0407	−3.93
TransitTrips	−0.0514	−4.43	—	—	—	—	−0.0296	−2.63
BikeTrips	−0.0483	−2.30	—	—	—	—	−0.0518	−2.64
WalkTrips	—	—	−0.0233	−2.17	—	—	—	—
DriveAccess	1.0000		—	—	—	—	0.7094	223.94
TransitAccess			1.0000		0.1668	16.25	−0.0729	−7.17
BikeAccess					1.0000		—	—
WalkAccess							1.0000	
Goodness-of-fit	observations		AIC		BIC		Likelihood (mean)	
	13,613		1551927.59		1557517.20		−775220.38	

TABLE B.5 Single or couple adults with one or more kids: Exogenous Variable Effects on Modal trip frequencies

Variable definitions	Auto Trips		Transit Trips		Bicycle Trips		Walk Trips	
	Estimate	z value	Estimate	z value	Estimate	z value	Estimate	z value
<i>(Base: Age category 0-15 years)</i>								
Presence in age category 16-25 years	—	—	0.5114	6.46	—	—	—	—
Presence in age category 26-35 years	—	—	−0.0789	−2.03	—	—	−0.1552	−3.36
Presence in age category 36-45 years	—	—	—	—	—	—	0.0956	2.37
Presence in age category 46-55 years	0.1370	4.38	0.2556	7.43	—	—	0.1622	3.82
Presence in age category 56-65 years	—	—	—	—	—	—	—	—
Presence in age category 66 and above years	—	—	—	—	—	—	0.2506	2.01
Presence of students in the household	0.2023	6.07	0.3316	8.63	—	—	0.3321	8.16
Distance to CBD (km)	0.0114	3.26	0.0092	2.54	−0.1073	−17.86	−0.0317	−9.47
Population Density (Standardized value)	−0.0488	−2.50	—	—	−0.1214	−3.07	—	—
Employment Density (Standardized value)	−0.0566	−2.65	−0.0620	−3.39	—	—	0.0579	2.97
<i>(Base: Income less than 15,000\$)</i>								
Income 15,000\$-40,000\$	—	—	—	—	—	—	0.1261	2.55
Income 40,000\$ to 60,000\$	—	—	0.0897	1.58	—	—	—	—
Income 60,000\$ to 100,000\$	—	—	0.1633	3.11	0.1132	1.71	—	—
Income 100,000\$ to 125,000\$	0.1755	3.90	0.2209	3.71	—	—	—	—
Income greater than 125,000\$	0.2892	8.06	0.1297	2.42	—	—	0.1024	2.80
<i>(Base: Vehicle availability per licensed adult is zero)</i>								
Vehicle availability per licensed adult is between zero and one	1.7215	28.18	−0.5539	−10.77	−0.2064	−2.62	−0.2062	−4.00
Vehicle availability per licensed adults as one or more	2.0362	33.23	−1.0408	−20.21	−0.7252	−8.85	−0.5379	−10.32
<i>(Base: Transit pass ownership less than one per adult in the household)</i>								
Transit pass ownership one or more per adult in the household	−0.4767	−12.10	0.7720	18.89	—	—	—	—
<i>(Base: No worker household)</i>								
Single worker household	—	—	0.4550	6.72	—	—	—	—
Multi-worker household	0.1793	6.04	0.6902	9.91	—	—	—	—
<i>(Base: dwelling type house or townhouse)</i>								
Apartment type dwelling	−0.1382	−3.75	0.1231	3.38	−0.5461	−7.08	—	—

TABLE B.6 Single or couple adults with one or more kids: Exogenous Variable Effects on Modal Accessibilities

152

Variable definitions	Auto Accessibility		Transit Accessibility		Bicycle Accessibility		Walk Accessibility	
	Estimate	z value	Estimate	z value	Estimate	z value	Estimate	z value
<i>(Base: Age category 0-15 years)</i>								
Presence in age category 16-25 years	-0.1434	-2.00	-0.1589	-1.80	—	—	—	—
Presence in age category 26-35 years	—	—	—	—	—	—	0.0557	1.86
Presence in age category 36-45 years	—	—	—	—	—	—	—	—
Presence in age category 46-55 years	—	—	0.0598	1.96	—	—	—	—
Presence in age category 56-65 years	—	—	—	—	—	—	—	—
Presence in age category 66 and above years	—	—	—	—	—	—	—	—
Presence of students in the household	—	—	—	—	—	—	0.0607	2.15
Distance to CBD (km)	-0.0305	-12.10	-0.1957	-51.79	-0.0797	-28.14	0.0144	5.73
Population Density (Standardized value)	0.2774	17.39	0.2278	12.66	0.0422	2.81	0.0517	2.92
Employment Density (Standardized value)	-4.5830	-69.49	0.5446	30.97	0.0812	4.67	-0.1869	-13.21
<i>(Base: Income less than 15,000\$)</i>								
Income 15,000\$-40,000\$	—	—	—	—	—	—	—	—
Income 40,000\$ to 60,000\$	—	—	-0.1369	-2.48	—	—	—	—
Income 60,000\$ to 100,000\$	—	—	-0.0852	-1.73	0.1521	3.72	—	—
Income 100,000\$ to 125,000\$	0.0622	1.73	-0.1863	-3.36	0.2569	5.25	—	—
Income greater than 125,000\$	—	—	-0.1798	-3.68	0.3555	8.92	—	—
<i>(Base: Vehicle availability per licensed adult is zero)</i>								
Vehicle availability per licensed adult is between zero and one	—	—	—	—	-0.1083	-3.89	—	—
Vehicle availability per licensed adults as one or more	—	—	-0.2032	-6.93	—	—	—	—
<i>(Base: Transit pass ownership less than one per adult in the household)</i>								
Transit pass ownership one or more per adult in the household	—	—	-0.0751	-1.85	-0.1077	-2.90	—	—
<i>(Base: dwelling type house or townhouse)</i>								
Apartment type dwelling	-0.3682	-10.68	-0.0758	-2.06	-0.2031	-5.90	-0.3865	-10.95

Appendix B. Appendix: Chapter 4

TABLE B.7 Single or couple adults with one or more kids: Endogenous Variable Effects

Variable definitions	Auto Trips		Transit Trips		Bicycle Trips		Walk Trips	
	Estimate	z value	Estimate	z value	Estimate	z value	Estimate	z value
<i>(Base case: Drive Access Very Low)</i>								
Drive Access Low	—	—	—	—	—	—	—	—
Drive Access Medium	—	—	—	—	—	—	0.1205	3.22
Drive Access High	0.1272	2.88	—	—	—	—	—	—
Drive Access Very High	—	—	—	—	—	—	—	—
<i>(Base case: Transit Access Very Low)</i>								
Transit Access Low	—	—	—	—	—	—	—	—
Transit Access Medium	—	—	—	—	—	—	—	—
Transit Access High	−0.1847	−4.48	0.1298	3.02	—	—	—	—
Transit Access Very High	−0.2574	−5.03	0.1336	2.54	0.1724	3.57	—	—
<i>(Base case: Bike Access Very Low)</i>								
Bike Access Low	—	—	—	—	—	—	—	—
Bike Access Medium	—	—	—	—	—	—	—	—
Bike Access High	—	—	—	—	—	—	—	—
Bike Access Very High	0.0662	1.79	—	—	—	—	0.1098	2.72
<i>(Base case: Walk Access Very Low)</i>								
Walk Access Low	—	—	—	—	−0.3357	−3.85	—	—
Walk Access Medium	—	—	—	—	−0.1736	−2.30	—	—
Walk Access High	−0.1187	−2.68	—	—	—	—	—	—
Walk Access Very High	—	—	—	—	—	—	—	—

TABLE B.8 Single or couple adults with one or more kids: Correlation Effects and Goodness-of-fit

Variable definitions	Auto Trips		Transit Trips		Bicycle Trips		Walk Trips	
	Estimate	z value	Estimate	z value	Estimate	z value	Estimate	z value
<i>Thresholds</i>								
Intercept 1 2	1.1966	14.31	0.5299	5.55	—	—	0.2109	2.70
Intercept 2 3	2.0196	24.07	0.6904	7.22	—	—	0.4059	5.18
Intercept 3 4	2.6194	31.01	1.6333	16.85	0.5940	6.91	0.9880	12.52
Intercept 4 5	3.2851	38.22	2.5227	25.40	1.1627	12.14	1.6765	20.61
<i>Correlation Terms</i>								
DriveTrips	1.0000		−0.2507	−17.92	−0.2227	−8.65	−0.1249	−7.79
TransitTrips			1.0000		−0.2798	−12.05	−0.1081	−6.33
BikeTrips					1.0000		—	—
WalkTrips							1.0000	
Variable definitions	Auto Accessibility		Transit Accessibility		Bicycle Accessibility		Walk Accessibility	
	Estimate	z value	Estimate	z value	Estimate	z value	Estimate	z value
<i>Thresholds</i>								
Intercept 1 2	−0.7790	−21.16	−3.7166	−51.01	−1.7463	−33.91	−0.7575	−20.94
Intercept 2 3	−0.1003	−2.74	−2.8147	−40.56	−1.0691	−21.25	−0.3705	−10.16
Intercept 3 4	0.5966	16.60	−1.8248	−27.73	−0.5061	−10.17	0.1431	3.94
Intercept 4 5	1.3850	37.22	−0.7112	−11.19	0.1950	3.94	0.8383	22.40
<i>Correlation Terms</i>								
DriveTrips	—	—	—	—	0.0192	1.68	—	—
TransitTrips	—	—	—	—	—	—	—	—
BikeTrips	—	—	—	—	—	—	—	—
WalkTrips	—	—	—	—	—	—	—	—
DriveAccess	1.0000		—	—	—	—	0.7725	168.46
TransitAccess			1.0000		0.0915	6.27	−0.1282	−8.77
BikeAccess					1.0000		—	—
WalkAccess							1.0000	—
Goodness-of-fit	observations		AIC		BIC		Likelihood (mean)	
	6,526		866000.98		872004.35		−432115.50	

TABLE B.9 Multi adults with no kids: Exogenous Variable Effects on Modal trip frequencies

Variable definitions	Auto Trips		Transit Trips		Bicycle Trips		Walk Trips	
	Estimate	z value	Estimate	z value	Estimate	z value	Estimate	z value
<i>(Base: Age category 0-15 years)</i>								
Presence in age category 16-25 years	0.0937	2.47	0.2013	4.83	—	—	—	—
Presence in age category 26-35 years	—	—	—	—	—	—	—	—
Presence in age category 36-45 years	—	—	—	—	—	—	0.1169	2.04
Presence in age category 46-55 years	—	—	—	—	−0.1830	−2.40	—	—
Presence in age category 56-65 years	0.0722	2.44	−0.2112	−2.56	—	—	−0.1558	−3.57
Presence in age category 66 and above years	—	—	−0.0949	−2.50	−0.3613	−4.03	−0.2219	−4.45
Presence of students in the household	0.0861	2.29	0.4667	12.13	—	—	0.2818	6.33
Distance to CBD (km)	0.0101	2.65	−0.0113	−3.78	−0.0652	−7.87	−0.0311	−6.45
Population Density (Standardized value)	—	—	—	—	−0.1818	−2.43	—	—
Employment Density (Standardized value)	—	—	−0.0785	−3.20	—	—	0.1124	3.64
<i>(Base: Income less than 15,000\$)</i>								
Income 15,000\$-40,000\$	—	—	—	—	—	—	—	—
Income 40,000\$ to 60,000\$	0.1159	2.27	—	—	—	—	—	—
Income 60,000\$ to 100,000\$	0.2349	4.96	0.1074	2.88	0.2481	2.28	—	—
Income 100,000\$ to 125,000\$	0.3655	6.63	—	—	0.2657	2.02	—	—
Income greater than 125,000\$	0.5499	10.92	0.0695	1.81	0.4374	3.92	—	—
<i>(Base: fraction of male adults in the household)</i>								
fraction of female adults in the household	—	—	0.2253	3.03	−0.5511	−3.03	—	—
<i>(Base: Vehicle availability per licensed adult is zero)</i>								
Vehicle availability per licensed adult is between zero and one	1.7489	24.58	−0.7648	−13.26	−0.4690	−4.36	−0.3625	−5.47
Vehicle availability per licensed adults as one or more	1.8161	24.75	−1.0752	−17.90	−0.8886	−7.08	−0.5787	−8.03
<i>(Base: Transit pass ownership less than one per adult in the household)</i>								
Transit pass ownership one or more per adult in the household	−0.5758	−8.14	0.7991	11.79	—	—	−0.3427	−3.69
<i>(Base: No worker household)</i>								
Single worker household	—	—	0.4041	5.61	—	—	—	—
Multi-worker household	0.1823	5.25	0.7590	11.15	—	—	0.1873	5.71
<i>(Base: dwelling type house or townhouse)</i>								
Apartment type dwelling	−0.1850	−5.01	0.1391	3.70	−0.3506	−3.05	—	—

TABLE B.10 Multi adults with no kids: Exogenous Variable Effects on Modal Accessibilities

156

Variable definitions	Auto Accessibility		Transit Accessibility		Bicycle Accessibility		Walk Accessibility	
	Estimate	z value	Estimate	z value	Estimate	z value	Estimate	z value
<i>(Base: Age category 0-15 years)</i>								
Presence in age category 16-25 years	—	—	—	—	—	—	—	—
Presence in age category 26-35 years	—	—	—	—	—	—	—	—
Presence in age category 36-45 years	—	—	—	—	—	—	0.0669	1.94
Presence in age category 46-55 years	—	—	—	—	—	—	—	—
Presence in age category 56-65 years	—	—	—	—	—	—	—	—
Presence in age category 66 and above years	—	—	0.0923	2.83	—	—	—	—
Presence of students in the household	—	—	—	—	—	—	—	—
Distance to CBD (km)	−0.0308	−11.31	−0.1785	−46.15	−0.0946	−34.32	0.0114	4.42
Population Density (Standardized value)	0.2894	17.12	0.2676	14.01	—	—	0.0327	1.87
Employment Density (Standardized value)	−4.5559	−58.60	0.7652	28.06	0.1038	4.72	−0.2364	−14.33
<i>(Base: Income less than 15,000\$)</i>								
Income 15,000\$-40,000\$	—	—	—	—	—	—	—	—
Income 40,000\$ to 60,000\$	—	—	−0.0867	−2.06	—	—	—	—
Income 60,000\$ to 100,000\$	—	—	—	—	—	—	—	—
Income 100,000\$ to 125,000\$	—	—	—	—	—	—	—	—
Income greater than 125,000\$	—	—	—	—	0.2238	6.4497	—	—
<i>(Base: fraction of male adults in the household)</i>								
fraction of female adults in the household	—	—	0.1694	2.16	—	—	—	—
<i>(Base: Vehicle availability per licensed adult is zero)</i>								
Vehicle availability per licensed adult is between zero and one	—	—	—	—	—	—	—	—
Vehicle availability per licensed adults as one or more	—	—	−0.0699	−2.22	—	—	—	—
<i>(Base: Transit pass ownership less than one per adult in the household)</i>								
Transit pass ownership one or more per adult in the household	—	—	—	—	—	—	—	—
<i>(Base: No worker household)</i>								
Single worker household	—	—	—	—	—	—	—	—
Multi-worker household	−0.0591	−2.13	—	—	−0.0408	−1.24	—	—
<i>(Base: dwelling type house or townhouse)</i>								
Apartment type dwelling	−0.3361	−7.82	−0.0778	−1.96	−0.2119	−6.27	−0.3266	−7.75

Appendix B, Appendix: Chapter 4

TABLE B.11 Multi adults with no kids: Endogenous Variable Effects

Variable definitions	Auto Trips		Transit Trips		Bicycle Trips		Walk Trips	
	Estimate	z value	Estimate	z value	Estimate	z value	Estimate	z value
<i>(Base case: Drive Access Very Low)</i>								
Drive Access Low	—	—	—	—	—	—	0.2029	3.04
Drive Access Medium	—	—	—	—	—	—	0.1415	2.07
Drive Access High	—	—	—	—	—	—	—	—
Drive Access Very High	—	—	0.0619	1.76	—	—	—	—
<i>(Base case: Transit Access Very Low)</i>								
Transit Access Low	−0.1317	−3.29	0.0919	2.39	—	—	—	—
Transit Access Medium	−0.1821	−4.11	0.1199	2.95	—	—	—	—
Transit Access High	−0.2937	−4.72	0.1090	2.08	—	—	0.1719	2.48
Transit Access Very High	−0.4822	−6.49			0.2759	2.62	0.4256493	5.12
<i>(Base case: Bike Access Very Low)</i>								
Bike Access Low	—	—	—	—	—	—	—	—
Bike Access Medium	—	—	—	—	−0.1991	−1.93	—	—
Bike Access High	—	—	—	—	—	—	0.1314	2.59
Bike Access Very High	—	—	—	—	—	—	—	—
<i>(Base case: Walk Access Very Low)</i>								
Walk Access Low	−0.1627	−3.51	—	—	—	—	—	—
Walk Access Medium	—	—	—	—	—	—	−0.1617	−2.32
Walk Access High	—	—	—	—	—	—	—	—
Walk Access Very High	—	—	—	—	—	—	—	—

TABLE B.12 Multi adults with no kids: Correlation Effects and Goodness-of-fit

Variable definitions	Auto Trips		Transit Trips		Bicycle Trips		Walk Trips	
	Estimate	z value	Estimate	z value	Estimate	z value	Estimate	z value
<i>Thresholds</i>								
Intercept 1 2	1.0069	9.61	—	—	—	—	0.4553	4.22
Intercept 2 3	1.8277	17.19	—	—	—	—	0.6905	6.41
Intercept 3 4	2.4441	22.81	0.9750	9.76	0.8981	4.39	1.4981	13.55
Intercept 4 5	3.0969	28.64	1.8175	17.89	1.4040	6.96	2.0735	17.93
<i>Correlation Terms</i>								
DriveTrips	1.0000		−0.2817	−19.77	−0.1971	−5.94	−0.0724	−3.34
TransitTrips			1.0000		−0.2030	−6.32	−0.0954	−4.46
BikeTrips					1.0000		−0.2189	−7.16
WalkTrips							1.0000	
Variable definitions	Auto Accessibility		Transit Accessibility		Bicycle Accessibility		Walk Accessibility	
	Estimate	z value	Estimate	z value	Estimate	z value	Estimate	z value
<i>Thresholds</i>								
Intercept 1 2	−0.6664	−13.33	−3.4865	−48.59	−2.0532	−41.63	−0.7619	−19.18
Intercept 2 3	−0.1113	−2.21	−2.4184	−35.92	−1.3183	−27.71	−0.3997	−9.98
Intercept 3 4	0.4927	9.97	−1.3178	−21.13	−0.7082	−15.22	0.1135	2.85
Intercept 4 5	1.1977	23.84	−0.2670	−4.42	0.0822	1.80	0.7322	17.92
<i>Correlation Terms</i>								
DriveTrips	—	—	—	—	—	—	—	—
TransitTrips	—	—	—	—	—	—	−0.0436	−3.21
BikeTrips	0.1489	4.26	—	—	—	—	−0.1409	−4.18
WalkTrips	0.0480	2.14	—	—	—	—	−0.0365	−1.71
DriveAccess	1.0000		—	—	—	—	0.8181	181.47
TransitAccess			1.0000		—	—	−0.1105	−6.95
BikeAccess					1.0000		−0.0380	−2.47
WalkAccess							1.0000	—
Goodness-of-fit	observations		AIC		BIC		Likelihood (mean)	
	5,660		715207.36		721050.56		−356723.83	

TABLE B.13 Multi adults with one or more kids: Exogenous Variable Effects on Modal trip frequencies

Variable definitions	Auto Trips		Transit Trips		Bicycle Trips		Walk Trips	
	Estimate	z value	Estimate	z value	Estimate	z value	Estimate	z value
<i>(Base: Age category 0-15 years)</i>								
Presence in age category 16-25 years	0.1240	2.15	0.3998	6.51	—	—	0.1186	1.82
Presence in age category 26-35 years	—	—	—	—	−0.5448	−3.20	−0.1497	−2.01
Presence in age category 36-45 years	—	—	—	—	—	—	—	—
Presence in age category 46-55 years	0.2095	3.85	—	—	—	—	—	—
Presence in age category 56-65 years	—	—	—	—	—	—	—	—
Presence in age category 66 and above years	—	—	−0.1773	−2.86	—	—	—	—
Presence of students in the household	0.1718	2.04	0.5612	5.94	—	—	0.5141	4.86
Distance to CBD (km)	0.0273	5.97	−0.0193	−3.91	−0.0965	−9.70	—	—
Population Density (Standardized value)	—	—	—	—	—	—	—	—
Employment Density (Standardized value)	—	—	−0.2868	−2.79	—	—	—	—
<i>(Base: Income less than 15,000\$)</i>								
Income 15,000\$-40,000\$	—	—	0.1772	1.97	—	—	—	—
Income 40,000\$ to 60,000\$	—	—	0.2369	2.72	—	—	—	—
Income 60,000\$ to 100,000\$	—	—	0.3162	4.01	—	—	—	—
Income 100,000\$ to 125,000\$	0.3120	3.87	—	—	—	—	—	—
Income greater than 125,000\$	0.4804	7.89	0.1649	2.05	—	—	—	—
<i>(Base: fraction of male adults in the household)</i>								
fraction of female adults in the household	—	—	0.3776	2.52	−1.0723	−3.23	—	—
<i>(Base: Vehicle availability per licensed adult is zero)</i>								
Vehicle availability per licensed adult is between zero and one	1.7255	12.72	−0.8008	−7.49	—	—	—	—
Vehicle availability per licensed adults as one or more	1.9199	13.89	−1.1095	−10.14	—	—	−0.1334	−2.40
<i>(Base: Transit pass ownership less than one per adult in the household)</i>								
Transit pass ownership one or more per adult in the household	−0.7694	−5.25	1.0604	7.34	—	—	−0.3992	−2.65
<i>(Base: No worker household)</i>								
Single worker household	—	—	—	—	—	—	—	—
Multi-worker household	—	—	0.3100	5.06	—	—	—	—

TABLE B.14 Multi adults with one or more kids: Exogenous Variable Effects on Modal Accessibilities

160

Variable definitions	Auto Accessibility		Transit Accessibility		Bicycle Accessibility		Walk Accessibility	
	Estimate	z value	Estimate	z value	Estimate	z value	Estimate	z value
<i>(Base: Age category 0-15 years)</i>								
Presence in age category 16-25 years	—	—	—	—	—	—	—	—
Presence in age category 26-35 years	—	—	—	—	−0.1528	−2.71	—	—
Presence in age category 36-45 years	—	—	—	—	—	—	—	—
Presence in age category 46-55 years	—	—	—	—	—	—	—	—
Presence in age category 56-65 years	—	—	—	—	—	—	—	—
Presence in age category 66 and above years	—	—	—	—	—	—	—	—
Presence of students in the household	—	—	—	—	—	—	—	—
Distance to CBD (km)	−0.0361	−9.03	−0.1799	−27.56	−0.0938	−19.94	—	—
Population Density (Standardized value)	0.3738	17.40	0.2148	5.74	—	—	—	—
Employment Density (Standardized value)	−6.3324	−41.63	1.1796	7.64	0.6777	5.33	−0.2968	−11.84
<i>(Base: Income less than 15,000\$)</i>								
Income 15,000\$-40,000\$	—	—	—	—	—	—	—	—
Income 40,000\$ to 60,000\$	—	—	—	—	—	—	—	—
Income 60,000\$ to 100,000\$	—	—	—	—	—	—	—	—
Income 100,000\$ to 125,000\$	0.1364	1.81	—	—	—	—	0.2583	3.47
Income greater than 125,000\$	—	—	—	—	—	—	—	—
<i>(Base: fraction of male adults in the household)</i>								
fraction of female adults in the household	−0.2705	−1.79	—	—	—	—	−0.3000	−2.03

TABLE B.15 Multi adults with one or mode kids: Endogenous Variable Effects

Variable definitions	Auto Trips		Transit Trips		Bicycle Trips		Walk Trips	
	Estimate	z value	Estimate	z value	Estimate	z value	Estimate	z value
<i>(Base case: Drive Access Very Low)</i>								
Drive Access Low	—	—	−0.2557	−2.73	—	—	—	—
Drive Access Medium	—	—	−0.2310	−2.55	—	—	—	—
Drive Access High	—	—	—	—	—	—	—	—
Drive Access Very High	—	—	—	—	—	—	—	—
<i>(Base case: Transit Access Very Low)</i>								
Transit Access Low	—	—	—	—	—	—	—	—
Transit Access Medium	—	—	—	—	—	—	—	—
Transit Access High	—	—	—	—	—	—	0.3658	4.52
Transit Access Very High	—	—	—	—	—	—	0.7669353	7.39
<i>(Base case: Bike Access Very Low)</i>								
Bike Access Low	—	—	—	—	0.3360	2.74	—	—
Bike Access Medium	—	—	—	—	—	—	—	—
Bike Access High	—	—	—	—	—	—	—	—
Bike Access Very High	—	—	—	—	—	—	—	—
<i>(Base case: Walk Access Very Low)</i>								
Walk Access Low	—	—	0.2252	2.40	—	—	—	—
Walk Access Medium	—	—	0.2292	2.41	—	—	—	—
Walk Access High	—	—	—	—	—	—	0.1262	2.02
Walk Access Very High	—	—	—	—	—	—	—	—

TABLE B.16 Multi adults with one or mode kids: Correlation Effects and Goodness-of-fit

Variable definitions	Auto Trips		Transit Trips		Bicycle Trips		Walk Trips	
	Estimate	z value	Estimate	z value	Estimate	z value	Estimate	z value
<i>Thresholds</i>								
Intercept 1 2	1.2463	7.86	—	—	—	—	0.8635	7.88
Intercept 2 3	2.0840	12.80	—	—	—	—	1.0312	9.38
Intercept 3 4	2.6300	16.00	0.7223	4.02	0.4390	2.16	1.6526	14.74
Intercept 4 5	3.1616	19.13	1.4851	8.22	0.8014	3.82	2.3605	20.60
<i>Correlation Terms</i>								
DriveTrips	1.0000		−0.2584	−10.44	−0.1301	−2.55	−0.1685	−6.01
TransitTrips			1.0000		−0.1690	−3.09	−0.1580	−5.60
BikeTrips					1.0000		—	—
WalkTrips							1.0000	
Variable definitions	Auto Accessibility		Transit Accessibility		Bicycle Accessibility		Walk Accessibility	
	Estimate	z value	Estimate	z value	Estimate	z value	Estimate	z value
<i>Thresholds</i>								
Intercept 1 2	−0.3671	−3.37	−3.6112	−32.43	−2.2269	−28.56	−0.9210	−10.69
Intercept 2 3	0.1815	1.68	−2.6058	−25.81	−1.5055	−20.14	−0.6176	−7.12
Intercept 3 4	0.7624	7.10	−1.4742	−16.46	−0.8468	−11.83		
Intercept 4 5	1.4820	13.62	−0.3642	−4.30			0.5060	5.83
<i>Correlation Terms</i>								
DriveTrips	—	—	—	—	—	—	−0.0510	−1.96
TransitTrips	—	—	—	—	—	—	—	—
BikeTrips	—	—	—	—	—	—	—	—
WalkTrips			—	—	—	—	0.0516	1.81
DriveAccess	1.0000		—	—	−0.0485	−1.82	0.8106	115.13
TransitAccess			1.0000		—	—	−0.0945	−3.55
BikeAccess					1.0000		−0.0500	−1.93
WalkAccess							1.0000	−1.93
Goodness-of-fit	observations		AIC		BIC		Likelihood (mean)	
	2,138		290019.88		293930.37		−144319.97	

Appendix C

Appendix: Chapter 5

First, we present results for the MNL model addressing endogeneity in residential location choice due to *natural log of price of a 2-bedroom apartment at outer postcode level* using the instruments *natural log of price of a 2-bedroom apartment at borough level* and *public transport accessibility level (PTAL) at the borough level*. This model is validated using the (modified) refutability tests. This is done in two-steps, first a regression model is estimated with outcome variable *natural log of price of a 2-bedroom apartment at outer postcode level* and regressors as instruments and all other exogenous variables available in the data. The resulting residual of this regression model is included as an exogenous variable in the choice model as a *control function*, along with the price variable and all other exogenous variables from the data.

TABLE C.1 First stage model to estimate residuals for price endogeneity

Variable definitions	Estimate	t-stat
(Intercept)	5.790937	20.022
log(postcode population density)	0.638474	47.971
log(postcode crime density)	-0.578153	-40.169
log(number of GPs)	-0.06407	-8.218
Household income >£50,000 and <£100,000	—	—
Household income >£100,000	0.048583	3.732
Ethnicity = non-white	—	—
log(PTAL at outer postcode level)	0.04093	4.052
log(HH Vehicle ownership + 1)	-0.054943	-2.919
<i>(Interaction: log(crime density (per km²)) *)</i>		
<i>(Base: Ethnicity = white)</i>		
Ethnicity = non-white	-0.037058	-2.971

Notes: "—" signifies statistically not significant; log of variables denote the natural logarithm to the base 'e' = 2.7182.

TABLE C.2 Control function (for price endogeneity) choice model result

Variable definitions	Multinomial Logit Estimate	t-stat
<i>(Alternative-specific attributes)</i>		
log(price of a 2-bedroom apartment)	-0.188635	-3.65771
log(number of GPs)	0.156003	5.53247
log(postcode population density)	0.564581	7.99435
log(postcode crime density)	—	—
log(vehicle miles travelled)	3.136318	7.5915
<i>(Interaction: log(price of a 2-bedroom apartment) *)</i> <i>(Base: Household income < £50,000)</i>		
Household income > £50,000 and < £100,000	0.113657	2.28833
Household income > £100,000	—	—
<i>(Base: Household vehicles = 0)</i>		
Households vehicles 1	8.294899	9.9109
Households vehicles 2	7.857882	7.13942
Households vehicles 3 or more	6.702281	4.47904
<i>(Interaction: log(population density (per km²)) *)</i> <i>(Base: Household vehicles = 0)</i>		
log(population density):non-zero vehicles	-0.518534	-11.14575
<i>(Interaction: log(crime density (per km²)) *)</i> <i>(Base: Ethnicity = white)</i>		
log(crime density):non-white ethnicity	0.414148	9.24036
log(vmt) * log(population density)	-0.225452	-7.87006
log(vmt) * log(crime density)	0.165327	5.35639
Residual price	-4.172031	-7.84026

Notes: "—" signifies statistically not significant; log of variables denote the natural logarithm to the base 'e' = 2.7182.

Then, we present results for the MNL model addressing endogeneity in residential location choice due to *natural log of vehicle miles travelled* using the instruments *natural log of public transport accessibility level (PTAL) at the outer postcode level* and *natural log of household vehicle ownership*. This model is validated using the (modified) refutability tests. This is done in two-steps, first a regression model is estimated with outcome variable *natural log of vehicle miles travelled* and regressors as instruments and all other exogenous variables available in the data. The resulting residual of this regression model is included as an exogenous variable in the choice model as a *control function*, along with the price variable and all other exogenous variables from the data.

TABLE C.3 First stage model to estimate residuals for travel behaviour endogeneity

Variable definitions	Estimate	t-stat
(Intercept)	2.37088	3.102
log(postcode population density)	-0.13515	-1.969
log(postcode crime density)	—	—
log(number of GPs)	—	—
Household income >£50,000 and <£100,000	0.19284	4.035
Household income >£100,000	0.11123	1.893
Ethnicity = non-white	—	—
log(PTAL at outer postcode level)	-0.23112	-4.44
log(HH Vehicle ownership + 1)	1.84026	39.83
<i>(Interaction: log(crime density (per km²)) *)</i>		
<i>(Base: Ethnicity = white)</i>		
Ethnicity = non-white	—	—

Notes: "—" signifies statistically not significant; log of variables denote the natural logarithm to the base 'e' = 2.7182.

TABLE C.4 Control function (for travel behaviour endogeneity) choice model result

Variable definitions	Multinomial Logit Estimate	t-stat
<i>(Alternative-specific attributes)</i>		
log(price of a 2-bedroom apartment)	-0.06837	-1.321
log(number of GPs)	0.15485	5.507
log(postcode population density)	0.53169	7.559
log(postcode crime density)	0.09309	1.27
log(vehicle miles travelled)	15.04811	6.168
<i>(Interaction: log(price of a 2-bedroom apartment) *)</i> <i>(Base: Household income < £50,000)</i>		
Household income > £50,000 and < £100,000	-0.12329	-2.055
Household income > £100,000	-0.12789	-2.537
<i>(Base: Household vehicles = 0)</i>		
Households vehicles 1	-9.50945	-2.745
Households vehicles 2	-19.43405	-3.567
Households vehicles 3 or more	-28.61799	-3.96
<i>(Interaction: log(population density (per km²)) *)</i> <i>(Base: Household vehicles = 0)</i>		
log(population density):non-zero vehicles	-0.55264	-11.787
<i>(Interaction: log(crime density (per km²)) *)</i> <i>(Base: Ethnicity = white)</i>		
log(crime density):white ethnicity	0.35911	7.968
log(vmt) * log(population density)	-0.2084	-7.284
log(vmt) * log(crime density)	0.14382	4.666
Residual VMT	-12.15325	-5.096

Notes: "—" signifies statistically not significant; log of variables denote the natural logarithm to the base 'e' = 2.7182.