METHODOLOGICAL AND EMPIRICAL CHALLENGES IN MODELLING RESIDENTIAL LOCATION CHOICES

Alireza Zolfaghari
June 2013

A thesis submitted as fulfilment of the requirements for the degree of Doctor of Philosophy of Imperial College London

Centre for Transport Studies
Imperial College London
DECLARATION

I hereby certify that all material in this thesis is my own work. Any quotation from, or description of the work of others is acknowledged by reference to the sources, whether published or unpublished.

Alireza Zolfaghari

The copyright of this thesis rests with the author and is made available under a Creative Commons Attribution Non-Commercial No Derivatives licence. Researchers are free to copy, distribute or transmit the thesis on the condition that they attribute it, that they do not use it for commercial purposes and that they do not alter, transform or build upon it. For any reuse or redistribution, researchers must make clear to others the licence terms of this work.
ABSTRACT

The modelling of residential locations is a key element in land use and transport planning. There are significant empirical and methodological challenges inherent in such modelling, however, despite recent advances both in the availability of spatial datasets and in computational and choice modelling techniques.

One of the most important of these challenges concerns spatial aggregation. The housing market is characterised by the fact that it offers spatially and functionally heterogeneous products; as a result, if residential alternatives are represented as aggregated spatial units (as in conventional residential location models), the variability of dwelling attributes is lost, which may limit the predictive ability and policy sensitivity of the model. This thesis presents a modelling framework for residential location choice that addresses three key challenges: (i) the development of models at the dwelling-unit level, (ii) the treatment of spatial structure effects in such dwelling-unit level models, and (iii) problems associated with estimation in such modelling frameworks in the absence of disaggregated dwelling unit supply data. The proposed framework is applied to the residential location choice context in London.

Another important challenge in the modelling of residential locations is the choice set formation problem. Most models of residential location choices have been developed based on the assumption that households consider all available alternatives when they are making location choices. Due to the high search costs associated with the housing market, however, and the limited capacity of households to process information, the validity of this assumption has been an ongoing debate among researchers. There have been some attempts in the literature to incorporate the cognitive capacities of households within discrete choice models of residential location: for instance, by modelling households’ choice sets exogenously based on simplifying assumptions regarding their spatial search behaviour (e.g., an anchor-based search strategy) and their characteristics. By undertaking an empirical comparison of alternative models within the context of residential location choice in the Greater London area this thesis investigates the feasibility and practicality of applying deterministic choice set formation approaches to capture the underlying search process of households. The thesis also investigates the uncertainty of choice sets in residential location choice modelling and proposes a simplified probabilistic choice set formation approach to model choice sets and choices simultaneously.

The dwelling-level modelling framework proposed in this research is practice-ready and can be used to estimate residential location choice models at the level of dwelling units without requiring independent and disaggregated dwelling supply data. The empirical comparison of alternative exogenous choice set formation approaches provides a guideline for modellers and land use planners to avoid inappropriate choice set formation approaches in practice. Finally, the proposed simplified choice set formation model can be applied to model the behaviour of households in online real estate environments.
ACKNOWLEDGMENTS

First and foremost, I would like to express my gratitude to my supervisors, Dr Aruna Sivakumar and Professor John Polak, for their invaluable guidance, assistance and financial support throughout my PhD research at Imperial College London. I have been privileged to have had the opportunity to work with them for the past four years.

I would also like to thank my examiners Professor Eric Miller and Professor Nilay Shah for the time and care they have taken in providing me with valuable feedback. My very grateful thanks are also extended to my colleagues and friends at the Centre for Transport Studies for being the best office mates and for the ideas we shared during the course of my PhD. I must also thank Ms Jackie Sime and Ms Fionnuala Donovan for their administrative assistance.

I am immensely indebted to my parents Ali-mohammad and Shahla for their never-ending love, patience and support in my life; without them this journey would have been impossible. Last but not least, I would like to thank my partner in life Pegah for continuously supporting me and encouraging me when I was at my lowest.
# TABLE OF CONTENTS

DECLARATION .................................................................................................................................................. i

ABSTRACT ........................................................................................................................................................ ii

ACKNOWLEDGMENTS ....................................................................................................................................... iii

TABLE OF CONTENTS ....................................................................................................................................... iv

LIST OF TABLES .................................................................................................................................................. ix

LIST OF FIGURES ............................................................................................................................................... x

CHAPTER 1 INTRODUCTION ............................................................................................................................. 1

1.1 Background and Motivation ....................................................................................................................... 1

1.2 Research Objectives ................................................................................................................................... 6

1.3 Structure of the Thesis ............................................................................................................................... 6

CHAPTER 2 LITERATURE REVIEW .................................................................................................................... 8

2.1 Overview ..................................................................................................................................................... 8

2.2 Urban Land Use Models: Timeline of Land Use Modelling Evolution .................................................... 9

2.2.1 Simulation Models .................................................................................................................................. 9

2.2.2 Optimisation Models .......................................................................................................................... 18

2.3 Residential Location Modelling: Theoretical Perspective ........................................................................ 19

2.3.1 Gravity-type Models ........................................................................................................................... 20

2.3.2 Bid-rent Models .................................................................................................................................. 21

2.3.3 Location Choice Models ..................................................................................................................... 22

2.4 Discrete Choice Analysis of Residential Location Choices ...................................................................... 25

2.4.1 Basic Concepts ..................................................................................................................................... 25

2.4.2 Important Model Structures .............................................................................................................. 27

2.4.3 Stated Preferences vs. Revealed Preferences .................................................................................... 30

2.4.4 Model Estimation ................................................................................................................................ 31

2.4.5 Research Challenges in Modelling Residential Location Choices .................................................. 33
4.2.1 The MNL Formulation .................................................................................. 74
4.2.2 The GEV Formulation .................................................................................. 75
4.2.3 The Mixed Logit Formulation ..................................................................... 82
4.2.4 Estimation Challenges ................................................................................ 86
4.3 Constructing the Universal Choice Set in Dwelling-level Models: A Dwelling Synthesising Approach ............................................................................... 87
4.4 Validation of the Choice Set Synthesising Approach .................................... 90
  4.4.1 Monte Carlo Experiment .......................................................................... 90
  4.4.2 Monte Carlo Results ................................................................................ 94
4.5 Empirical Application .................................................................................... 96
  4.5.1 Data ......................................................................................................... 97
  4.5.2 Results .................................................................................................... 100
4.6 Conclusion ..................................................................................................... 103

CHAPTER 5 SPATIAL SEARCH AND CHOICE SET FORMATION IN RESIDENTIAL LOCATION CHOICE MODELLING ................................................................. 104

  5.1 Overview ...................................................................................................... 104
  5.2 Modelling Framework ................................................................................ 105
    5.2.1 Spatial Search Strategies and Choice Set Formation ............................ 105
    5.2.2 Screening Models ................................................................................ 108
    5.2.3 Hazard-based Screening Model ............................................................ 109
    5.2.4 Different Choice Set Formation Approaches ....................................... 111
  5.3 Empirical Analysis ....................................................................................... 113
    5.3.1 Data Preparation for the Residential Choice Model ............................. 114
    5.3.2 Data Preparation for the Hazard-based Screening Model .................... 117
  5.4 Model Estimation Results ........................................................................... 118
    5.4.1 Hazard-based Screening Model ............................................................ 118
    5.4.2 Residential Location Choice Model ..................................................... 120
C.2. Variable Descriptions ............................................................. 178

APPENDIX D LIST OF ACRONYMS AND ABBREVIATIONS ............... 185

APPENDIX E PUBLICATIONS ................................................................ 186
  E.1. Journal Articles........................................................................ 186
  E.2. Conference Papers .................................................................. 186
LIST OF TABLES

Table 2-1- Comparison of Different Urban Simulation Models......................................................18
Table 2-2 - Previous Research Findings in Zonal-based Residential Location Choice Models.........................................................................................................................37
Table 2-3 – Different Levels of Aggregation of Residential Alternatives in Developed Models.................................................................................................................................42
Table 2-4 – Observed and Unobserved Correlation and Heterogeneity in Residential Location Choice Models..........................................................................................................................46
Table 4-1- Dissimilarity and Allocation Parameters of the Example Model ......................79
Table 4-2- Allocation Matrix of the Example Model.................................................................79
Table 4-3- An Example of Dwelling-level and Zonal-level Disturbances ......................83
Table 4-4- Notation Reference for Variables in the Utility Function and True Parameters .................................................................................................................................90
Table 4-5- Variables and Parameters of the Hedonic Price Model ........................................92
Table 4-6- Validation of the Choice Set Synthesising Approach for Disequilibrium and Equilibrium Market Clearing.................................................................................................................95
Table 4-7- Counts of Remaining Cases in Data Clearing.........................................................97
Table 4-8- Dwelling-level Residential Location Choice Model Estimation Results ..........101
Table 5-1- Kolmogorov-Smirnov Statistics of Different Distributions............................117
Table 5-2- Hazard-based Screening Model Estimation Results............................................119
Table 5-3- Summary Statistics of Expected Choice Sets.......................................................121
Table 5-4 – Residential Location Choice Models Estimation Results..................................122
Table 5-5- Prediction Test on the Hold-out Sample...............................................................127
Table 6-1 – Case Study 1 - Estimated Models ......................................................................142
Table 6-2 – Estimated Parameters and Prediction Tests of Case Study 1 .........................143
Table 6-3 – Estimated Parameters and Prediction Tests of Case Study 2 .......................145
LIST OF FIGURES

Figure 1-1- Urban and Rural Population in the UK, 1950 to 2050--Percentage of Total Population (United Nations, 2012) ................................................................. 1
Figure 1-2- Interaction of Land Use and Transport System (Wegener, 2004) ............... 3
Figure 2-1- Different Approaches in Land Use Modelling ............................................. 11
Figure 2-2- Timeline Showing the Evolution of Urban Simulation Models ................. 12
Figure 2-3- Parcels (Green), 150m Gridcells (Red) and Travel Analysis Zones (Blue) for the City of London ................................................................. 17
Figure 2-4- Theoretical Frameworks in Residential Location Modelling ..................... 19
Figure 2-5- Relationship of Location Choice and Bid-rent Models ............................... 24
Figure 2-6- Explanatory Variables and Data Sources Used in Modelling of Residential Location Choices ................................................................. 35
Figure 3-1- A Comprehensive Conceptual Framework for the Housing Market ............ 62
Figure 3-2 – Different Choice Structures for Mobility/Refurbishment/Tenure Decisions ............................................................................................................. 64
Figure 3-3- Dwelling Supply in Rental and Owner’s Market ........................................ 67
Figure 3-4 – Research Programme and Thesis Chapters ............................................. 68
Figure 4-1-An Example of Random Assignment of Parcels to Dwellings within a Zone 72
Figure 4-2- An Example of Residential Choice among 8 Dwellings Scattered in 4 Zones 77
Figure 4-3- Generalised Nested Structure for Capturing Spatial Correlation in Dwelling-Level Models ..................................................................................... 77
Figure 4-4- An Example of Cross-correlated Random Fields Used in the Simulation of the Synthetic City ......................................................................................... 91
Figure 4-5- Distribution of Dwelling Prices in the Synthetic City (Disequilibrium Approach) ........................................................................................................ 93
Figure 4-6- Distribution of Dwelling Prices in the Synthetic City (Equilibrium Approach) ........................................................................................................ 93
Figure 4-7- Validation of the Choice Set Synthesising Approach for Disequilibrium and Equilibrium Market Clearing ................................................................. 96
Figure 4-8- LHS 2002 Observations in the Greater London Zoning by Travel Analysis Zones (TAZs) ................................................................................................. 98
Figure 5-1 – Exogenous Choice Set Formation Approaches in Spatial Context ........... 107
Figure 5-2- LATS 2001 Households’ Spatial Distribution Zoning by Travel Analysis Zones (TAZs)...................................................................................................................................................... 115
Figure 5-3- Distribution of Work Commute Time for Londoners (Data Source: LATS 20001).................................................................................................................................................................................. 118
Figure 5-4- Distribution of Housing Prices in London for Year 2001 (Data Source: Land Registry)........................................................................................................................................................................................................... 118
Figure 5-5- Sensitivity of Housing Price Parameter and Rho Square to the Choice of Percentile .............................................................................................................................................................................................................. 124
Figure 5-6 – Sensitivity of Commute Time Parameter and Rho Square to the Choice of Percentile .............................................................................................................................................................................................................. 125
Figure 6-1- The KBS model vs. the Proposed Model ................................................................................................................................................................................................................................. 138
CHAPTER 1

INTRODUCTION

1.1 Background and Motivation

Worldwide, cities and urban areas are gaining an estimated 60 million people per year - over 1 million every week. In many developing countries, cities are growing two or three times faster than the overall population. In Europe, the percentage of the population living in urban areas is expected to rise from 73% in 2000 to approximately 80% by 2030. More than 80% of the UK population now lives in urban areas, with an annual urbanisation rate of 0.7% (United Nations, 2012).

Figure 1-1: Urban and Rural Population in the UK, 1950 to 2050—Percentage of Total Population (United Nations, 2012)

Population growth and urban expansion in metropolitan regions will over the next few decades result in an increase in demand for major urban infrastructures including housing, commercial and office spaces, transportation for people and goods, utilities, etc. The accelerating rate of urbanisation throughout the world is also an important
factor in climate change since the CO₂ emissions and other environmental impacts associated with cities make the most direct contribution to climate change. A major challenge for governments and regional authorities, therefore, is to support the economic and population growth in urban areas in a sustainable way by: (i) making the best use of current urban infrastructures, (ii) reducing environmental impacts and CO₂ emissions, (iii) optimising traffic fluidity, (iv) improving energy efficiency, (v) increasing quality of life and social equity.

In order to meet the goals and aspirations of the community, governments and regional authorities guide the development of cities by setting land use and transportation policies and investing in public facilities such as development of new transportation infrastructures. Land use policies in general involve attempts to improve the attractiveness of urban or suburban areas as locations for investment. However, it is necessary to target these policies towards the types of investment sectors that are regarded to be most sensitive to small spatial scale variations, such as the real estate and property development sectors. Since the transactions in the real estate market take place within complex legal and institutional frameworks, policy makers can guide the spatial pattern of private sector investments (e.g., favour property development in less developed locations) by changing or relaxing these legal and institutional frameworks in order to achieve the community goals (McCann, 2001).

There is a wide recognition that transport and land use are interrelated via accessibility. Transport strategies can significantly improve the accessibility of an area, so that area becomes attractive for households and businesses which generate new developments in the area. On the other hand, the relocation of households and businesses to an area, as well as the property developments in that area, affects the pattern of trips and the performance of the transport system (see Figure 1-2). There is, therefore, a two-way interaction between transport and land use (Ortuzar and Willumsen, 2011).

Following the development of the monocentric model of urban land uses in urban economics (Alonso, 1960), urban economists were actively engaged in the development of theoretical models of urban spatial structure. Although these theoretical models provided powerful general qualitative insights into many crucial policy questions, they were not able to provide the quantitative answers needed for specific urban policies
and plans. Subsequently, urban land use modelling emerged as an applied field within urban economics with the aim of developing tools that can be used to make sound policy decisions about the allocation of resources in urban space (Anas, 1982).

![Figure 1-2- Interaction of Land Use and Transport System (Wegener, 2004)](image)

Urban land use models are mainly developed as analytical tools for planners and urban policy makers in order to analyse and forecast the effects of land use policies such as land use zoning policies¹, regeneration policies², etc. Recent urban land use models are capable of predicting changes in the distribution of population/employment and land use patterns and they can be combined with transportation models in order to capture the interaction of land use and transportation (i.e., integrated urban models).

Residential location models in general are one of the major components of land use models since a large proportion of urban land use is residential and most of the trips conducted in an urban area are home-based trips (i.e., a trip that starts or ends at home). In early generation land use models gravity models were traditionally used as a location model to predict the spatial distribution of population in cities in an aggregate manner (see Wilson, 1970). Following the seminal works of Lerman (1976) and

---

¹ Land use zoning polices regulate the types of investment and development activities at particular locations (zones).
² Regeneration policies are designed to encourage redevelopment of the city centres of urban areas, and to counter the out-migration and depopulation of city centres.
McFadden (1978), however, the disaggregate approach of modelling residential locations based on microeconomic choice models (i.e., discrete choice models) has been the dominant approach in the literature.

Models of residential location choice are developed as a part of an urban or regional case study in order to help us understand how residential location choices are made (see Kim et al., 2005), or as an important component of integrated land use-transport model systems, since they predict the medium-term dynamics in the urban area and help to determine how the urban landscape is shaped over time (see Waddell et al., 2003). Residential location choice models quantify how and to what extent factors such as accessibility, distance from work, price, school quality and safety concerns have an impact on where people choose to live and help us to study phenomena such as racial segregation, gentrification, etc. in urban areas (see Bayer et al., 2005; Guo and Bhat, 2007a).

Despite the recent advances in the availability of spatial datasets and improvements in computing power and choice modelling techniques, significant empirical and methodological challenges remain in modelling residential location choices. These challenges arise because of the fact that the residential location choice process is one of the sub-processes of the housing market and there are some unique characteristics of the housing market (e.g., heterogeneous and immobile products, high search costs, etc.) that need to be considered in the development of an empirical model of the residential location choice process.

One of the most important of these challenges concerns spatial aggregation. While implementation of spatially aggregated residential location choice models requires less computational resources and disaggregated datasets, this aggregation implies an incomplete representation of the heterogeneity of dwelling unit attributes. The housing market is characterised by the fact that it offers spatially and functionally heterogeneous products; therefore, by representing residential alternatives as aggregated spatial units, as in conventional models, the variability of dwelling attributes is lost, and this may limit the predictive ability of the model. Recently, some authors have attempted to develop residential location choice models at the level of dwelling

---

3 Gentrification is a process in which poorer residents of an area are gradually displaced by an influx of wealthier residents.
units, assuming that all dwellings are differentiated based on their observed and unobserved attributes (see Habib and Miller, 2009). Estimation of such models requires spatially disaggregated housing supply data, which is not available for many metropolitan regions.

The locational fixity characteristic of the housing market leads to an important feature of residential location choice models in that alternatives are either spatial units or dwelling units augmented to spatial units. This association of alternatives to space leads to another important issue in modelling: the IIA (Independent from Irrelevant Alternatives) assumption (in a conventional Multinomial Logit (MNL) structure) is no longer invalid. There is a dependency between different residential alternatives since it is expected that closer alternatives will be more similar than farther alternatives. Previous research in this area has led to the development of more relaxed model structures in order to incorporate this spatial dependency within spatially aggregated residential location choice models (see Guo and Bhat, 2004a). Incorporating spatial effects for models developed at the level of dwelling units is not explored in the literature, however.

Since there are a large number of alternatives involved in the residential location choice process and since the ability of households in gathering and processing information in the housing market is limited (i.e., due to the high search costs), discrete choice modelling of households’ residential location choices are accompanied by a choice set formation problem. In other words, conceptually, residential location choice is a dynamic spatial search process, in which households are exposed to a dynamically changing set of residential alternatives from which they assemble and evaluate a choice set of credible alternatives, and ultimately make a final decision (Habib and Miller, 2007). Search behaviour is characterised by the limitations inherent in human perceptual abilities and by subjectively defined knowledge about urban spatial environment: limitations that severely constrain the choice. Since this underlying spatial search and choice process is typically unobserved, conventional models typically assume that households have perfect information about all available alternatives. Some authors, however, have acknowledged the existence of an underlying search process and have thus attempted to formulate discrete choice models incorporating the choice set formation process in the context of residential location choice (see Martínez et al.,
To date, the performance of different choice set formation approaches has not been evaluated in the residential location choice literature and it is not clear whether models incorporating the choice set formation process perform better in forecasting than the conventional models based on the perfect information assumption.

1.2 Research Objectives

The unique characteristics of the housing market, as explored above, result in several methodological and empirical challenges in the modelling of residential location choices.

The heterogeneity and immobility of housing market products give rise to fundamental questions of how to represent residential alternatives in modelling and how the spatial nature of alternatives affects the model structure. The large number of alternatives and the high search costs of the housing market provoke the main question of how the choice sets should be constructed and which alternatives should be included in households’ choice sets. The main objective of this thesis is to answer these fundamental questions and thus to improve the representation of the housing market within empirical modelling of residential location choices.

This research will, therefore, address two critical challenges in the modelling of residential location choices; one concerns the representation of alternatives and spatial structure effects, and the other concerns the choice set formation problem. The details of the research programme are presented in chapter 3.

1.3 Structure of the Thesis

This thesis has seven chapters. Chapter 2 presents a comprehensive literature review of urban land use models in general and of discrete choice modelling of residential location choices in particular, covering critical gaps in residential location choice modelling from both theoretical and empirical perspectives. Chapter 3 investigates residential location choices from a wider perspective, and proposes a conceptual framework for the housing market where residential location choices are identified as one of the sub-processes of the housing market. This chapter also describes the research programme of this thesis building on the conceptual model and the critical
gaps identified in chapter 2. Chapter 4 proposes a residential location choice model at
the dwelling unit level incorporating spatial effects at the zonal-level. This chapter also
proposes an innovative approach to estimate such models in the absence of
disaggregated dwelling supply data. Chapter 5 investigates the feasibility and
practicality of applying deterministic choice set formation approaches in a residential
location choice context for capturing the underlying search process of households
within the housing market. Chapter 6 proposes a simplified probabilistic choice set
formation model in a residential location choice context. Finally, chapter 7 concludes
with a summary of contributions and suggestions for future research in this area.
CHAPTER 2

LITERATURE REVIEW

2.1 Overview

Questions such as how and to what extent urban development induces travel demand and how and to what extent the performance of the transport system induces urban development have motivated researchers and planners to develop integrated urban models in order to capture this long-term behaviour of urban systems. In this chapter, we begin by reviewing the evolution of urban land use models from the earliest and simplest developments to the most advanced dynamic microsimulation models. Having discussed the general overview of land use model systems, we review different modelling frameworks of residential location components that have been applied in different urban model systems. In order to establish the theoretical framework of this study we identify and review different approaches to the modelling of residential locations from aggregated gravity-type models to behavioural location choice models.

Since residential location choice models based on random utility theory are one of the major components of state-of-the-art microsimulation urban models, the main body of this chapter is devoted to a review of the methodological and empirical challenges in modelling residential location choices based on the discrete choice framework. The main goal of this chapter, therefore, is to understand the current challenges in existing models of residential locations and to identify potential opportunities to develop improved models taking advantage of advances in computing power and econometric methods.

The chapter is structured as follows: Section 2.2 briefly reviews the history and the evolution of operational urban land use models. Section 2.3 reviews the different residential location modelling frameworks and their underpinning theories proposed in the literature. Section 2.4 elaborates on modelling residential locations based on discrete choice theory and identifies fundamental research challenges. Section 2.5
narrowed down the scope of the review to the challenges involved in representing alternatives and incorporating the spatial structure effects into residential location choice models. The problem of massive universal choice set and the choice set formation problem in the residential location choice context are discussed in section 2.6. Finally, section 2.7 concludes this chapter by summarising the critical gaps in discrete choice residential location modelling literature.

2.2 Urban Land Use Models: Timeline of Land Use Modelling Evolution

This section presents the history and the evolution of operational urban land use models. The purpose of this section is to review briefly the most important developments in order to gain a general overview of urban land use modelling and to identify potential areas for research.

Urban land use models in general can be classified into two major categories as: (i) Simulation Models, (ii) Optimisation Models. Simulation models describe how urban processes actually work (i.e., descriptive models) and can be used for forecasting, scenario building and policy analysis. Optimisation models, on the other hand, are prescriptive and suggest how urban processes should work (i.e., normative models). The main focus of this section is on reviewing the evolution of simulation models, however, optimisation models will also be briefly discussed.

2.2.1 Simulation Models

The emergence of simulation models dates back to the work of Lowry (1964) who developed an operational model of urban development based on the spatial interactions of employment and population. During the past decades, the principles set out by Lowry have informed the development of many operational land use models, including DRAM/EMPAL, MEPLAN, TRANUS, and PECAS (Echenique et al., 1969; Putman, 1983; de la Barra, 1989; Hunt and Abraham, 2003). We refer to these models as Lowry-type models because of their similarities with the original Lowry model.

Despite its simple formulation, the Lowry model depicted the relationships between transportation and land use well. It has been seriously criticised, however, for its lack of

---

4 This section is not meant to be an exhaustive review of all previous developments as there are many good reviews available that comprehensively review operational urban land use models (see Wegener, 1994; Oryani and Harris, 1996; Miller et al., 2004; Wegener, 2004; Hunt et al., 2005).
behavioural consistency (i.e., model components are not derived from microeconomic principles) and lack of systematic consistency (i.e., model components are not linked together based on microeconomic principles) (see Anas, 1982). The behavioural inconsistency of the location model components of the original Lowry model have been eliminated in subsequent Lowry-type models by making use of microeconomic models, such as discrete choice models, rather than the gravity formulation used originally. The systematic inconsistency has, however, not been completely eliminated in the extensions to the Lowry model.\(^5\)

In order to eliminate the systematic inconsistency of Lowry-type models, researchers developed urban land use models from an economic viewpoint by linking the behavioural location models on the demand-side and the supply-side based on the notion of equilibrium. These models focus on the equilibrium solution to a system of models and are referred to as equilibrium land use models (see Anas, 1994; Martinez, 1996; Martínez and Henríquez, 2007). These models are indebted to the work of Alonso (1960) who introduced the bid-rent theory for urban land uses and Scarf (1977) who introduced an algorithm to compute the equilibrium prices. Equilibrium models are aggregate in nature as agents are categorised into a small number of categories based on the representative agent paradigm.

More recent studies are based on the microsimulation approach (i.e., modelling at the level of disaggregate agents such as households and firms) to incorporate more behavioural realism in modelling urban land uses (see Waddell et al., 2003; Salvini and Miller, 2005; Hurtubia et al., 2011; Simmonds et al., 2011). While equilibrium models describe a hypothetical long-term market condition, microsimulation models drop the equilibrium assumption and acknowledge different speeds of processes in both the demand-side and the supply-side of the market. For example, the speed of housing supply development might be less than the growth in housing demand in the short-term (i.e., supply lag), causing disequilibrium conditions. This is why microsimulation models

\(^5\) It should be noted that some of the extensions of the Lowry model made use of spatial input-output models in order to overcome the systematic inconsistency of the Lowry model from a macroeconomic perspective (e.g., MEPLAN, TRANUS and PECAS). These models can be also viewed as equilibrium models since an input-output framework is in fact a simplified formulation of the Walrasian equilibrium. These models are classified as Lowry-type models here, however, because of their similarities with the original Lowry model. Zhao and Kockelman (2004) examined the existence and uniqueness of equilibrium solutions for spatial input-output models and provided a general solution algorithm for such models.
are also referred to as dynamic microsimulation models in the literature. Dropping the equilibrium assumption also has computing advantages for microsimulating urban processes since in dynamic microsimulation models there is no need to solve the fixed-point equations that represent the equilibrium.

The rapid growth of computing power, the emergence of new technologies for collecting, processing and visualising geographical data, such as Geographical Information Systems (GIS), and advances in modelling techniques have each paved the way for researchers to move from gravity-type models to behavioural econometric models, such as location choice models, and from aggregate equilibrium models to disaggregate microsimulation models.

Figure 2-1 depicts the different approaches to land use modelling, while Figure 2-2 summarises the chronological evolution of urban simulation models following Waddell, (2005). Lowry-type models, equilibrium models and microsimulation models are discussed comprehensively in the following sub-sections.

![Figure 2-1 - Different Approaches in Land Use Modelling](image-url)
2.2.1.1 Lowry-type Models

Lowry (1964) developed a simple model system capable of predicting population and employment distributions for a given future year based on the gravity model formulation. Lowry identified three sectors that occupy an urban area: (a) Basic sector, (b) Retail Sector, and (c) Residential sector.

The basic sector includes the employment involved in the production of goods and services that meet non-local demand and are exported outside the urban area (e.g., industrial activities). The location and size of basic sector activities do not, therefore, depend on local market size. The basic sector is an exogenous element of the Lowry model and the spatial distribution of employment in the basic sector is assumed to be known to the analyst. The retail sector includes the employment involved in production of goods and services that meet local demand. The location and size of these activities depend on local market size. The retail sector is an endogenous element of Lowry
model. The residential sector comprises the resident population in the urban area which is also an endogenous element of the model.

The core assumption of the Lowry model is that regional and urban growth (or decline) is a function of the expansion (or contraction) of the basic sector. Consecutively, expansion or contraction of the basic sector affects the retail sector and residential sector. The model, therefore, begins by establishing the impacts of the basic sector over the residential sector and retail sector and then determines the distribution of the resident population and retail employment based on the gravity formulation.

The Lowry model has been extended and applied by several authors in different metropolitan regions. The most used successors to the Lowry model is the Integrated Transportation and Land Use Package (ITLUP) developed by Putman (1983). ITLUP consists of two main sub-models: (i) Disaggregate Residential Allocation Model (DRAM), which forecasts household locations by household types (employed residents) in relation to employment locations in a future year, and (ii) Employment Allocation Model (EMPAL) which forecasts employment locations by type in relation to an attractiveness measure.

One of the critical deficiencies of the Lowry model is its inability to handle linkages between different economic sectors. The development of the Input-Output model by Leontief (1941) paved the way for the future extension of the Lowry model by providing a computational framework for representing the interdependencies of different economic sectors. MEPLAN (Echenique et al., 1969) and its close substitute TRANUS (de la Barra, 1989) each adopt the multiregional input-output modelling approach in order to extend Lowry's modelling approach into a more comprehensive multisectoral approach. By defining households as sectors producing labour and consuming commodities, the origins and destinations of commodity flows can be derived from an intersectoral input-output table.

Another important model developed in this area is PECAS (the Production, Exchange and Consumption Allocation System), developed by Hunt and Abraham (2003). PECAS contains two principal sub-models: (i) Activity Allocation (AA) which is an aggregate,
spatial input-output model, and (ii) Spatial Development (SD) which is a disaggregate land use transition model.\(^6\)

**2.2.1.2 Equilibrium Models**

Motivated by Alonso's bid-rent theory in urban economic literature, researchers attempted to apply microeconomic principles to link, based on the notion of equilibrium, the demand-side and supply-side sub-models of urban land use models. Equilibrium models are based on the assumption that prices instantaneously adjust so that excess demands are zero (i.e., the Walrasian equilibrium). The notion of equilibrium provides a simple and powerful market clearing approach for urban land uses which is useful for long-term analysis. Equilibrium models can also incorporate the potential of feedback influence from the market equilibrium on individual choices (e.g., the utility of a location depends in part on the number or attributes of other individuals that choose the same or nearby locations).

Equilibrium models usually involve a non-linear equation system and require an iteration algorithm to solve the equilibrium (see Scarf, 1977). Although the existence and uniqueness of equilibrium can be proved in most models, based on the Brouwer fixed-point theorem, the solution algorithms are usually computationally intensive. This computational cost means that the application of equilibrium models tends to be limited to aggregated analysis (i.e., households and firms are clustered into categories, and land is divided into zones and dwellings into types).

Equilibrium urban simulation models evolved from three pioneering models that developed in the 1970s including: (i) the National Bureau of Economic Research (NBER) model which was developed by Ingram et al. (1972), (ii) the Harvard Urban Development Simulation (HUDS) model which was the modified and extended version of the NBER model, developed by Kain et al. (1976), (iii) the Urban Institute Model (UIM) which was developed by deLeeuw and Struyk (1975).

Alex Anas and his colleagues developed many equilibrium based urban simulation models for different metropolitan regions in the United States, including: CATLAS

---

\(^6\) Land use transition models in general refer to a variety of analytical tools that help to understand the spatial and temporal patterns of land conversion [see Irwin and Geoghegan (2001) and Plantinga and Irwin (2006) for a review of alternative approaches].

METROSIM evolved from previous developments by Anas and contains three components of market equilibrium: (i) labour market equilibrium and job assignment, (ii) housing market equilibrium, and (iii) commercial space equilibrium. The model iterates between these markets and the transportation system for equilibrium of land use and transportation flows. Another important equilibrium model of transport, land use and regional economy is RELU-TRAN, which was recently developed for the Chicago metropolitan area (see Anas and Liu, 2007). All of these models used a discrete choice framework to model residential locations (see Section 2.3.3).

MUSSA (Martinez, 1996) is another equilibrium model which is designed to forecast the expected location of residents and firms in an urban area. However, MUSSA presents an alternative framework for modelling residential locations by adopting Ellickson’s (1981) stochastic bid-rent approach (see Section 2.3.2). Here, the equilibrium is defined by the market clearing condition that all consumers are located somewhere, which is attained by adjusting the consumers’ utility levels (as opposed to adjusting the price levels in a discrete choice framework). MUSSA II (Martinez et al., 2007) extended the previous version by updating the supply model and the way that constraints are dealt with. The supply is produced by real estate developers that behave as profit maximisers. The supply model in MUSSA II has changed from an aggregate deterministic econometric model to a stochastic behavioural model based on Logit formulation.

Equilibrium models are attractive because the equilibrium assumption provides an easily operationalised theoretical framework through which to adjust the prices and match the supply and demand. In spite of the advantages of the equilibrium approach for clearing the market, it has been criticised for oversimplification of the market characteristics and behaviour of the agents in the market (see Farooq and Miller, 2012).

2.2.1.3 Microsimulation Models

The equilibrium approach for clearing the market relies on several unrealistic assumptions about the behaviour of the agents and the market including: (i) agents have perfect information about the market and there is no friction in the market (i.e., no
search or transaction costs), (iii) agents with different preferences can cooperate to achieve a cooperative outcome (i.e., a market clearing condition), and (iii) the outcome is not path-dependent.

None of these assumptions are perfectly true in reality, however, since: (i) agents have limited information about the market and their decisions are constrained by search and transaction costs, (ii) agents in a real housing market make decisions independently and are thus non-cooperative among each other, (iii) the market clearing condition in reality is path-dependent as the choices of agents at a given point in time are limited by the decisions made in the past.

The microsimulation approach was originally proposed by Orcutt (1957) to avoid aggregation bias7 with the aim of analysing the impact of policies in demographic models at the individual level. Recent urban land use models have eschewed the equilibrium assumption in favour of the microsimulation approach as a means of minimising the unrealistic characteristics of equilibrium models (see Miller et al., 2004). This is achieved, however, at the cost of losing the equilibrium model’s ability to take account of feedback influence.

There are several operational land use models which use the microsimulation approach, such as IRPUD (Wegener, 2011b), UrbanSim (Waddell et al., 2003) and ILUTE (Salvini and Miller, 2005). The IRPUD model (Wegener, 2011b) is dynamic microsimulation model of location and mobility decisions in a metropolitan area which was initially designed and implemented in 1977 at the Institute of Spatial Planning of the University of Dortmund (IRPUD). The spatial and temporal dimensions in IRPUD are incorporated by the subdivision of the study area into zones, and the subdivision of time into periods of one or more years' duration. IRPUD is also implemented within the ILUMASS (Integrated Land Use Modelling and Transportation System Simulation) project for modelling land uses (Wagner and Wegener, 2007).

UrbanSim (Waddell et al., 2003) is another dynamic microsimulation land use model used to simulate the spatial and temporal evolution of land use and the locations of households and jobs within metropolitan areas. In 2005, a new Open Platform for Urban

---

7 I.e. error entered into a model as result of using average characteristics of the population and ignoring the potential heterogeneity that might exist among members of the population.
Simulation (OPUS) was implemented to support further development of UrbanSim. According to recent surveys, UrbanSim has become the most widely used land use model system by planning agencies in the United States. Geographically, UrbanSim can operate at the level of zones, gridcells, and parcels (see Figure 2-3).

![Figure 2-3: Parcels (Green), 150m Gridcells (Red) and Travel Analysis Zones (Blue) for the City of London](image)

ILUTE (Integrated Land Use, Transportation, Environment) is another important microsimulation dynamic urban model system (Salvini and Miller, 2005). This model system has been under development at the University of Toronto for the last several years. ILUTE operates at the most disaggregate level of dwelling units and accommodates multiple spatial resolutions.

As discussed earlier, equilibrium models adjust prices to match the supply and demand to clear the market. The market clearing process is not a straightforward task in microsimulation models, however, and is still a critical research challenge. UrbanSim assumes that agents are price takers and determines prices exogenous to households’ location choices based on a hedonic regression model (i.e., disequilibrium market
clearing). Unfortunately, it is not clear whether the coupling of hedonic regression models with the location choice models is appropriate for market clearing in microsimulation models (Waddell, 2009).

Farooq and Miller (2012) argued that price determination is the outcome of the supply and demand interaction even in the disequilibrium market clearing and proposed an endogenous approach for market clearing in dynamic microsimulation models. They proposed to simulate the transaction prices within a certain range of the asking prices (i.e., predicted based on a hedonic model) by matching the available supply with the demand based on households’ choice sets. This approach is computationally very expensive and is heavily dependent on the validity of the behavioural rules assumed to generate households’ choice sets.

The market clearing problem in dynamic microsimulation models has also been examined by Hurtubia et al. (2011). They proposed a method to model location choice and real estate prices simultaneously based on the stochastic bid-rent approach, acknowledging that both location choices and prices are a function of the household’s preferences. Table 2-1 summarises the different market clearing approaches applied in urban simulation models.

<table>
<thead>
<tr>
<th>Simulation Models</th>
<th>Market Clearing</th>
<th>Representation of Agents</th>
<th>Representation of Alternatives</th>
<th>Feedback</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lowry-type Models</td>
<td>Input-Output Framework</td>
<td>Aggregate</td>
<td>Aggregate (Zones)</td>
<td>No</td>
</tr>
<tr>
<td>Equilibrium Models</td>
<td>Walrasian Equilibrium</td>
<td>Aggregate</td>
<td>Aggregate (Zones)</td>
<td>Yes</td>
</tr>
<tr>
<td>Microsimulation Models</td>
<td>Ad hoc</td>
<td>Disaggregate</td>
<td>Aggregate (Zones) / Disaggregate (Dwellings)</td>
<td>No</td>
</tr>
</tbody>
</table>

### 2.2.2 Optimisation Models

Application of optimisation models in land use planning dates back to the work of Herbert and Stevens (1960) and Harris (1962). The main purpose of optimisation land

---

8 Some of the Lowry-type models, including the original Lowry model, used ad hoc equilibration approaches, such as uniform distribution of excess demands among the geographic zones of the urban area (see Anas, 1983b).
use models is to find the most efficient allocation of land uses in a city. Therefore, they cannot be used to assess the impacts of transport and land use policies on the evolution of existing urban areas. More recent optimisation models consider multiple objectives and constraints simultaneously and are developed at the level of parcels (see Masoomi et al., 2012).

2.3 Residential Location Modelling: Theoretical Perspective

The previous sections have provided a brief introduction to different approaches to developing urban simulation model systems. This section elaborates more on residential location components of urban simulation models. As briefly mentioned in section 2.2, residential location components of early generation urban simulation models are based on gravity-type formulations. Although the gravity-type models provided a practical approach for modelling residential locations, they have been seriously criticised for their aggregate design and lack of underpinning economic theory. As a result, various different modelling frameworks based on microeconomic theory have been proposed to incorporate behavioural aspects in residential location modelling (see Figure 2-4). The main objective of this section is to review these different modelling frameworks from a theoretical perspective.

![Figure 2-4: Theoretical Frameworks in Residential Location Modelling](image-url)
2.3.1 Gravity-type Models

Gravity-type models illustrate the macroscopic relationships between locations. They provided a powerful tool to model transport flows (i.e., estimate trips between pairs of origins and destinations) and location allocations (i.e., allocating households around workplaces), as discussed by Wilson (1970). The main assumption of Gravity-type models is that the interaction between two locations declines with increasing distance between them (i.e., distance decay function), but is positively associated with the size of activities at each location. If the distance decay function is defined as the inverse of the square distance, the model takes the form of Newton’s law of universal gravitation. The standard form of gravity model for trip distribution takes the following form:

\[ T_{ij} = kO_i D_j f(d_{ij}) \]  

(2-1)

where, \( T_{ij} \) is the total number of trips between zone \( i \) and zone \( j \), \( k \) is a constant, \( O_i \) is the total number of trips originated from zone \( i \), \( D_j \) is the total number of trips attracted by zone \( j \), and \( f(d_{ij}) \) is a distance decay function.

In a location allocation context, Lowry (1964) assumed that the interaction between residence and workplace determines residential location and developed the following spatial interaction model:

\[ R_{ij} = kE_j f(d_{ij}) \]  

(2-2)

\[ P_i = \sum_j R_{ij} = k \sum_j E_j f(d_{ij}) \]  

(2-3)

where, \( R_{ij} \) is the total number of residents living in zone \( i \) and working in zone \( j \), \( k \) is a constant, \( E_j \) is the total number of employment in zone \( j \), \( f(d_{ij}) \) is a distance decay function, and \( P_i \) is the residential population of zone \( i \).

Motivated by the entropy concept in statistical mechanics, Wilson (1970) provided a theoretical basis for the gravity model. The entropy-maximising principle offers a general tool and extends the standard gravity model to incorporate cost constraints as well as other attributes rather than just two size variables. Kapur (1982) considers cost as a function to be minimised and proposed another class of gravity-type model which
is based on the simultaneous maximisation of entropy and minimisation of cost function.

2.3.2 Bid-rent Models

Alonso’s (1960) bid-rent model of the urban land market laid the foundation for microeconomic analysis of urban modelling. Alonso’s bid-rent model explains how, in a city with almost all employment in a single centre (i.e., a monocentric city), households compete for residential land and locate in concentric rings at densities which decline in relation to distance from the centre. Hence, the household’s decision to locate nearer or farther from the central business district (CBD) entails trading off commuting time and land price.

Many authors have attempted to extend this monocentric model by proposing generalised theoretical models that take account of multiple employment centres, local amenities, etc. These models in general result in a very irregular bid-rent surface (see Richardson, 1977a). Monocentric-based models still have a great importance in the empirical analysis of urban forms. Recently, Bertaud and Malpezzi (2003) estimated a series of population density measures based on negative exponential density gradients (i.e., as implied by Alonso model) for 48 major metropolitan areas around the world and concluded that the standard monocentric urban model fits the data quite well.

The dispersion of location choices across a population (or a segment of population) in bid-rent models is only described by transportation costs to the Central Business District (CBD). Standard bid-rent models, therefore, capture the heterogeneity in locational preferences and the willingness to pay for a location in a very limited way. This is partly because the standard bid-rent model does not incorporate localised amenities (e.g., good schools) and disamenities (e.g., high crime rate) as discussed in generalised bid-rent models and also because Alonso’s bid-rent model is founded on the non-stochastic assumption of the price and usage of land.

Ellickson (1981) extended the bid-rent theory and formulated the estimation method of the stochastic bid-rent function. In this approach, the bid function is formulated as the willingness-to-pay for a particular location as following:

\[ B_{nz} = WP_{nz} = I_n - V_{nz}^{-1}(\bar{U}, X_n, p) \]  

(2-4)
where, $V_{nz}$ is the indirect utility function of location $z$, $X_z$ is the vector of attributes of location $z$, $\bar{U}$ is the utility level, $p$ is the price of the composite good, and $I_n$ is the income of household $n$.

Assuming that bid function has an error term which is distributed according to an extreme value distribution, the probability of household $n$ being the best bidder for a location $z$ is given as (see Martínez, 2008):

$$P_{nz} = \frac{\exp(\mu B_{nz})}{\sum_{n=1}^{N} \exp(\mu B_{nz})}$$  \hspace{1cm} (2-5)

Under the auction market assumption, rent of a location is the expected maximum bid for the location, which can be formulated as the logsum of the bids (see Martínez, 2008):

$$r_i = \frac{1}{\mu} \ln\left(\sum_{n'=1}^{N} \exp(\mu B_{n'zs})\right)$$  \hspace{1cm} (2-6)

Many authors have attempted to extend Ellickson’s stochastic bid-rent model by specifying different bid functions and error structures (see Lerman and Kern, 1983; Wang, 2009; Hurtubia et al., 2011).

2.3.3 Location Choice Models

Lerman (1976) and McFadden (1978) adopted discrete choice models for analysing the demand for residential locations. Random utility discrete choice models provide a powerful analytical tool to model residential locations at the level of individual decision makers (i.e., households) and at higher spatial resolutions compared to the early-generation gravity-type models. Discrete choice models have been the dominant approach in modelling residential locations because they allow for the relationship between spatial choices and zonal characteristics as well as the attributes of households. The remainder of this section presents a brief introduction to household location choice models, sections 2.4, 2.5 and 2.6 comprehensively review different aspects of discrete choice modelling in general and the residential location choice context in particular.

In location choice models, households are assumed to choose their residential location so as to maximise their utility, subject to the price they have to pay for a given location (i.e., households are price takers). Assuming that the utility function consists of a
systematic term and an error term (i.e., a Random Utility Model (RUM)) which is distributed according to Gumbel distribution (i.e., a Multinomial Logit (MNL) model), the probability that household $n$ chooses location $z$ is given as:

$$\begin{align*}
P_{nz} &= \frac{\exp(\mu V_{nz})}{\sum_{z=1}^{Z} \exp(\mu V_{i,z})} \\
\end{align*}$$

(2-7)

where, $P_{nz}$, is the probability that household $n$ selects location $z$, and $C_n$ is the choice set of locations available to the household.

### 2.3.3.1 Relationship of Location Choice and Gravity-type Models

Anas (1983b) investigated the similarity of doubly-constrained gravity models derived from entropy maximisation principles with the MNL location choice model derived from random utility theory and proved that these two approaches are in fact equivalent and a behaviourally valid gravity-type model can be estimated from disaggregated data on individual choices. He concluded that the behavioural realism of models estimated by either approach is not determined by any inherent structural property of the models.

### 2.3.3.2 Relationship of Location Choice and Bid-rent Models

Alonso's bid-rent theory and the random utility models of household location choice both share the support of microeconomic theory. Anas (1990) compared the classical bid-rent models with random utility location choice models and illustrated that when the variance of the error term of the utility function tends to zero, the discrete choice location choice model converges towards the monocentric model. He concluded, therefore, that the discrete choice approach can be viewed as an extension to the standard monocentric bid-rent model. Martinez (1992) also investigated the theoretical comparison of Ellickson's stochastic bid-rent model and the discrete location choice models. He demonstrated that in perfectly competitive land markets (i.e., equilibrium condition), these approaches are equivalent; therefore they should be viewed as complementary rather than as alternatives (see Figure 2-5).

Martinez (1992) also formulated the bid-choice model by combining Ellickson's stochastic bid-rent approach and Lerman's choice approach based on the argument that the urban land market is a case with common values. This effectively means that households may behave as close to price takers as desired (i.e., the random utility
approach), without invalidating the assumption that the final price is defined by the highest bidder's bid and the housing market is an auction (i.e., the bid-rent approach).

<table>
<thead>
<tr>
<th>Stochastic Bid-rent Model</th>
<th>Household Location Choice Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maximising Willingness-to-Pay (Bid)</td>
<td>Maximising Utility</td>
</tr>
<tr>
<td>Assign a household to a location (or a dwelling)</td>
<td>Assign a location (or a dwelling) to a household</td>
</tr>
<tr>
<td>Utility levels are adjusted to clear the market</td>
<td>Prices (Rents) are adjusted to clear the market</td>
</tr>
</tbody>
</table>

\[
P_{nz} = \frac{\exp(\mu B_{nz})}{\sum_{n=1}^{N} \exp(\mu B_{nz})}
\]

\[
P_{nz} = \frac{\exp(\mu V_{nz})}{\sum_{z=1}^{Z} \exp(\mu V_{nz})}
\]

**Figure 2-5- Relationship of Location Choice and Bid-rent Models**

### 2.3.3.3 Relationship of Location Choice and Hedonic Price Models

In hedonic price models (see Rosen, 1974), prices (average zonal prices or dwelling prices) are modelled by decomposing the price into its constituent characteristics, such as locational attributes and dwelling attributes, using a regression model. Hedonic price models are used to obtain estimates of the amount that each attribute contributes to the prices.

Discrete choice residential location choice models can also be viewed as an extension to the hedonic housing price model (Bayer et al., 2005). Consider a specification of a utility function without any non-idiiosyncratic heterogeneity in preferences (e.g., household attributes) and endowments (e.g., employment locations) as:
\[ U_z = \beta_1 X_z - \beta_2 P_z + \varepsilon_z \]  

(2-8)

where, \( X_z \) is a vector of observable attributes of alternative \( z \) other than price, and \( P_z \) is the price.

In this case, the equilibrium condition implies that prices adjust so the mean utility of each alternative is identical and we can write:

\[ \forall z : \beta_1 X_z - \beta_2 P_z + \varepsilon_z = K \]  

(2-9)

\[ P_z = \frac{\beta_1}{\beta_1} X_z + \frac{1}{\beta_1} \varepsilon_z \]  

(2-10)

This is equivalent to Rosen's hedonic regression model. Therefore, we can say that the hedonic price model returns the equilibrium prices if households’ preferences are assumed to be only idiosyncratically different.

2.4 Discrete Choice Analysis of Residential Location Choices

As discussed in section 2.3, discrete choice models have been the dominant approach in the modelling of residential locations. This section provides a fuller discussion of the discrete choice approach, including, first, a refresher on random utility theory and discrete choice models, and then, a description of various model structures and estimation techniques that have been applied in residential location choice modelling. Data requirements for developing a residential location choice model are also discussed. Finally, the section sets out the most important and fundamental research challenges in discrete choice analysis of residential location choices. Section 2.5 and 2.6 will go on to describe these challenges in more detail.

2.4.1 Basic Concepts

In general, discrete choice models are concerned with modelling the behaviour of individual decision makers (e.g., individual people, households, firms, etc.) when making choices from a set of discrete alternatives, or the choice set (see Ben-Akiva and Lerman, 1985; Train, 2003). Many choice situations naturally give rise to discrete alternatives (e.g., choice of mode of travel), or to alternatives that can be meaningfully discretised (e.g., choice of residential location in which space is discretised into zones).
Decision makers evaluate different alternatives based on their attributes and apply a decision rule to make a choice among the alternatives. There are a wide range of decision rules that can be applied to develop discrete choice models, such as compensatory decision rules, non-compensatory decision rules, etc. (see Ben-Akiva and Lerman, 1985; Train, 2003). The most common decision rule applied in discrete choice literature is the compensatory decision rule, which assumes that a negative evaluation of an attribute can be compensated by a positive evaluation of another attribute. Microeconomic consumer theory provides the most common framework for developing compensatory decision rules by introducing the concept of utility and assuming that decision makers choose the alternatives that maximise their utilities.

Mathematically speaking, each alternative \( i: 1, \ldots, I \) in the choice set is characterised by a utility \( U_{ni} \) which is unique for each decision maker because of variations in the attributes of alternatives and the characteristics of decision makers, such that decision maker \( n \) will choose alternative \( i \) if and only if \( U_{ni} > U_{nj} \ \forall j \neq i \).

Since we may not be able to observe or measure every attribute of the alternatives and every characteristic of the decision makers, the utility cannot be measured exactly. In discrete choice models based on random utility theory, therefore, we assume that the utility is composed of a deterministic component (i.e., a systematic term) which can be calculated based on the observed attributes of alternatives and characteristics of decision makers, and a stochastic component (i.e., an error term) which is unobserved. Hence, the utility that decision maker \( n \) attributes to alternative \( i \) can be written as:

\[
U_{ni} = V_{ni} + \varepsilon_{ni} \tag{2-11}
\]

where, \( V_{ni} \) represents the measurable component, and \( \varepsilon_{ni} \) represents the random error component, from the modeller’s perspective.

Finally, the probability that decision maker \( n \) will choose alternative \( i \) is given as:

\[
P_{ni} = \text{prob}[U_{ni} > U_{nj}, \forall j = 1, \ldots, I]
\]

\[
= \text{prob}[V_{ni} + \varepsilon_{ni} > V_{nj} + \varepsilon_{nj}, \forall j = 1, \ldots, I]
\]

\[
= \text{prob}[\varepsilon_{nj} - \varepsilon_{ni} < V_{ni} - V_{nj}, \forall j = 1, \ldots, I] \tag{2-12}
\]

26
In order to determine the choice probabilities, we need to make an assumption regarding the distribution of $\varepsilon$. This will result in the development of different model structures in the literature. Reviewing all existing model structures is not within the scope of this study [see Hess (2005) for a comprehensive discussion on this topic], however, the next section briefly describes the model structures that have been used most commonly both in general and in the particular context of residential location choice modelling context.

2.4.2 Important Model Structures

2.4.2.1 The Multinomial Logit (MNL) model

The Multinomial Logit model is based on the assumption that the unobserved part of the utility is distributed according to a type I extreme value (Gumbel) distribution. The MNL choice probability that decision maker $n$ will choose alternative $i$ is given as:

$$P_{ni} = \frac{\exp(V_{ni})}{\sum_{j=1}^{J} \exp(V_{nj})}$$

(2-13)

There are three main reasons that have led to the popularity of the MNL models in empirical residential location choice modelling (Guo, 2004). First, MNL models are computationally very efficient in both estimation and prediction because of their straightforward and closed-form choice probability function. Secondly, MNL models exhibit the IIA (Independence from Irrelevant Alternatives) property which allows the parameters of the model to be estimated consistently using a random sampling technique, as we shall see later in this section. Finally, the IIA property of MNL models implies a proportionate substitution phenomenon which allows for introducing an alternative into or eliminating from the choice set without re-estimating the model.

While the IIA property of MNL models is realistic in some choice situations, it has a limitation where alternatives are close substitutes for each other (see Ben-Akiva and Lerman, 1985). This is an important issue in modelling residential location choices since alternatives are expected to be spatially correlated (see Section 2.5.4).

2.4.2.2 The Multinomial Probit (MNP) model

If we assume that $\varepsilon$ is distributed according to a multivariate normal distribution, as would seem natural, the resulting model is the Multinomial Probit (MNP) model. MNP
models can accommodate any correlation structure among unobserved attributes of the alternatives, hence they can represent any substitution pattern. The only limitation of MNP models is that all unobserved components of utility are assumed to be normally distributed. Although representing error terms by normal distributions is appealing in most cases, this might be inappropriate in situations where the error term is constrained to be positive (see Train, 2003).

Unfortunately, MNP choice probabilities do not have a closed-form expression, hence, the estimation of MNP models requires the numerical approximation of a multidimensional integral. This will result in a very high computation cost for estimation of MNP models which essentially limits the use of MNP models to choice situations with a very small number of alternatives. This has motivated researchers to propose different distributions for unobserved attributes of the alternatives in order to reduce the estimation cost of MNP models and relax the IIA property of MNL models.

2.4.2.3 The GEV class of models

The GEV class of models extend the MNL model by relaxing the independence assumption between the error terms of alternatives. McFadden (1978) identified a class of distributions which allows for various levels of correlation among the unobserved part of utility. The choice probabilities for this class of distributions take a closed-form expression. Considering the following multivariate Generalised Extreme Value (GEV) cumulative distribution function:

\[ F(\varepsilon_1, ..., \varepsilon_l) = \exp(-G(e^{-\varepsilon_1}, ..., e^{-\varepsilon_l})) \]  \hspace{1cm} (2-14)

Under certain conditions\(^9\) for the generating function \(G\), the choice probabilities for alternative \(i\) are given by:

\[ P_{ni} = \frac{Y_i G,(Y_1, ..., Y_l)}{G(Y_1, ..., Y_l)} \]  \hspace{1cm} (2-15)

where,

\[ Y_i = \exp(V_i) \]  \hspace{1cm} (2-16)

\(^9\) Certain conditions are required to ensure that \(F(\varepsilon)\) is indeed a cumulative distribution function [see Train (2003) for a comprehensive discussion].
MNL, Nested Logit, and many other choice models can be shown to be special cases of GEV class models involving appropriate specification of the generating function $G$ (see Train, 2003). The most commonly used member of the GEV class of models is the Nested Logit model in which alternatives are divided into nests of alternatives in a tree form and where the alternatives within a nest are assumed to be correlated (see Ben-Akiva and Lerman, 1985). In fact, many other members of the GEV family can also be represented by a nesting structure in a tree form, such as Cross Nested Logit (CNL), Paired Combinatorial Logit (PCL), and Generalised Nested Logit (see Vovsha, 1997; Koppelman and Wen, 2000; Wen and Koppelman, 2001). The GEV class models are not restricted to those that can be represented by a tree form of nesting structure, however. Karlstrom (2001), for example, specified an arbitrary $G$ and showed that the goodness of fit of the resulting model was better than other conventional model structures such as MNL and Nested Logit. Daly and Bierlaire (2003) also proposed a network-based framework to characterise the underlying correlation structure.

### 2.4.2.4 The Mixed Logit model

Similar to GEV class models, the Mixed Logit model generalises the MNL model to accommodate for flexible substitution patterns. The Mixed Logit (or Logit kernel or Mixed MNL) model is derived from the assumption that the unobserved portion of utility consists of a part that follows any distribution that allows flexible substitution patterns across alternatives as well as a part that is distributed according to a Gumbel distribution (i.e., error components derivation\(^\text{10}\)). Unlike the GEV class of models, the choice probabilities of the Mixed Logit model do not have a closed-form expression and involve integration of the MNL choice probabilities over the distribution of the additional error term (see Train, 2003). The utility of the Mixed Logit model is given by:

$$U_{ni} = V_{ni} + \eta_{ni} + \varepsilon_{ni}$$  \hspace{1cm} (2-18)
where, $\eta_{ni}$ is the additional error (distributed over the alternatives) with the probability density function of $f(\eta_n)$ which is specified by the researcher. The choice probability of the Mixed Logit model takes the following form for the given distribution of $\eta$:

$$P_{ni} = \int_{\eta_n} \frac{\exp(V_{ni} + \eta_{ni})}{\sum_{j=1}^{J} \exp(V_{nj} + \eta_{nj})} f(\eta_n) \, d\eta_n$$

Mixed Logit models can approximate any discrete choice model derived from random utility maximisation (McFadden and Train, 2000), however, similar to MNP models, the parameters of the Mixed Logit models need to be estimated with the help of simulation estimation methods (see Bhat, 2001; Sivakumar et al., 2005). The mixing approach described here for MNL models can also be extended for the GEV class of models. The resulting model is called a Mixed GEV (or GEV kernel) model in which the mixing distribution is superimposed on the GEV model structure (see Hess et al., 2005).

### 2.4.3 Stated Preferences vs. Revealed Preferences

In general, two types of data can be used to estimate discrete choice models, i.e., revealed preferences and stated preferences data. Revealed preferences data relates to people’s actual choices (i.e., revealed choices) in real situations. Stated preferences data are data collected in experimental or survey situations where respondents are presented with hypothetical choice situations (see Train, 2003).

Residential location choice models are conventionally estimated based on revealed preferences data or households’ observed residential choices. Despite the fact that revealed preferences data reflect households’ actual choices, it can be argued that households’ revealed choices might not be their optimum choices due to high mobility and transaction costs in the housing market. This may result in a deviation from the utility maximisation framework. Alternatively, the estimation sample can be defined as recent mover households in order to make sure that they are located at their optimum locations. The estimation sample in this case, however, will not be representative of the whole population, which may result in biased parameter estimates for residential location choices of the entire population. Furthermore, modelling households’ residential location preferences based on revealed preference data may be biased since revealed preferences data are shaped by compromises made between what is desired and what is available in the market (Pagliara et al., 2010).
Some authors have suggested using stated preferences data in order to arrive at a more reliable estimate of the utility of households' residential location preferences, (see Hunt et al., 1994; Marcucci et al., 2011; Torres et al., 2013). Stated preferences data is deficient, however, in that people often do not actually do what they say they would do in hypothetical choice situations. Given the uncertainty that exists about how people respond to the survey in the stated preferences approach, it is often argued that revealed preferences data are more suitable for modelling demand (see Ben-Akiva and Lerman, 1985).

2.4.4 Model Estimation

The choice probability of MNL and GEV models takes a closed form as described earlier; hence, the estimation of the parameters of these models can be conducted based on the Maximum Likelihood (ML) method. The conventional approach for estimation of complex econometric models involving intractable integrals such as MNP and Mixed Logit models is based on the Maximum Simulated Likelihood (MSL) approach (see Bhat, 2001; Sivakumar et al., 2005). The main barrier to the application of complex discrete choice models for spatial choices, such as the MNP and Mixed Logit model, is the computation cost of this MSL approach since the number of dimensions of integration increases with the number of alternatives.

Recently, Bhat (2011b) introduced the Maximum Approximate Composite Marginal Likelihood (MACML) estimation approach for both MNP and Mixed Logit models which is based on the analytical approximation of the multidimensional integral (instead of the numerical approximation in the conventional MSL approach). The MACML approach requires much less computation effort compared to the MSL approach and allows for estimation of several models that were infeasible to estimate using the MSL approach. Estimation of Mixed Logit models using the MACML approach requires a normal scale mixture representation for the extreme value error terms, and adds an additional layer of computation effort (Bhat, 2011a).

Despite recent advances in computing power, estimation of residential choice models is still burdensome due the massive size of the universal choice set. McFadden (1978) proved that the IIA property of MNL models permits consistent estimation of such models based on a random sample of the alternatives.
Let $D_n$ be a random sample of the decision maker $n$’s choice set $C_n$ ($D_n \subseteq C_n$), and let $\pi_n(D_n|j)$ be the probability that subset $D_n$ is sampled. According to Bayes’ theorem of conditional probability, the probability of choosing alternative $i$ by decision maker $n$ is:

$$P(i|D_n \subseteq C_n) = \frac{\pi_n(D_n|i) \cdot P_{i,n}}{\sum_{j \in D_n} \pi_n(D_n|j) \cdot P_{j,n}}$$  \hspace{1cm} (2-20)

The conditional probability $P(i|D_n)$ exists if $\pi_n(D_n|j) > 0$, $\forall j \in D_n$. By substituting the MNL choice probability in this expression, we obtain:

$$P(i|D_n \subseteq C_n) = \frac{e^{V_{i,n} + \ln(\pi_n(D_n|i))}}{\sum_{j \in D_n} e^{V_{j,n} + \ln(\pi_n(D_n|j))}}$$  \hspace{1cm} (2-21)

The term $\ln(\pi_n(D_n|j))$ in the above equation represents the alternative-specific correction terms that are to be computed based on the sampling probability which are added to the utility of each alternative prior to estimation in order to achieve an unbiased estimation of parameters. It should be noted that the simplification in deriving the above equation is based on the cancellation of the denominator $P_{i,n}$ in the MNL model. This simplification does not occur for non-MNL models. A simple random sampling strategy is characterised by the following uniform conditioning property:

$$\pi_n(D_n|i) = \pi_n(D_n|j), \ \forall i, j \in D_n$$  \hspace{1cm} (2-22)

This property implies that the alternative-specific correction terms are equal and therefore cancel out in choice probabilities. Thus, in the formal MNL model, simple random sampling of alternatives allows a consistent estimate of parameters.

Ben-Akiva and Lerman (1985) argued that the simple random sampling approach is not necessarily an efficient approach for the estimation of discrete choice models with large choice sets since the majority of the alternatives may have very small choice probabilities. They suggested, therefore, applying the importance sampling approach based on the preliminary estimates of the choice probabilities in order to estimate the parameters of the model more efficiently. It should be noted that if importance sampling is adopted, the alternative-specific correction terms should be computed and added to the utility function according to the sampling weights (see Ben-Akiva and Lerman, 1985; Frejinger et al., 2009).
Recent work by Guevara and Ben-Akiva (2013) extends the sampling approach for the GEV family of choice models and provides a consistent estimator for GEV class models on a subset of alternatives. However, there is no theoretical support examining how sampling of alternatives in Mixed Logit models affects the empirical accuracy and extent of bias of the estimated parameters (Azaiez, 2010). Some authors have therefore used numerical experiments to examine the effect of the sampling of alternatives in the parameters estimated by Mixed Logit models. These numerical results show that the sampling of alternatives does not affect significantly the estimation of parameters in Mixed Logit models (see Nerella and Bhat, 2004; Azaiez, 2010; Guevara and Ben-Akiva, 2012; Lemp and Kockelman, 2012).

2.4.5 Research Challenges in Modelling Residential Location Choices

One of the fundamental issues in modelling residential location choices is associated with the appropriate representation and the level of aggregation of alternatives. The increasing availability of high-resolution spatial data, such as parcel data, in recent years has motivated researchers to develop models at higher resolutions (see Waddell, 2009). Arguing that households actually choose individual dwelling units rather than aggregated spatial units, some authors have also attempted to model residential location choices at the level of dwelling units (see Habib and Miller, 2009). Section 2.5 investigates previous studies based on their level of aggregation of alternatives and sheds light on the advantages and disadvantages of these different approaches.

Some authors have also tried to improve the specification of the choice process by accommodating complex patterns of unobserved spatial correlation amongst residential alternatives (see Guo and Bhat, 2004a; Sener et al., 2010). Section 2.5, therefore, also reviews different model structures proposed in the literature to capture spatial structure effects in residential location choice context.

Another fundamental issue in the modelling of residential location choices is associated with the large number of alternatives which is threefold: (i) the computational burden of estimation and application of residential location choice models, (ii) the increase in the stochastic noise due to flattening of choice probability distributions, (iii) the choice set formation problem as the number of alternatives evaluated is constrained by individuals’ limited capacity for gathering and processing information (Fotheringham et
al., 2000). Section 2.6 investigates these different issues and specifically focuses on different approaches proposed in the literature to tackle the choice set formation problem in a residential location choice context.

2.5 Representation of Alternatives and Spatial Structure Effects

Household surveys in general provide information on the chosen alternatives and the characteristics of households. As discussed earlier, however, in order to model residential location choices, the attributes of chosen and non-chosen alternatives are also required. The attributes of chosen and non-chosen alternatives are usually derived from different data sources (see Figure 2-6). Depending on the availability of these data sources, and the level of spatial granularity of the datasets, different models have been developed at different levels of aggregation of alternatives.

Most residential location choice models represent alternatives as zones and incorporate spatially aggregated data across zones in the modelling. These models are referred to in the literature as zonal-based, subarea-based, or meso-scale residential location choice models. The availability of detailed spatial data in recent years has enabled researchers to model residential locations by representing individual entities in the real world (e.g., land parcels and buildings) as residential alternatives. These models are referred to as parcel-based, object-based, or micro-scale, residential location choice models. There have also been some attempts in the literature to model residential location choices at the most disaggregate level of dwelling units.

Although recent developments in this area all suggest modelling residential location choices at the parcel-level, or even at the dwelling-level, there is no empirical evidence demonstrating the effect of level of aggregation of alternatives on the prediction performance of residential location choice models. In order to shed light on this fundamental issue, we have here classified the existing residential location choice models in the literature based on their level of aggregation of alternatives.

The workhorse of residential location choice modelling is the Multinomial Logit model (MNL). However, the association of alternatives to space in residential location choice models violates the IIA assumption of the MNL model since it is expected that the closer alternatives will be more similar than the farther alternatives, according to Tobler’s (1970)
first law of geography. Potential spatial effects, such as the spatial correlation among residential alternatives, should therefore be considered in the modelling of residential location choices.

This section first reviews the existing residential location choice models in the literature based on their level of aggregation of alternatives. Then, different model structures which have been proposed to incorporate spatial structure effects are reviewed.

![Diagram of Explanatory Variables and Data Sources Used in Modelling of Residential Location Choices](image)

*Figure 2-6: Explanatory Variables and Data Sources Used in Modelling of Residential Location Choices*
2.5.1 Zonal-based Models

The zonal-based approach involves aggregating elementary alternatives (i.e., dwelling units) into spatially defined alternatives such as Travel Analysis Zones (TAZs), Census Tracts, etc. Since alternatives are spatial units in zonal-based models, locational attributes can be directly specified in the utility function, subject to consistency in the level of aggregation of variables (locational attributes that are available at lower spatial granularity should be aggregated to the level of aggregation of the spatial units that are assumed to be alternative locations). On the other hand, in order to be able to incorporate dwelling attributes in the utility function of zonal-based models, they need to be aggregated to the level of spatial units in order to be consistent with other locational attributes. As a result of this spatial aggregation, the variability of dwelling attributes in spatial units is lost. This is effectively equivalent to assuming that dwelling units are homogeneous within each zone.

There are many studies focused on empirical analysis of various determinants of household residential location choice decisions (see Farley, 1995; Timmermans et al., 1996; Waddell, 1996; Abraham and Hunt, 1997; Clark and Ware, 1997; Ben-Akiva and Bowman, 1998; Nechyba and Strauss, 1998; Molin et al., 1999; Sermons and Koppelman, 2001; Barrow, 2002; Srour et al., 2002; Guo and Bhat, 2007a; Cho et al., 2008). Guo (2004) and Schirmer et al. (2012) comprehensively reviewed previously developed zonal-based residential location choice models from an empirical perspective and Table 2-2 summarises some of the influencing attributes of residential location choice decisions alongside their effects on the utility function.

In order to have a model that is not sensitive to the level of aggregation of elementary alternatives, modellers have incorporated the number or quantity of elementary alternatives, known as size variables (e.g., number of dwellings in zones), in the utility function.\textsuperscript{11} The problem with incorporating the size variables is that it requires the representative utility to be non-linear with respect to size variable parameters, as discussed in the following (see Train, 2003).

\textsuperscript{11} Another problem in modelling spatial choices such as choices of residential locations is the Modifiable Areal Unit Problem (MAUP). The MAUP is known as the sensitivity of the statistical inference of a spatial phenomenon to the way in which the zones are defined. This is a long-standing issue in spatial analysis and a general solution to the problem does not yet exist (see Openshaw, 1977; Guo and Bhat, 2004b).
Table 2-2 – Previous Research Findings in Zonal-based Residential Location Choice Models

<table>
<thead>
<tr>
<th>Attributes</th>
<th>Effects on Utility</th>
<th>Study</th>
</tr>
</thead>
<tbody>
<tr>
<td>Population Density</td>
<td>Positive</td>
<td>Ben-Akiva and Bowman, 1998; Srour et al. 2002; Guo and Bhat, 2004</td>
</tr>
<tr>
<td>School Quality</td>
<td>Positive</td>
<td>Barrow, 2002</td>
</tr>
<tr>
<td>Crime Rate</td>
<td>Negative</td>
<td>Ben-Akiva and Bowman, 1998; Nechyba and Strauss, 1998</td>
</tr>
<tr>
<td>Housing Price</td>
<td>Negative</td>
<td>Abraham and Hunt, 1997</td>
</tr>
<tr>
<td>Housing Space</td>
<td>Positive</td>
<td>Molin et al. 1999</td>
</tr>
<tr>
<td>Commute time to Workplace</td>
<td>Negative</td>
<td>Waddell. 1996; Sermons and Koppelman, 2001</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Molin et al. 1999; Guo and Bhat, 2004</td>
</tr>
<tr>
<td>Accessibility to Employment</td>
<td>Positive</td>
<td>Timmermans et al. 1996; Cho et al. 2008</td>
</tr>
<tr>
<td>Accessibility to Shopping</td>
<td>Mixed findings</td>
<td>Timmermans et al. 1996; Guo and Bhat, 2004 find it positive, but</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Hunt et al. 1994 and Srour et al. 2002 found it negative.</td>
</tr>
<tr>
<td>Accessibility to Recreational Facilities</td>
<td>Negligible</td>
<td>Guo and Bhat, 2007</td>
</tr>
<tr>
<td>Accessibility to Open Space</td>
<td>Positive</td>
<td>Farely, 1995; Clark and Ware, 1997</td>
</tr>
</tbody>
</table>

Consider a region which comprises $Z$ zones $z: 1, \ldots, Z$, and in which zone $z$ contains $D_z$ dwellings (i.e., size variable). If the combined zone of the two zones $i$ and $j$ is labelled zone $c$ the number of dwellings in the combined zone is necessarily the sum of that in the two original zones (i.e., $D_i + D_j = D_c$). If the attractiveness of a zone depends only on the number of dwellings in a zone, the model should give a probability of choosing the combined zone which is the same as the sum of the probabilities of choosing the two original zones. Therefore, the model must satisfy:

$$P_i + P_j = P_c$$  \hspace{2cm} (2-23)

This equality holds only when (assuming MNL choice model):

$$\exp(V_i) + \exp(V_j) = \exp(V_c)$$  \hspace{2cm} (2-24)

We can now write:

$$\exp(\ln(\alpha D_i)) + \exp(\ln(\alpha D_j)) = \exp(\ln(\alpha D_c))$$  \hspace{2cm} (2-25)

$$D_i + D_j = D_c$$  \hspace{2cm} (2-26)
Therefore, the parameter of the size variable $D$ should be specified inside an $\ln$ operation in order to develop a model that is insensitive to the level of granularity. Since the choice probabilities remain unchanged by adding a constant (i.e., $\ln(\alpha)$), the parameter $\alpha$ cannot be identified and should be set to a constant number (usually one). Daly (1982) extended the size variable approach for multiple size variables; in this case, the coefficient of one of the size variables should be set to one.

de Palma et al. (2005) proposed an importance sampling approach to estimate the parameters of a utility function incorporating one size variable more efficiently. They argued that if the probability that alternative $k$ is included in the sampled choice set is proportional to $D_k$, the alternative-specific correction term (which should be added to the utility function in order to obtain consistent parameter estimates based on an importance sampling protocol), and the term $\ln(D_k)$ exactly compensate. Hence, no correcting factor is necessary to obtain consistent estimates.\(^{12}\)

Lerman (1975) and McFadden (1978) also considered the spatial aggregation problem in the case of a non-homogeneity assumption and introduced a correlation parameter that measures the level of similarity among unobserved attributes of dwellings in an aggregated spatial unit. The main disadvantage of this generalisation is the need to retain the zonal definitions to define the correlation parameter (Ben-Akiva and Watanatada, 1981).

While zonal-based models capture spatial distribution of residential locations reasonably well, they have been criticised by some authors. Quigley (1985) argued that the aggregated alternatives model in most cases is only an approximation of the “theoretically correct choice model”, where dwelling units are considered as a choice. Besides, the heterogeneity of households’ residential preferences with respect to dwelling attributes are as important as the location attributes, hence ignoring the variability of dwelling attributes in zonal-based models may reduce the predictive ability of the model.

\(^{12}\) It should also be noted that alternative-specific constants are not included in estimation of residential location choice models for practical reasons as there are usually a large number of alternatives involved (see Anas, 1982).
2.5.2 Parcel-based Models

As discussed earlier, the spatial scales of zonal-based models are usually too coarse to evaluate policies applied at the fine spatial scales. Recently, therefore, there have been some attempts to develop residential location choice models at the level of land parcels or buildings.

Lee et al. (2009) and Lee and Waddell (2010) developed a parcel-based residential location choice model for the Puget Sound region. Availability of x-y coordinates of all households’ residential locations in the household survey enabled the authors to assign each household to a parcel. This was accomplished by taking advantage of parcel data from county tax assessment offices. This parcel data includes information on the number of dwellings in the parcel as well as average values of dwelling attributes. The locational attributes available at the TAZ level, such as accessibility, are assigned to the matching parcels. The universal choice set in this study comprised 986,157 parcels covering the whole study area, and the model estimate is achieved based on the random sampling of alternatives strategy. It should be noted that parcel-based models also incorporate many locational attributes such as workplace travel time, school quality, etc. that are only defined or available for zones.

While parcel-based models are useful in analysing policies that require a higher spatial resolution, the difficulties of working with high-resolution spatial data in terms of computation time, as well as the lack of availability of parcel data for many metropolitan regions, inhibit wide use of parcel-based models (Waddell, 2009).

2.5.3 Dwelling-level Models

Similar to zonal-based models, models developed at the level of parcels require a restrictive assumption that all dwelling attributes are the same for a given parcel. In order to incorporate the variability of dwelling attributes and circumvent the aggregation bias of spatially aggregated models, some authors propose to model residential location choice at the level of dwelling units. Since dwelling units are intrinsically immobile, when a household chooses a dwelling unit, it also chooses the location (zone or parcel) that comprises the dwelling unit.
The spatial resolution of models developed at the dwelling-unit level is limited by the level of spatial granularity of available locational attributes (e.g., parcel-level data or zonal-level data), as well as the accuracy of spatial locations of dwelling units in the household survey. Dwelling-level models can therefore be classified into two categories based on their level of spatial granularity: (a) parcel-based dwelling-level models in which alternatives are dwelling units that are identified by their parcels in space, and (b) zonal-based dwelling-level models in which alternatives are dwelling units that are identified by their zones in space.

Parcel-based dwelling-level models are the most disaggregate residential location choice models as alternatives are dwelling units and locations of dwelling units are determined by dwellings' parcels. Similar to parcel-based models, however, parcel-based dwelling-level models also require parcel data, which are not available for many metropolitan regions. It should be noted that parcel-based dwelling-level models also include zonal-level location attributes for variables that are only defined or available for zones such as school quality, travel time to workplace locations, etc.

While alternatives in zonal-based dwelling-level models are dwelling units, the location of dwelling units are determined by the zones. As a result, zonal-based dwelling-level models cannot be used to analyse phenomena and policies that require a higher spatial resolution, as mentioned earlier.

Zhou and Kockelman (2008) estimated a parcel-based dwelling-level residential location choice for Austin, Texas. Household characteristics, as well as dwelling preferences, were acquired from a survey of Austin movers conducted in 2005. Dwelling locations were determined by the parcels to which they belong. The locational attributes were available at the school district level and were assigned to the parcels within each district. In order to generate the choice set, they took a sample from all chosen dwelling units in the entire household survey, which in essence means that the source of non-chosen alternatives and chosen alternatives was the same. This approach, however, could potentially introduce biases in the parameter estimates of the model, as will be discussed in more detail in chapter 4.

Habib and Miller (2009) developed a parcel-based dwelling-level model of residential location choice for the Greater Toronto Region (GTA). Households' characteristics, as
well as dwelling preferences, were acquired from a retrospective residential search survey (RSS) which was administered as a mail-back questionnaire survey. Dwelling locations were determined by their parcels and the locational attributes were available at the level of Census Tracts and were assigned to the matching parcels. To construct the universal choice set, Habib and Miller (2009) used independent dwelling supply data containing 262,669 dwellings advertised for sale during an eight-year period, obtained from the Toronto Real Estate Board (TREB). This approach requires a very rich disaggregated dwelling supply data in order to determine dwelling locations at the parcel level.

Guevara (2010) developed a zonal-based dwelling-level model of residential location choice for Lisbon Metropolitan Area (LMA) in which the dwelling locations were determined by zones rather than parcels. Household characteristics, as well as dwelling preferences, were acquired from a small convenient online survey (SOUTR) conducted by Martínez et al. (2010). This survey collected information on residential locations, dwelling attributes and household characteristics from 750 households across the entire LMA. Dwelling locations were determined by Freguesia.13 Guevara (2010) used an independent dwelling supply data containing 12,358 dwellings advertised for sale in Feb 2007 obtained from Imokapa (www.imokpa.com) in order to construct the universal choice set. Another important zonal-based dwelling-level residential location choice model was developed by Bayer et al. (2005). Table 2-3 summarises the previous studies on residential location choice modelling discussed above.

All models developed at the level of dwelling units are based on the assumption that households choose a unique type of dwelling in a certain location. Since we cannot observe all of the attributes of a dwelling that make it differentiated from other dwellings, some authors have followed the specification of Berry, Levinsohn and Pakes (1995), and incorporated alternative-specific unobservable attributes in the utility function (see Bayer et al., 2005; Guevara and Ben-Akiva, 2006). Estimation of such models is accompanied by the endogeneity problem, due the correlation of the omitted attributes with the dwelling prices.

13 Freguesia is a secondary local administrative unit in Portugal.
14 Endogeneity refers to situations where observed explanatory variables are correlated with error terms, so that standard estimation procedures that rely on independent errors cannot be used directly.
### Table 2-3 – Different Levels of Aggregation of Residential Alternatives in Developed Models

<table>
<thead>
<tr>
<th>Level of Aggregation</th>
<th>Study</th>
<th>Alternatives</th>
<th>Locations</th>
<th>Universal Choice Set</th>
<th>Dwelling Attributes</th>
<th>Zonal Attributes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zonal-based</td>
<td>Most of Studies in Residential Location Choice Modelling</td>
<td>Zones (TAZs)</td>
<td>Zones (TAZs)</td>
<td>All zones in the region</td>
<td>Zonal Averages</td>
<td>Zones (TAZs)</td>
</tr>
<tr>
<td>Parcel-based</td>
<td>Lee and Waddell, 2010; Lee et al. 2009</td>
<td>Parcels</td>
<td>Parcels</td>
<td>All parcels in the region</td>
<td>Parcel Averages</td>
<td>Zones (TAZs)</td>
</tr>
<tr>
<td>Parcel-based</td>
<td>Zhou and Kockelman, 2008</td>
<td>Dwelling Units</td>
<td>Parcels</td>
<td>Sampled from the household survey</td>
<td>Disaggregate Measure</td>
<td>Zones (School District)</td>
</tr>
<tr>
<td>Parcel-based</td>
<td>Habib and Miller, 2009</td>
<td>Dwelling Units</td>
<td>Parcels</td>
<td>Sampled from independent dwelling supply data</td>
<td>Disaggregate Measure</td>
<td>Zones (Census Tracts)</td>
</tr>
<tr>
<td>Zonal-based</td>
<td>Guevara, 2010; Bayer et al. 2005</td>
<td>Dwelling Units</td>
<td>Zones</td>
<td>Sampled from independent dwelling supply data</td>
<td>Disaggregate Measure</td>
<td>Zones</td>
</tr>
</tbody>
</table>

#### 2.5.4 Model Structures and Spatial Structure Effects

Alternatives in residential location choice models are either spatial units or dwelling units augmented to spatial units. It is expected, therefore, that the closer alternatives will be more similar than the farther alternatives according to the first law of geography, which is stated by Tobler (1970) as: everything is related to everything else, but near things are more related than distant things. Despite its advantages, a simple MNL model is consequently not a good candidate to model residential location choice because it cannot incorporate spatial structure effects such as spatial autocorrelation.

The spatial econometric literature identifies three spatial structure effects that need to be considered in modelling: (i) spatial autocorrelation, (ii) spatial heteroscedasticity, and (iii) spatial heterogeneity. Ignoring these spatial structure effects in econometric modelling will result in mis-estimated standard errors in linear models and inconsistent parameter estimation and structural instability in non-linear models (see Anselin, 1988; LeSage and Pace, 2009).
In a residential location choice context, spatial autocorrelation refers to the potential correlation among the error terms of nearby spatial units. For example, consider a zone with a very high attractiveness (which is not observed by the analyst), other zones adjacent (or close) to this zone have the additional benefit of being adjacent (or close) to the attractive zone. Since this unobserved attractiveness decreases with distance to the highly attractive zone, the errors from nearby alternatives are likely to be correlated (Guo, 2004).

Guo and Bhat (2004a) developed a zonal-based residential location choice model for Dallas County, Texas based on the Generalised Extreme Value (GEV) family models in order to capture the potential spatial autocorrelation among contiguous alternatives. Later, Sener et al. (2010) extended Guo and Bhat’s (2004a) model by relaxing the model structure to capture the spatial autocorrelation across various alternatives which may not necessarily be neighbours. These models are derived from the following GEV generator function:

\[
G(e^{V_{n1}}, \ldots, e^{V_{ni}}) = \sum_{i=1}^{l} \sum_{j=i+1}^{l} \left( \left( \alpha_{i,ij} e^{V_{ni}} \right)^{1/\mu} + \left( \alpha_{j,ij} e^{V_{nj}} \right)^{1/\mu} \right)^{\mu}
\]

where,

\[V_{ni}\] is the systematic utility function of alternative \(i\) perceived by household \(n\),

\(\mu\) is a dissimilarity parameter capturing the correlation between alternatives,

\(\alpha_{i,ij}\) is an allocation parameter representing the portion of alternative \(i\) assigned to the nest \(ij\).

The allocation parameter (function) in Guo and Bhat (2004a) is defined as the proportion of number of contiguous alternatives as:

\[
\alpha_{i,ij} = \frac{\omega_{ij}}{\sum_k \omega_{ik}}
\]

where, \(\omega_{ij}\) is 1 if zone \(i\) is contiguous to zone \(j\), and 0 otherwise.

In Sener et al. (2010), meanwhile, the allocation function takes a more general form in order to accommodate spatial autocorrelation among alternatives that are close, but not
necessarily contiguous to one another. Sener et al. (2010) used a distance decay function for the empirical application of their GSCL model as follows:

\[ \alpha_{i,j} = \frac{d_{ij}^\phi}{\sum_k d_{ik}^\phi} \]  \hspace{1cm} (2-29)

where,

\( d_{ij} \) is the distance between zone \( i \) and \( j \),

\( \phi \) is parameter of the allocation function.

The main advantage of the GEV formulation is that it allows relaxations of the independence assumption among alternative error terms while maintaining closed-form expressions for the choice probabilities. The correlations among alternatives are captured in this class of models in a very limited way, however, based on a pre-assumed correlation structure. Hence, some authors suggest applying more general model structures such as the Multinomial Probit (MNP) model and the Mixed Logit model in order to provide a more general correlation structure in spatial discrete choice models (see Bolduc, 1992; Ben-Akiva and Bolduc, 1996; Vichiensan et al., 2005).

Vichiensan et al. (2005) developed a zonal-based residential location choice model, based on a Mixed Logit formulation, for the city of Sendai in Japan. They compared different configurations of zones based on their distances in order to investigate the effect of spatial autocorrelation in each situation. They also limited the number of zones to five in order to speed up the computation. Although the structure of Mixed Logit and MNP models provides a flexible modelling framework to capture spatial autocorrelation, their application in practice is limited, compared to GEV models, because of the computation costs involved in estimating and applying these models.

Some authors have introduced a similarity factor into the systematic part of the utility function in order to incorporate the dependency of alternatives that exist in the spatial dimension. These models are context-specific and are mainly applied in route choice and destination choice problems (see Schuessler and Axhausen, 2007). This approach has also been extended in the residential location choice context based on dominance variables (Cascetta et al., 2007).
Another important spatial structure effect that should be considered in modelling is spatial heteroscedasticity, which refers to differences in variances of the error terms across spatial units. Since spatial units may differ in important characteristics, homoscedasticity is an assumption that may not hold in a residential location choice context. For example, it is unlikely that a household associates a very high level of utility with a zone with the high crime rate (assuming that the crime rate is unobserved). Therefore, the variability in the utilities (i.e., variance of the error term) that a household associates with a zone with a low crime rate is expected to be more than a zone with a high crime rate.

Spatial heterogeneity is another spatial structure effect described earlier and refers to differences in the data-generating process in relation to space due to location-specific effects (Fotheringham et al., 1996). If spatial heterogeneity exists, a single global relationship (i.e., model parameters) in a study region may not reflect the process appropriately in any local part of the study region. Fotheringham and Brunsdon (1999) described two potential sources for these location-specific variations as: (i) intrinsic behavioural differences in the process across spatial units, (ii) lack of information (from the analyst's perspective) regarding some process-related or spatial-unit related attributes.

Spatial heterogeneity is usually addressed in the geography literature for linear models based on a multi-level modelling paradigm. In linear models, spatial heterogeneity refers to differences in relationships (i.e., model parameters) between the dependent variable of interest and the independent variables across spatial units (see Bullen et al., 1997). There are also a few studies that have examined spatial heterogeneity for non-linear models by extending the multi-level modelling paradigm for non-linear cases (see Bhat, 2000; Bhat and Zhao, 2002).

In residential location choice models developed at the level of dwelling units, spatial heterogeneity can potentially exist and can be seen as the differences in households' taste in relation to different attributes of dwellings across spatial units. In location choice problems, however, the elementary alternatives (e.g., dwelling units) are clustered into spatial zones, unlike previous studies where decision makers are clustered into zones.
Although, as discussed earlier, some models have been developed to capture spatial autocorrelation in zonal-based residential location choice contexts, other spatial structure effects, namely spatial heteroscedasticity and spatial heterogeneity, have not been addressed in the literature. In particular, the treatment of spatial structure effects in models developed at the dwelling-unit level has not previously been examined. Chapter 4 elaborates more on this issue.

In addition to spatial structure effects, some authors have developed residential location choice models based on Mixed Logit or Mixed GEV formulations in order to accommodate unobserved taste heterogeneity across decision makers (see Guo and Bhat, 2004a; Habib, 2009). Table 2-4 summarises the different sources of spatial correlation and spatial heterogeneity incorporated in residential location choice models, both in functional form and in model structure.

Most residential location choice models (including all the models discussed earlier) are static, in the sense that they do not capture the temporal dimension of location choice decisions. Taking advantage of panel-retrospective data, Habib and Miller (2009) acknowledged the temporal dimension of location choice decisions and developed a residential location choice model that takes into account the effects of households’ previous locations in modelling.

<table>
<thead>
<tr>
<th>Source</th>
<th>Representation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed Taste Heterogeneity</td>
<td>Functional Form: Interaction of Dwelling Attributes with Households’ Attributes</td>
</tr>
<tr>
<td>Unobserved Taste Heterogeneity</td>
<td>Model Structure: Random Parameters across Households</td>
</tr>
<tr>
<td>Observed Spatial Heterogeneity</td>
<td>Functional Form: Locational Attributes</td>
</tr>
<tr>
<td>Unobserved Spatial Heterogeneity</td>
<td>Not developed yet</td>
</tr>
<tr>
<td>Observed Spatial Correlation</td>
<td>Functional Form: Dominance Variables</td>
</tr>
<tr>
<td>Unobserved Spatial Correlation</td>
<td>Model Structure: Spatial Autocorrelation (GEV, Mixed Logit, MNP)</td>
</tr>
<tr>
<td>Spatial Heteroscedasticity</td>
<td>Not developed yet</td>
</tr>
</tbody>
</table>

Table 2-4 – Observed and Unobserved Correlation and Heterogeneity in Residential Location Choice Models
2.5.5 Concluding Remarks

Most of the previous works in discrete choice modelling of residential location have been developed at the level of aggregated spatial alternatives mainly because of a lack of disaggregated dwelling supply data. These models suffer from aggregation bias and cannot consistently capture the heterogeneity of households in respect to dwelling attributes in choosing a residential location. Very few studies acknowledge that people choose dwelling units, rather than zones, and thus develop residential location choice models at the level of dwelling units. Overall, there are still many methodological and empirical challenges involved in developing state-of-the-art, dwelling-level, residential location choice models.

From a methodological perspective, relaxing the IIA assumption and incorporating spatial effects, such as spatial autocorrelation, has been one of the research agendas in zonal-based residential location choice models. Such models developed at the dwelling-level are deficient, however, in that they do not address the potential spatial effects arising from the association of dwelling units to space. It should be borne in mind that ignoring spatial effects might introduce biases, especially in the case of non-linear models.

Empirically, estimation and application of models developed at dwelling-unit level is limited because of a lack of independent dwelling supply data for many metropolitan regions. As a result, some studies have used a sample of dwellings from household surveys in order to generate the universal choice set in the absence of disaggregated dwelling supply data. This approach is also prone to biases, as we shall see in chapter 4.

2.6 The Problem of Massive Universal Choice Set and the Choice Set Formation Problem

Depending on the granularity of the location choices, the universal set of alternatives can range from hundreds in zonal-based models to hundreds of thousands or even millions in parcel-based and dwelling-level location choice models. As discussed in section 2.4.4, random sampling of alternatives is typically applied in order to cope with the estimation costs in models with a massive universal choice set. The sampling of alternatives approach provides a powerful tool for estimation of spatial models;
however, there are other issues in the modelling of choices among large numbers of alternatives which are discussed in the following.

Forecasting of spatial choices is one of the most important issues as forecasting typically involves multiple runs of a Monte Carlo sampling process based on the choice probabilities that are calculated on the full set of alternatives. This can increase the simulation run time to the extent that application of complex choice models (even when it is possible to estimate them) becomes infeasible in large scale models. Kikuchi et al. (2003) proposed a Markov Chain Monte Carlo (MCMC) algorithm to reduce the computation time required to forecast spatial choices with a massive universal choice set for MNL models. Further research is required to investigate the feasibility of this approach for non-MNL models.

Instability of prediction results in models with a massive universal choice set (i.e., stochastic variation) is another important issue. Wegener (2011a) argued that the predictive power of models with a large number of alternatives is reduced to due to the flattening of choice probability distributions. Using a Monte Carlo simulation, he showed that the stochastic noise quickly becomes larger if the number of agents becomes smaller or the number of alternatives becomes larger. Waddell (2009) suggested that models which incorporate a choice set formation process may reduce the severity of this probability flattening problem.

Another important issue in models with a massive universal choice set concerns the fact that decision makers do not choose from the universal choice set when they are facing a large number of alternatives. This issue is known in the literature as the choice set formation problem. The inappropriate use of the universal choice set in modelling can lead to significant misspecification errors where the actual choice sets considered by decision makers are different from the universal choice set (see Stopher, 1980).

Experimental research has suggested that in a complex choice situation (i.e., choice from a large number of alternatives), decision makers adopt non-compensatory screening strategies (e.g., elimination-by-aspects, see Tversky, 1972) to reduce the number of alternatives they consider to a smaller number before using a compensatory decision rule to make a final decision (see Manrai and Andrews, 1998). This has led to the view
that decisions may be made in two stages: (a) non-compensatory stage (b) compensatory stage.

This can be explained from both a psychological and from an economic perspective. From the psychological perspective, individuals’ cognitive capacities are limited (i.e., bounded rationality) and they employ a simplifying heuristic to mitigate the cognitive demand arising from choices from a large set of alternatives. From the economic perspective, individuals facing choices from a large set of alternatives search for new information until the benefits of information acquisition exceed the associated search cost (see Roberts and Lattin, 1991).

In summary, the high computation cost involved in models with a massive universal choice set, the stochastic variation and instability of forecasting of such models, and the behavioural aspects of choice set formation lead to an important practical question of how to identify the actual choice sets of decision makers from the set of all available alternatives. Since the observed choice data do not reveal any information about the actual choice sets, researchers have proposed different methodologies to deal with the issue of choice set formation based on decision theory. These different approaches to deal with choice set formation problem are reviewed in the following sub-sections.

### 2.6.1 Deterministic Choice Set Formation

In the deterministic (or exogenous) choice set formation approach, it is assumed that decision makers’ choice sets can be determined deterministically and exogenous to the choice process. The choice set for each decision maker, in the deterministic approach, is typically generated by restricting the universal choice set using deterministic constraints on one or more attributes of alternatives. The deterministic constraint approach attempts to address the issue of choice set formation in a more behavioural manner by limiting the universal choice set to feasible alternatives for each decision maker.

Many applications of this approach exist in the literature for different choice contexts. In the mode choice context, for example, knowledge that an individual has no driving

---

15 There have been some attempts to model choices as an outcome of a search process based on the search theory rather than the decision theory and using the “choice process” data rather than the choice data (see Lerman and Mahmassani, 1985; Caplin and Dean, 2011).
licence may lead the analyst to an a priori removal of the auto-mode from the individual’s choice set (see Ben-Akiva and Lerman, 1974; Train, 1980). In the destination/activity location choice context, a distance threshold, for example, is determined deterministically and used to generate the choice sets based on the hypothesis that individuals do not consider alternatives that would imply travelling more than a specified maximum distance (see Parsons and Hauber, 1998; Termansen et al., 2004; Scott, 2006).

Some authors have also proposed applying the importance sampling approach (without considering the alternative-specific correction terms) to incorporate more behavioural realism in the choice set formation stage for spatial models. For example, in the residential location choice component of ILUTE, households’ choice sets are constructed by taking 75% of a random sample from dwellings that were within 15 km of their previous locations and the remaining 25% from dwellings that were not within the 15 km threshold (Farooq and Miller, 2012). Elgar et al. (2009) also applied the importance sampling approach (without alternative-specific correction terms) in a firm location choice context by oversampling of alternatives around the anchor-points (anchor-points are points of significance to decision makers – such as work locations – which lead decision makers to assign more importance to contiguous locations when they are looking for a residential location).

As discussed in section 2.4.4, the generation of choice sets based on the importance sampling of alternatives when the alternative-specific correction terms are considered is asymptotically equivalent to the universal choice set approach, hence, sampling approaches do not in general capture any non-trivial behavioural mechanisms associated with the choice set formation process. The importance sampling approach, without considering the alternative-specific correction terms, and the deterministic constraint approach, can be considered to be equivalent. In fact, the deterministic constraint approach can be viewed as an importance sampling protocol where alternatives are assigned a weight of 0, if they are outside the thresholds, and a weight of 1, if they are inside the thresholds.

In general, the definition of choice sets in the deterministic constraint approach (or the importance sampling approach without correction) depends on the analysts’
assumptions as to which variable and what thresholds should be applied to define the choice sets. In fact, these assumptions reflect the search behaviour of decision makers. Chapter 5 elaborates more on exogenous choice set formation approaches in the context of residential location choice modelling, namely the deterministic constraint approach and the importance sampling approach.

### 2.6.2 Two-stage Choice Set Formation Models

Some authors have argued that deterministically constructing choice sets based on an analyst’s judgement involves uncertainty and have thus attempted to incorporate the uncertainty involved in the choice set formation process by probabilistic modelling of choice sets and choices based on the Manski (1977) formulation. The general form of Manski’s two-stage model is as follows:

\[
P_{i,n} = \sum_{C_n \in G} P(i|C_n). Q_n(C_n) \tag{2-30}
\]

where,

- \( P_{i,n} \) is unconditional probability of alternative \( i \) being chosen,
- \( P(i|C_n) \) is conditional probability of alternative \( i \) being chosen given choice set \( C_n \),
- \( Q_n(C_n) \) is probability that individual \( n \)’s choice set being \( C_n \), and
- \( G \) is a set of all non-empty subsets of the universal choice set \( M \).

In a marketing research and brand choice context, different choice set generation strategies are studied. Manrai and Andrews (1998) reviewed different specifications of choice generation in marketing literature and classified various choice set generation models such as: (a) memory-based or stimulus-based, (b) attribute-based or brand-based, etc. In transportation and environmental planning literature on the other hand, the scope of choice set generation strategies is mainly limited to an attribute-based approach where an individual is assumed to generate her choice set by eliminating alternatives based on one or multiple constraints on the attributes of alternatives (i.e., a constraint-based model).

Swait and Ben-Akiva (1987) formulated a two-stage probabilistic model through a constraint-based non-compensatory process for the choice set generation process and
the utility maximisation compensatory process for the actual choice (i.e., a semi-compensatory choice set formation model). They assumed that individual \( n \) is considered to reject alternative \( i \) if the value of any constraint variable \( r \), denoted by \( x_{n,i}^r \), exceeds a threshold \( T_{n,i}^r \). Since choice sets in general are a latent construct to the analyst, and there is usually nothing observed about them except the most preferred alternative for each individual, the thresholds are assumed to be random variables (i.e., a random constraint model). Therefore, the probability of an alternative \( i \) being included in individual \( n \)'s choice set and the probability that individual \( n \)'s choice set being \( C_n \) (assuming thresholds are independent) are given as:

\[
q_n(i \in C_n) = \prod_r Pr(x_{n,i}^r \leq T_{n,i}^r)
\]

\[
Q_n(C_n) = K \cdot \prod_{i \in C_n} q_n(i \in C_n) \cdot \prod_{i \notin C_n} (1 - q_n(i \in C_n))
\]

where, \( K \) is normalisation constant in order to allow the choice set probabilities over \( G \) to sum to one. The normalisation constant \( K \) can be computed as:

\[
K = \frac{1}{1 - Q_n(\emptyset)}
\]

where, \( Q_n(\emptyset) \) denotes the probability that the choice set is empty.

In order to incorporate the heterogeneity of individuals in the choice set formation stage, one can parameterise the means of distributions of thresholds as a function of individual characteristics. Thresholds are assumed to follow different distributions in the literature, with normal distributions being assumed in Swait and Ben-Akiva (1987), Andrews and Srinivasan (1995), Cantillo and Ortúzar (2005) and logistic distributions in Ben-Akiva and Boccara (1995) and Basar and Bhat (2004).

In random constraint models, the variables determining the choice set probability and the choice probability may overlap partially, completely, or not at all. Haab and Hicks (1997) partitioned the full set of observable determinants into those influencing the choice set probability and those that influence the choice probability. The simulation study by Andrews and Manrai (1998b) found no adverse consequences from including the same variables as determinants of both the choice set probability and the choice probability.
The major drawback of random constraint models is that they are computationally intensive since the number of theoretically possible choice sets increases exponentially with the number of alternatives \( |G| = 2^{\mid M \mid} - 1 \). This approach is therefore impractical for choice situations with a large number of alternatives, such as choices of residential locations. Some authors have tried to overcome the computation burden of the Manski model while still accommodating the probabilistic nature of choice sets by restricting the composition of set \( G \) based on some simplifying assumptions regarding the behavioural realism of the choice set formation process (see Swait and Ben-Akiva, 1987; Siddarth et al., 1995; Andrews and Manrai, 1998a; Zheng and Guo, 2008; Kaplan et al., 2009; Hicks and Schnier, 2010). These models are discussed further in chapter 6.

2.6.3 Single-stage Choice Set Formation Models

Following the competing destination model of Fotheringham (1983), some authors have viewed choice sets as fuzzy sets so as to be able to incorporate the uncertainties of choice sets in discrete choice models (see Cascetta and Papola, 2001; Cascetta et al., 2007). Wu and Rangaswamy (2003) provided a detailed discussion on this approach.

In a residential location choice context, Martinez et al. (2009) developed the Constrained Multinomial Logit (CMNL) model where each alternative has a degree of membership to the choice set which is determined based on some constraints on the attributes of alternatives. This approach is also referred to as the soft constraint approach, in which the constraints can be violated to some extent (i.e., as opposed to the crisp constraint approach of the Manski model). Bierlaire et al. (2010) compared the single-stage framework with the two-stage modelling framework using synthetic data and concluded that the CMNL model developed by Martinez et al. (2009) is unable to reproduce the results of the Manski model and should be considered as a model in its own right.

2.6.4 Taste-driven Choice Set Formation Models

Horowitz and Louviere (1995) hypothesised that the choice set may simply be a reflection of preference rather than a separate construct. They argued that choice sets provide no information beyond that contained in the utility function. The utility function, however, is never known with certainty to the analyst and must be estimated from the data. Since any information about preferences may be useful for improving
estimation efficiency, Horowitz and Louviere (1995) argue that the specifications of choice sets can be used to improve estimation efficiency. Hence, they employed a range of survey questions to elicit individuals’ choice sets and used that to improve the specification of the utility function (see Horowitz and Louviere, 1995; Parsons et al., 1999).

Swait (2001) proposed a new model of choice set generation belonging to the GEV (Generalised Extreme Value) family of discrete choice models (i.e., the GenL model). The choice sets in the GenL model are taste-driven, similar to the proposal of Horowitz and Louviere (1995). An interesting feature of the GenL model is that the model neither uses exogenous information nor incorporates latent variables describing the choice sets.

2.6.5 Concluding Remarks

This section has shown that there are other problems associated with the massive size of the universal choice set in residential location choice models than the computational cost of estimation inherent in such models, including, the computational cost of forecasting, problems associated with the flattening of choice probabilities, and the choice set formation problem. While the random sampling of alternatives approach provides a statistical solution for estimation in models with a large number of alternatives, it disregards any behavioural considerations regarding the search process and the cognitive capacities of households.

There are also different approaches proposed in the literature to deal with the choice set formation problem. To summarise, deterministic choice set formation approaches assume the choice sets can be constructed based on some behavioural rules. This approach is easy to implement in choice situations with a large number of alternatives. The performance of the deterministic choice set formation approach within the residential location choice context has not been evaluated in the literature, however. Furthermore, the deterministic choice set formation approach is also unable to incorporate the uncertainty inherent in the choice set formation process.

There have been some attempts in the literature to incorporate the uncertainty of the choice set formation process based on the Manski formulation (i.e., two-stage models) and the fuzzy-based approach (i.e., single-stage models). Unfortunately, the implementation of a theoretically sound, two-stage discrete choice modelling paradigm
incorporating probabilistic choice sets is impractical when the number of alternatives is large. The performance of single-stage choice set formation approaches should also be evaluated in forecasting scenarios and in a residential location choice context before applying them in practice. Chapters 5 and 6 examine the choice set formation problem in a residential location choice context in more detail.

2.7 Conclusion

This chapter contributes toward the identification of the challenges associated with the modelling of residential locations as applied in urban simulation models. Recent advances in computing power and econometric methods provide opportunities to address these challenges and to develop an improved model of residential location choices as discussed in the following chapters.

The chapter began by presenting an overview of operational land use models, emphasising the residential location component of these models. It also discussed different theories and modelling frameworks applied in residential location modelling.

The discrete choice approach is indisputably the most widely-applied approach in modelling residential location and has been comprehensively examined in the chapter.

The most conventional approach in the discrete choice analysis of residential location choices is the zonal-based approach in which alternatives are assumed to be aggregated spatial units. Zonal-based models, however, suffer from aggregation bias and do not incorporate the preference of households regarding the attributes of dwellings. This may significantly reduce the predictive ability of zonal-based models. There have, therefore, been some attempts in the literature to model residential location choices at the level of dwelling units, but there are still many methodological and empirical challenges involved in modelling at this level. Chapter 4 elaborates on the challenges arising in dwelling-level residential location choice models.

Additional, and still unsolved, problems in the modelling of residential location choices are the massive universal choice set and choice set formation. This chapter has reviewed alternative choice set formation approaches in order to provide the background context and identify potential challenges and critical gaps that are further examined in chapters 5 and 6 of this thesis.
CHAPTER 3

THE HOUSING MARKET AND RESIDENTIAL LOCATION CHOICES

3.1 Overview

Chapter 2 briefly presented the evolution of widely used urban simulation models and elaborated the various approaches for modelling residential locations within urban simulation models, focusing especially on discrete choice residential location models. This chapter considers residential location choices from a wider perspective as one of the sub-processes of the housing market.

As discussed in chapter 2, state-of-the-art urban simulation models attempt to include all the processes involved in the housing market (and other relevant markets), not only the residential location choices. This chapter starts by briefly reviewing previous attempts to model the different sub-processes of the housing market (in addition to the residential location choices discussed in chapter 2) such as residential mobility decisions, tenure decisions, etc. It then proposes a comprehensive conceptual framework for the housing market in order to describe how the market works in reality and to provide some insights into the interaction of the different processes and agents involved in the supply and demand sides of the market. Both consumption and investment aspects of the housing market are considered in the proposed conceptual framework. The role of market facilitators such as real estate agents and mortgage lenders are also considered. The proposed conceptual framework provides a basis for improving the representation of the housing market within residential location choice models, which will be discussed further in chapters 4, 5 and 6 of this thesis.

This chapter is organised as follows: a brief review of attempts to model different processes in the housing market is presented in section 3.2. This section should be considered as a complement to chapter 2. Section 3.3 presents a comprehensive conceptual modelling framework for the housing market, describing various agents and
processes involved in both the demand and supply sides of the market. Building on this conceptual framework, section 3.4 describes our programme of research to improve the representation of the housing market within residential location choice models. This section also clarifies the structure of the remaining chapters of the thesis. Finally, section 3.5 concludes the chapter.

3.2 Market-based Urban Simulation Models

As we have seen in chapter 2, state-of-the-art urban simulation models follow a market-based approach in attempting to simulate all the processes and agents involved in various markets, such as the housing market, labour market, etc. For practical reasons, however, state-of-the-art urban simulation models do not fully include all the processes involved in the housing market (and other relevant markets) but typically follow a simplified approach for these processes. For example, UrbanSim simulates residential mobility decisions by random selection of households based on the annual mobility rate (Waddell et al., 2003).

The general trend in the evolution of urban simulation models indicates that future models are likely to represent the housing market (and other markets) more completely by taking advantage of recent empirical developments relating to the different processes in these markets, as we shall see later in this section. ILUTE, for example, simulates residential mobility decisions based on a behavioural residential mobility model (Salvini and Miller, 2005). We have examined empirical modelling approaches for households’ location choice decisions comprehensively in chapter 2. Here, we briefly review empirical developments relating to other dimensions of the housing market.

Habib and Miller (2007) developed an empirical model of residential mobility decisions based on the residential stressor concept using the hazard model (i.e., a continuous time formulation) and the discrete choice model (i.e., a discrete time formulation). Potepan (1989) examined the decision between making home improvements and moving by estimating a binominal Logit model based on a 1979 panel study of income dynamics. Montgomery (1992) constructed a model where households choose the optimal level of housing and the means to achieve that level given their current housing such that they maximise utility. She estimated an ordered probit model based on a 1985 American housing survey.
Some authors have also investigated the housing tenure choice problem and attempted to incorporate the dual role of housing as a consumption and investment good in a behavioural model of tenure choice to determine the influencing factors in renting or buying decisions (see Henderson and Ioannides, 1983; Goodman, 1988; Haurin et al., 1996). The households’ mobility decisions and tenure decisions after moving have been also jointly modelled based on a dynamic discrete choice framework (Ioannides and Kan, 1996; Kan, 2000).

Unlike the demand-side of the market, agents in the supply side of the market are very diverse (from developers in the private sector to non-profit organisations and governments) and acquiring data about these different agents and the choices they make is cumbersome since most of this data is proprietary. This diversity of agents and the complicated decision making process, as well as a lack of data relating to the behaviour of housing suppliers, present significant barriers to the development of an empirical model for housing supply (Quigley, 1979; Dipasquale, 1999).

Recently, some authors have attempted to model the behaviour of developers on the location and type of new developments. Haider and Miller (2004) considered the behaviour of developers and proposed a location choice model of different development types using disaggregated data from newly built residential projects. Farooq et al. (2011) also developed a supply model from the developers’ perspective: they applied a discrete-continuous framework to model a collection of choices (i.e., location, development type and quantity) made by developers on individual sites to achieve maximum profits. Zollig and Axhausen (2012) also examined the behaviour of real estate developers and identified the heterogeneity in their behaviour in terms of decision criteria, information considered and executed tasks.

The decisions of dwelling owners to rent their dwellings or to keep them vacant is examined by Anas (1982). He proposed a binary choice model based on a profit maximisation framework arguing that owners decide to rent their properties if the profit that can be obtained by renting the dwelling is positive, and they decide to keep it vacant if the loss from an occupied dwelling is larger than the cost of maintaining a vacant unit.
The next section proposes a comprehensive conceptual framework for the housing market which helps us to understand better the relationship between the different studies mentioned in this section. The proposed conceptual framework also sheds light on potential improvements to existing empirical models as well as the development of new empirical models able to capture wider aspects of the housing market.

3.3 A Comprehensive Conceptual Framework for the Housing Market

The housing market is characterised by the complex integration of various agents that are involved in different decision making processes. In order to deal with this complexity, and in order to shed light on the interactions of agents and processes that are involved in the housing market, many authors have attempted to conceptualise the decision making processes and the behaviour of agents (see Brown and Moore, 1970; Cadwallader, 1992; Wong, 2003; Habib and Miller, 2007).

Conceptual frameworks of housing markets are important for piece-wise development of empirical models of different processes in the market as well as the development of urban simulation model systems since they identify how different agents and processes interact and break down different processes into a succession of steps, each representing the choices that agents face (Coulombel, 2010). None of these previous attempts achieves a comprehensive solution, however, since they do not consider the processes involved in the supply-side of the market and the interactions of these with the demand-side processes. The investment perspective of the housing market has also been ignored in previous conceptual models. This section, therefore, first characterises the housing market discussing the unique features which differentiate it from other markets, and goes on to propose a comprehensive conceptual framework for the housing market incorporating the processes involved in both the demand-side and the supply-side of the market.

3.3.1 Characterising the Housing Market

The housing market uniquely combines features of consumption and investment. Looking from the consumption perspective, households in the housing market have preferences for dwellings and neighbourhoods and choose their residences based on the attributes of these dwellings and neighbourhoods. On the other hand, from the
investment perspective, households have preferences regarding the performance of dwellings as a financial asset (i.e., price appreciation).

The housing market is also characterised by its high search cost since households conduct extensive search efforts in order to choose a dwelling once they decide to move. Hence, households do not have perfect information about the housing market when they make their decisions. As a result, real estate agents as well as other sources of information such as the internet, media and word of mouth play an important role in the housing market in so far as they facilitate the search process. The products of the housing market also have some unique features such as heterogeneity, locational fixity, durability and indivisibility which, together, differentiate the housing market from other markets.

Another important characteristic of the housing market is that supply cannot adjust quickly in response to market changes because construction projects usually take a long time and this leads to a lag in available supply. This phenomenon is referred to as “supply lag” in the literature (Dipasquale, 1999).

3.3.2 The Proposed Conceptual Framework

Figure 3-1 presents an overview of the proposed conceptual framework for the housing market. The different processes and agents depicted in the conceptual framework are described in more detail in the following.

On the demand-side of the market, consumers are households that make a set of long-term choices, including residential mobility/refurbishment/tenure decisions, and decisions on location and type of dwellings. Households on the demand-side are assumed to be utility maximisers and choose a dwelling unit from their choice sets (i.e., a set of vacant dwelling units in the market that the household is aware of) which has the highest utility. Section 3.3.2.1 describes the behaviour of households and the processes involved on the demand-side of the markets in more detail.

On the supply-side, agents are assumed to be profit maximisers who make their investment decisions considering the market conditions. There are mainly three agents in the supply-side of the market as: (i) developers who supply new dwellings to the market, (ii) professional refurbishers who supply refurbished dwellings to the market,
and (iii) investor owners who rent out the property that they own to someone else. These agents are involved in different processes as described in section 3.3.2.2.

There are also some agents that interact with the agents in both the demand-side and the supply-side of the market. These agents are called market facilitators and are depicted in the middle of Figure 3-1. The major market facilitators include financial institutions (i.e., lenders) who provide financial products (e.g., mortgages, business loans, bridging loans, etc.) and real estate agents who provide information to agents in both the demand- and supply-side of the market. There are other market facilitators such as solicitors, surveyors, etc. who facilitate the purchase and sale of houses but these are not included in the conceptual model since they do not have a significant impact on the how the market operates.

Governments also intervene in the housing market by imposing taxes and providing subsidies that change the behaviour of households (e.g., mortgage interest relief), owners (e.g., taxes on vacant dwellings in order encourage owners to rent their vacant properties), and agents involved in the supply-side of the market (e.g., tax policies designed to encourage the private construction of housing for rental). Governments may also change the behaviour of households and supply-side agents by investing in public facilities such as transportation infrastructures. Such investments can change housing prices and the attractiveness of a location within a region because of the location’s proximity to the new transit (Wardrip, 2011). Planning permissions, the release of surplus public land for new housing development and direct involvement in the construction of new housing are other forms of leverage which governments may employ to intervene in the housing market.

---

16 The fact that a transit line can increase property values so significantly provides a good opportunity to help finance a transit line project using novel taxation on properties that benefit from the proximity to the new transit.
Figure 3-1 - A Comprehensive Conceptual Framework for the Housing Market
3.3.2.1 Demand-side Processes and Agents

On the demand-side of the market, households adjust their housing consumption by moving or by refurbishing their current dwellings. Households also choose to rent or buy (i.e., tenure choice) in the housing market based on their financial status, a choice which is usually constrained by mobility decisions since in most cases a move has to be made in order to change the housing tenure. The decision to move, the decision to refurbish existing dwellings and tenure decisions are therefore interdependent, as has been discussed by some authors (see Potepan, 1989; Montgomery, 1992; Ioannides and Kan, 1996; Kan, 2000).

Residential mobility decisions can be conceptualised based on the concept of residential stressors first introduced by Rossi (1955). Households are assumed to make mobility decisions when the inadequacy between the current residence and the optimal one exceeds a stress threshold. Residential stresses can arise from changes in household composition (e.g., marriage or divorce), changes in different life cycle events (e.g., job change or income rise), and changes in the surrounding environment (e.g., changes in accessibility). Passive information, such as news about the market condition, can also change the level of stress of households.

The behaviours of owner-occupiers and renters are quite different with respect to mobility decisions. Renters are prone to move more often than owner-occupiers because home owners are subject to higher mobility costs (e.g., transfer taxes, legal fees, potential higher interest rates for mortgages) and constraints (e.g., selling a property can take a long time, depending on market conditions). On the other hand, unlike renters, owner-occupiers have the right to modify their dwellings and lands as they please subject to government regulations (i.e., planning permission). Owner-occupier households, therefore, may decide to refurbish their current dwellings rather than move to a new dwelling in order to satisfy their needs.

Households' decisions to refurbish their dwellings can be conceptualised based on the idea of maximising the value of net benefits from refurbishing, as discussed by Bogdon (1992). The benefits include housing consumption, which is the utility acquired from the refurbished dwelling unit (e.g., more space and less operational and maintenance
costs) and the return on the housing investment in the form of capital gains (Bogdon, 1992).

As mentioned earlier, households face tenure decisions in their housing career which involve both consumption and investment aspects of the housing market. Households choose their tenure type based on their financial status, their ability to obtain credit and the condition of the housing market. Home ownership has been an attractive type of tenure for households and has been promoted by governments as it has a wealth creating potential for households and it encourages households to invest more in their neighbourhoods (Saunders, 1990). Many households, however, cannot afford to buy a house unless they use private sector equity sharing (i.e., mortgages) and/or public sector equity sharing (i.e., shared ownership). Consequently, government housing policies and the behaviour of mortgage lenders play an important role in households’ tenure choices. On the other hand, mortgage market conditions are also affected by housing market conditions and by macroeconomic factors, which makes the whole process more complex.

In summary, renter households have the choice of moving or staying in their current dwellings subject to their contracts, and also have the choice of changing their tenure type if they decide to move. Owner-occupier households have the choice of moving, staying and refurbishing, and staying and not changing. They also have the choice of changing their tenure type if they decide to move. Different choice structures can be assumed to represent the interdependencies between mobility/refurbishment/tenure decisions. Further empirical work is required to identify the suitable choice structure for these interrelated decisions. Figure 3-2 depicts two potential choice structures assuming there are only two tenure types.17

![Figure 3-2 – Different Choice Structures for Mobility/Refurbishment/Tenure Decisions](image)

17 Different tenure types might be available in different countries. Social housing and shared ownership are other important types of tenure in the UK.
Households become active in the housing market when they decide to move and elect their tenure type. They enter into a search process with crude knowledge of the market acquired through daily activities like reading the newspaper, word of mouth, seeing rental advertisements and driving through a neighbourhood. During this search process, households collect information about neighbourhoods and available properties from different sources such as real estate agencies, the internet, rental advertisements and by physically visiting locations and properties. They consider this information as well as their own needs and constraints (e.g., required space, acceptable work commute time, acceptable price) in order to narrow down the alternatives and form their choice sets. Finally, households make their dwelling choices from their choice sets based on the attributes of dwellings such as dwelling size, type, etc., the attributes of neighbourhoods and locations in which dwellings are situated, such as travel time and cost, school quality, crime rate, etc., and the expected performance of the dwellings as a financial asset.

It should be noted that in the proposed conceptual model, it is assumed that households first form their choice sets by searching the market and then they make their dwelling choices. Some authors have criticised this approach and have argued that the residential location choice process is a search process wherein the process is terminated and the choice set is known only when the final choice is made (see Hall, 1980; Richardson, 1982). Lee and Waddell (2010) have also criticised the separation of mobility and location choice decisions and have argued that households' mobility decisions may partly be based on the price and quality of the available alternatives.

In the conceptual framework proposed here, however, we have separated the mobility/tenure/refurbishment choice process, the searching process, and the dwelling choice process for practical reasons. Feedback to previous stages is provided in the conceptual framework in order to illustrate the interdependency of these different stages.

3.3.2.2 Supply-side Processes and Agents

The available housing supply at each instance of time partly derives from the existing dwellings of potential movers (both owners and renters), and partly from new real estate developments. Some part of the housing supply also derives from redevelopment
of old dwellings. Housing supply, therefore, depends to some degree on the availability of land and is affected by the land market. Although private land ownership is permitted in most countries, the ultimate owner of land is the state which controls the land market by imposing regulations to control the usage and supply of land (Mahoney et al., 2007).

The supply of new dwellings is provided by the private sector with the incentive of making profit and by the public sector with the aim of achieving maximum public welfare, as discussed by Wilson (1970). The private profit-maximising system assumes that there are sufficient profit-maximising developers and that there is perfect competition among these developers. Developers assess their profits for different types of development in different locations in which they own or can buy land. Since the expectation of profits and the risk-taking attitude of developers are different, they select different locations and development types in order to maximise their profits. Developers’ decisions are constrained by their capital and their ability to obtain credit. In the public welfare-maximising system, government is directly involved in developing new housing supply with the aim of achieving maximum public welfare. It should be noted that the private profit-maximising system results in a different housing stock from the public one, since the most potentially profitable parts of the market could be overdeveloped.

Professional refurbishers in the supply-side of the market are involved in refurbishing the dwelling stock with the intention of making profit. Investor owners, who generate income from renting out the properties that they own, are also involved in the redevelopment market since they may decide to improve their properties in order to generate more profit from renting. Decisions as to different types of refurbishment activities (e.g., repairing, extending, etc.) are determined by the consumption requirements of households and their budget constraints and, in this respect, owner occupiers also play an important role in the evolution of housing stock in the supply-side of the market. Obviously, for all agents involved in the redevelopment market, including owner-occupiers (i.e., on the demand-side), investor owners and professional refurbishers (i.e., on the supply-side), capital gain is one the most important factors in refurbishment decisions (i.e., whether to refurbish or not and the type of refurbishment).
Another important decision that owners face is the decision as to whether to offer the dwelling unit to the market for sale/rent or to keep the unit vacant. This depends on market conditions, maintenance costs and the profit they may generate by renting or selling the property.

In summary, housing supply is determined not only by the production decision of developers of new housing stock but also by the decisions made by owners concerning conversion and redevelopment of the existing stock (see Figure 3-3).

**Figure 3-3- Dwelling Supply in Rental and Owner’s Market**

### 3.4 Research Programme

The conceptual framework presented in the previous section has attempted to provide a complete picture of the housing market and the different agents and processes involved in it. The conceptual framework described in the previous section sheds light on the unique characteristics of the housing market related to residential location choices, which can be summarised as comprising heterogeneous and immobile products, having high search costs, and combining features of both investment and consumption.

While this conceptual framework has attempted to provide a complete picture of the housing market and the different agents and processes involved in it, the objective of
this thesis is more narrowly focused on improving the empirical modelling of the residential location choice process, leaving other processes involved in the housing market for future work. To achieve this objective, this thesis attempts to incorporate the unique features of the housing market as described earlier into discrete choice modelling of residential location choices. To this end, the thesis focuses on the consumption aspects of the housing market and does not consider investment aspects in modelling residential location choices.\textsuperscript{18} The research programme has therefore been divided into two parts, as depicted in Figure 3-4.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{research-programme.png}
\caption{Research Programme and Thesis Chapters}
\end{figure}

Part I (chapter 4) acknowledges the heterogeneity and immobility of housing market products and proposes to model residential location choices at the level of dwelling

\textsuperscript{18} There is an interaction between the housing market and the transportation market which is well known in urban economics literature, as discussed in chapter 2. There is also a profound association between the housing market and the labour market since households’ choice of employment, employment locations, and other attributes of the labour market influence their choice of residence (see Roback, 1982). This thesis, however, does not consider the interaction of the housing market with other markets and assumes that choices related to other markets such as mode of travel and employment locations are exogenous to residential locations.
units. Various model structures to capture the spatial structure effects and the problems associated with construction of the universal choice set and estimation of such models are also explored in chapter 4. For reasons of simplicity, this chapter does not consider the high search costs of the housing market and assumes that households have perfect information on all alternatives in the market.

Part II (chapters 5 and 6) acknowledges the high search costs of the housing market and attempts to incorporate the underlying search process (as discussed in the conceptual framework) into the discrete choice modelling of residential location choices based on choice set formation strategies. This part of the thesis investigates both deterministic choice set formation approaches (i.e., chapter 5), and probabilistic choice set formation models (i.e., chapter 6) within the context of residential location choice modelling.

3.5 Conclusion

This chapter has characterised the complexity of the housing market as a result of the integration of various agents and processes that are involved in both the demand-side and supply-side of the market, and has proposed a comprehensive conceptual framework in order to shed light on how the housing market works in reality. This proposed conceptual model is unique in that it is the first of its kind that attempts to present a complete picture of the decision processes involved in the housing market.

The proposed conceptual framework is used in this study to identify the challenges and to propose improvements in empirical modelling of residential locations. These challenges and improvements are further explored in the following chapters. Undoubtedly, however, the proposed conceptual framework gives rise to many research questions in respect to the development of empirical models of the processes involved in the housing market and further empirical models could be developed in order to incorporate wider aspects of the housing market in applied urban simulation modelling. This will shape the future work linked to this study.
CHAPTER 4

MODELLING RESIDENTIAL LOCATION CHOICES AT THE LEVEL OF DWELLING UNITS

4.1 Overview

The gradual emergence of high-resolution spatial data has motivated researchers to extend zonal-based residential location choice models to parcel-based models. Parcel-based models are useful in analysing phenomena and policies that require a higher spatial resolution, such as environmental impacts, walking scale accessibility, etc. However, the difficulties of working with high-resolution spatial data in terms of computation time, as well as the lack of availability of such data for many metropolitan regions, inhibit wide use of these microscopic models (Waddell, 2009). Moreover, similar to zonal-based models, models developed at the level of parcels fail to recognise that households actually choose individual dwelling units rather than aggregated spatial units. This problem is often referred to in the literature as aggregation bias (or ecological fallacy).

As discussed in chapter 3, the housing market is characterised by the fact that it offers heterogeneous products. In fact, each dwelling unit is unique to a certain extent. By aggregating elementary alternatives (i.e., dwelling units) into spatial units (i.e., zones and parcels), the variability of dwelling attributes in spatial units is lost. Besides, empirical studies confirm the importance of dwelling attributes in households' residential preferences; therefore, ignoring the variability of dwelling attributes in modelling residential location choice may limit the predictive ability of the model. As a result, some authors have attempted to model residential location choices at the level of dwelling units, assuming that dwelling units are multivariate heterogeneous products (see Bayer et al., 2005; Zhou and Kockelman, 2008; Habib and Miller, 2009; Guevara, 2010).
As discussed in chapter 2, dwelling-level models can be categorised into two general categories based on how they represent the spatial locations of dwelling units, as: (a) parcel-based dwelling-level models, and (b) zonal-based dwelling-level models. Similar to parcel-based models, parcel-based dwelling-level models also require parcel data which are not available for many metropolitan regions.

Parcel data for urban modelling purposes is usually assembled from different data sources. In the United States, for example, it is derived from assessor tables from County Tax Assessor offices and from building tables (Waddell et al., 2004). Parcel data provides information on: (i) averages of attributes of dwellings that reside in a parcel such as average prices, average dwelling sizes, etc. (dwelling-specific variables), (ii) the attributes of the parcel itself, such as lot size, number of dwelling units, etc. (parcel-specific variables), (iii) the locations of dwellings at the parcel-level, which can be also used to measure quantities that require high spatial resolutions such as waking scale accessibility (parcel-level locational variables).

In order to exploit the advantages of parcel data, parcel-based dwelling-level models must be able to accommodate parcel-specific variables and parcel-level locational variables. It is self-evident that the locations of dwelling-units cannot be identified at the parcel-level if we do not incorporate parcel-specific and parcel-level locational variables in the model. In such cases, the locations of dwellings with similar attributes are differentiated based on their zonal-level attributes and the assignment of parcels to dwellings with similar attributes within the same zone will be random. Figure 4-1 presents an example using three dwellings with similar dwelling-level attributes situated in different parcels within a zone. In this case, if the parcel-level attributes are not included in the utility function, the choice model cannot differentiate the locations of dwellings at the level of parcels. Hence, any parcel within the zone can be assigned to the dwellings as there is no systematic difference between the utilities of these three dwellings. This is an important issue in parcel-based dwelling-level models which is not considered in some of the developed models in the literature. For instance, the model developed by Habib and Miller (2009) does not include any parcel-specific or parcel-level locational variables; hence, in this model, parcels of similar dwellings within a zone cannot be differentiated and the advantages of parcel data are not exploited.
The association of dwelling units to space in both zonal-based dwelling-level models and parcel-based dwelling-level models leads to another important issue in modelling since the IIA (Independent from Irrelevant Alternatives) assumption (in a conventional Multinomial Logit (MNL) structure) is no longer valid. If two dwellings are close to each other, they might have some common attributes that are not observed by the analyst, such as the air quality of the area or being close to a park. Therefore, a more flexible model structure is required in order to capture these spatial effects in dwelling-level residential location choice models. Chapter 2 reviewed different approaches to incorporating the spatial effects in zonal-based models; however, incorporating spatial effects in dwelling-level models has not been explored in the literature. This chapter

---

The parcel map have been extracted from OS MasterMap Topography Layer (available at: [www.ordnancesurvey.co.uk](http://www.ordnancesurvey.co.uk)) for one of the TAZ of London.
proposes a zonal-based dwelling-level residential location choice model capturing spatial effects based on both the GEV family and the Mixed Logit formulations.

An additional problem in developing dwelling-level models is the lack of independent disaggregated dwelling supply data for many metropolitan regions. Due to the lack of dwelling supply data, some studies have constructed the choice set by taking a random sample from the set of all chosen alternatives in the household survey (see Zhou and Kockelman, 2008). This approach is prone to biases, however. This chapter illustrates and discusses these biases and proposes a dwelling synthesising approach to correct the choice set construction bias in dwelling-level residential location choice models taking advantage of aggregated dwelling supply data.

The rest of the chapter is organised as follows: section 4.2 presents a zonal-based dwelling-level model and suggests a GEV and a Mixed Logit formulation to capture spatial structure effects. Section 4.3 presents the dwelling synthesising approach for choice set construction. The proposed approach is validated using a Monte Carlo experiment as discussed in section 4.4. The empirical application of the zonal-based dwelling-level model based on the proposed choice set construction approach is presented in section 4.5. Finally, section 4.6 concludes the chapter with a summary of contributions and remaining challenges.

4.2 Zonal-based Dwelling-level Residential Location Choice Models

Incorporating Spatial Effects

This section presents a zonal-based dwelling-level residential location choice model incorporating spatial structure effects such as spatial correlation and spatial heteroscedasticity. For clarity of presentation, the model based on the IIA assumption and the MNL formulation is presented first and then extended to propose a novel GEV model and a Mixed Logit model to capture spatial effects in zonal-based dwelling-level models. The challenges involved in the estimation of such complex model structures, and the opportunities that have become available with the recent advances in estimation techniques, are also discussed.
4.2.1 The MNL Formulation

The utility that a household associates with a dwelling unit depends on both the dwelling unit attributes and the attributes of the zone in which the dwelling resides. Therefore, for a given region with \( Z \) zones \( z \): 1, ..., \( Z \), and \( D \) dwellings \( d \): 1, ..., \( D \), the utility that household \( n \) associates with dwelling \( d \) in zone \( z \) can be written as:

\[
U_{nd} = V_{nd} + \nu_{nd}
\]  

(4-1)

where,

\( V_{nd} \) is the systematic utility that household \( n \) associates with dwelling \( d \) in zone \( z \).

\( \nu_{nd} \) is an unobserved standard extreme value random term that represents idiosyncratic individual differences which are independently and identically distributed across dwellings and households for an MNL formulation.

In zonal-based dwelling-level models, \( V_{nd} \) can be separated into two parts as:

\[
V_{nd} = V'_{nd} + V'_{nz(d)}
\]  

(4-2)

\( V'_{nd} \) is the part of the systematic utility the household associates with the dwelling.

\( V'_{nz(d)} \) is the part of the systematic utility that the household associates with the zone within which the dwelling is located.

Let \( X_d \) represent a vector of observable attributes of dwelling \( d \) other than price (i.e., dwelling size, type, age, etc.) and let \( P_d \) represent the price of dwelling \( d \). The observable attributes of zone \( z \) (i.e., accessibility, school quality, crime rate, etc.) is denoted by \( X_z \).

The travel distance or travel time measure between zone \( z \) and the zone of the primary workplace of household \( n \) is denoted by \( C_{nz} \).

Finally, let \( H_n \) represent a vector of the socioeconomic characteristics of household \( n \) (i.e., income, race, composition, etc.).

Then, the part of the systematic utility the household associate with the dwelling can be written as:

\[
V'_{nd} = X_d(\beta_1 + \beta_2H_n) - P_d(\beta_3 + \beta_4H_n)
\]  

(4-3)
And the part of the systematic utility that the household associates with the zone that comprises the dwelling can be written as:

\[ V'_{nz(d)} = X_k(\beta_5 + \beta_e H_n) - C_{nz}(\beta_7 + \beta_8 H_n) \]  

(4.4)

The characteristics of the household are also included in the utility function by interaction terms.

Assuming that all dwelling attributes that affect households' decisions are included in the utility function, the probability that household \( n \) chooses dwelling \( d \) in zone \( z \) is given by Multinomial Logit (MNL) probability as:

\[ P_{nd} = \frac{\exp(V_{nd})}{\sum_{d'} \exp(V_{nd'})} \]  

(4.5)

where, \( d' \) indices all possible dwelling choices in the region.

Accordingly, the likelihood function can be written as:

\[ L(\beta) = \prod_{n=1}^{N} \prod_{d} P_{nd}^{Y_{nd}} \]  

(4.6)

where, \( Y_{nd} \) takes the value 1 if household \( n \) chose dwelling \( d \), and 0 otherwise.

The next sections present more flexible model structures that allow the capture of spatial correlation and spatial heteroscedasticity in zonal-based dwelling-level models. It should be borne in mind that ignoring spatial correlation and spatial heteroscedasticity will result in biased and inconsistent parameter estimation and structural instability in non-linear models such as the proposed residential location choice model (see Anselin, 1988; LeSage and Pace, 2009).

4.2.2 The GEV Formulation

GEV family models have previously been used to capture spatial correlation in zonal-based models, as has been discussed in chapter 2. This section proposes a novel GEV family model in order to capture spatial correlation in zonal-based dwelling-level models. The proposed GEV model is a spatial case of the Generalised Nested Logit (GNL)

\[ \text{If an influencing dwelling attribute is not observed and is correlated with the dwelling price (assuming that dwelling price captures all dwelling and neighbourhood attributes), the price will be correlated with the error term which causes an endogeneity problem (see Bayer et al., 2005; Guevara and Ben-Akiva, 2006).} \]
model proposed by Wen and Koppelman (2001) and extends the Spatially Correlated Logit (SCL) model (Guo and Bhat, 2004a) for dwelling-level models.

As previously discussed in chapter 3, because of the association of alternatives to space in the housing market (i.e., locational fixity), there is a dependency between different residential alternatives (i.e., dwelling units) since it is expected that the closer alternatives will be more similar than farther alternatives. Since the locations of alternatives in zonal-based dwelling-level models are identified by zones, the distance of two dwellings can be at best determined by the distances between the centroids of the zones that comprise them. In the development of the GEV model, therefore, the following assumptions have been made: (i) dwellings that are in the same zone have some common unobserved attributes because they share some unobserved additional advantages or disadvantages, and (ii) dwellings that are not in the same zone also have some common unobserved attributes and the degree of their similarities decreases with the distances between the centroids of the zones that comprise them because the long-distance dwellings receive smaller shares from the unobserved additional advantages or disadvantages than close-by dwellings.

The representation of spatial correlation among dwelling units based on the zones in which they are located is highly sensitive to the way in which the zones are defined in the study area. This is a long standing and still unsolved issue in spatial analysis which is known in the literature as the Modifiable Areal Unit Problem (MAUP).

Considering the residential choice example of Figure 4-2, we expect the unobserved attributes of dwellings 2, 3 and 4 (and other dwellings that are in the same zone) to be correlated. We also expect that the unobserved attributes of dwellings 2, 3 and 4 (dwellings in zone 2), and dwellings 5 and 6 (dwellings in zone 3) to be correlated (and other dwellings that are in different zones), where the degree of their correlation decreases with the distance between zone 2 and zone 3.
Figure 4-2- An Example of Residential Choice among 8 Dwellings Scattered in 4 Zones

Figure 4-3- Generalised Nested Structure for Capturing Spatial Correlation in Dwelling-Level Models

This complex pattern of correlation can be represented by the generalised nested structure depicted in Figure 4-3. The following generator function is proposed to capture spatial correlation in dwelling-level models:

$$G(e^{V_{n1}}, \ldots, e^{V_{nd}}) = \sum_{i=1}^{Z-1} \sum_{j=i+1}^{Z} \left( \sum_{d \in N_{ij}} (\alpha_{d,ij} e^{V_{nd}})^{1/\mu} \right)^{\mu}$$  \hspace{1cm} (4-7)

where,

$V_{nd}$ is the systematic utility that household $n$ associates with dwelling $d$,

$N_{ij}$ is the set of all dwellings included in nest $ij$, 

77
\( \mu \) is the dissimilarity parameter capturing the correlation between dwellings units \((0 < \mu \leq 1)\).

\( \alpha_{d,ij} \) is the allocation parameter representing the portion of dwelling \( d \) assigned to the nest \( ij \).

The above GEV generator function meets McFadden’s (1978) conditions, hence it is consistent with utility maximisation. It should be noted that there are \((Z - 1) \times D\) allocation parameters in the proposed model (see Table 4-1) which means that even with the 4 zones and 8 dwellings in this simple example 24 allocation parameters should be estimated. Obviously, once there are a large number of dwellings and zones in an area the number of allocation parameters expands significantly. In addition, the identification problem might come into effect, since the maximum number of logsum parameters that can be identified is equal to the number of variance-covariance parameters that can be estimated for the MNP model (i.e., \( \frac{D \times (D - 1)}{2} - 1 \)).

Similar to Sener et al. (2010), therefore, we reparameterised the allocation parameters as a distance decay function given by:

\[
\alpha_{d,ij} = \frac{R_{ij}^\phi}{\sum_{\forall j \in N_{ij}} R_{ij}^\phi} \quad (4-8)
\]

\[
\sum_{\forall j \in N_{ij}} \alpha_{d,ij} = 1, \forall d \quad (4-9)
\]

where,

\( R_{ij} \) is the distance between zone \( i \) and \( j \),

\( \phi \) is the parameter to be estimated.

We would expect a negative sign for the parameter \( \phi \) since the correlation between dwellings reduces as the distance between the zones that comprise them increases. The distance decay function is normalised in such a way that the total allocation of dwelling \( d \) across all nests \( ij \) that comprise dwelling \( d \) is unity. This provides a useful interpretation of the allocation of each alternative to each nest, as discussed by Wen and Koppelman (2001).
Given that the allocation parameters for dwellings that are in the same zone are equal in the proposed model, \((Z - 1) \times Z\) unique allocation parameters (i.e., the allocation matrix) should be calculated for a given \(\phi\) based on the distance decay function (see Table 4-2). This is exactly equivalent to previously developed zonal-based models capturing spatial correlation (see Guo and Bhat, 2004a; Sener et al., 2010).

### Table 4-1- Dissimilarity and Allocation Parameters of the Example Model

<table>
<thead>
<tr>
<th>Nests</th>
<th>Dwellings</th>
<th>Dissimilarity Parameters</th>
<th>Allocation Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>1,2,3,4</td>
<td>(\mu)</td>
<td>(a(1)12, a(2)12, a(3)12, a(4)12)</td>
</tr>
<tr>
<td>13</td>
<td>1,5,6</td>
<td>(\mu)</td>
<td>(a(1)13, a(5)13, a(6)13)</td>
</tr>
<tr>
<td>14</td>
<td>1,7,8</td>
<td>(\mu)</td>
<td>(a(1)14, a(7)14, a(8)14)</td>
</tr>
<tr>
<td>23</td>
<td>2,3,4,5,6</td>
<td>(\mu)</td>
<td>(a(2)23, a(3)23, a(4)23, a(5)23, a(6)23)</td>
</tr>
<tr>
<td>24</td>
<td>2,3,4,7,8</td>
<td>(\mu)</td>
<td>(a(2)24, a(3)24, a(4)24, a(7)24, a(8)24)</td>
</tr>
<tr>
<td>34</td>
<td>5,6,7,8</td>
<td>(\mu)</td>
<td>(a(5)34, a(6)34, a(7)34, a(8)34)</td>
</tr>
</tbody>
</table>

### Table 4-2- Allocation Matrix of the Example Model

<table>
<thead>
<tr>
<th>Nests</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>(a(1)12)</td>
<td>(a(2)12)</td>
<td>(a(3)12)</td>
<td>(a(4)12)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>13</td>
<td>(a(1)13)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>(a(5)13)</td>
<td>(a(6)13)</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>14</td>
<td>(a(1)14)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>(a(7)14)</td>
<td>(a(8)14)</td>
</tr>
<tr>
<td>23</td>
<td>0</td>
<td>(a(2)23)</td>
<td>(a(3)23)</td>
<td>(a(4)23)</td>
<td>(a(5)23)</td>
<td>(a(6)23)</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>24</td>
<td>0</td>
<td>(a(2)24)</td>
<td>(a(3)24)</td>
<td>(a(4)24)</td>
<td>0</td>
<td>0</td>
<td>(a(7)24)</td>
<td>(a(8)24)</td>
</tr>
<tr>
<td>34</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>(a(5)34)</td>
<td>(a(6)34)</td>
<td>(a(7)34)</td>
<td>(a(8)34)</td>
</tr>
<tr>
<td>Sum</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Based on the proposed GEV generator function, the vector of the unobserved part of utility \([\varepsilon_{n1}, \ldots, \varepsilon_{nD}]\) has the following cumulative extreme-value distribution:

\[
F(\varepsilon_{n1}, \ldots, \varepsilon_{nD}) = \exp \left( -\sum_{i=1}^{Z-1} \sum_{j=i+1}^{Z} \left( \sum_{d \in N_{ij}} (\alpha_{d,ij} e^{-\varepsilon_{nd}})^{\frac{1}{\alpha}} \right)^{\mu} \right)
\]  
(4-10)

The marginal cumulative distribution function (CDF) of each stochastic element \(\varepsilon_{nd}\) can be written as:

\[
F(\varepsilon_{nd}) = \exp \left( -\sum_{d \in N_{ij}} \alpha_{d,ij} e^{-\varepsilon_{nd}} \right)
\]  
(4-11)

Since \(\sum_{d \in N_{ij}} \alpha_{d,ij} = 1\), \(\varepsilon_{nd}\) is distributed according to type I extreme value (Gumbel) distribution.
\[ F(\varepsilon_{nd}) = \exp(-e^{-\varepsilon_{nd}}) \] (4-12)

The bivariate marginal CDF for two stochastic elements \( \varepsilon_{nd} \) and \( \varepsilon_{nd} \) of two dwelling units \( d \) and \( \hat{d} \) that are situated in the same zone is given by:\(^{21}\)

\[ H(\varepsilon_{nd}, \varepsilon_{nd}) = \exp \left( -\sum_{\nu \in \mathcal{N}_{ij}} \alpha_{d,ij} \left( e^{-\varepsilon_{nd}} \frac{1}{\mu} + \varepsilon_{nd} \right) \right) \] (4-13)

Given that \( \alpha_{d,ij} = \alpha_{d,ij} \) for dwellings that are in the same zone, we can write:

\[ H(\varepsilon_{nd}, \varepsilon_{nd}) = \exp \left( -\sum_{\nu \in \mathcal{N}_{ij}} \alpha_{d,ij} \left( e^{-\varepsilon_{nd}} \frac{1}{\mu} + e^{-\varepsilon_{nd}} \frac{1}{\mu} \right) \right) \] (4-14)

Since \( \sum_{\nu \in \mathcal{N}_{ij}} \alpha_{d,ij} = 1 \), we can write:

\[ H(\varepsilon_{nd}, \varepsilon_{nd}) = \exp \left( -\left( e^{-\varepsilon_{nd}} \frac{1}{\mu} + e^{-\varepsilon_{nd}} \frac{1}{\mu} \right) \right) \] (4-15)

This is equivalent to bivariate CDF for the NL model where \( d \) and \( \hat{d} \) are in a nest (i.e., the same zone). Therefore, the correlation between dwellings that are situated in the same zone is not a function of allocation parameters and distances. When the dissimilarity parameter \( \mu = 1 \) (i.e., no correlations among dwellings that are in the same zone), the bivariate CDF collapse to the independent bivariate CDF which is the case of the MNL model.

The bivariate marginal CDF for two stochastic elements \( \varepsilon_{nd} \) and \( \varepsilon_{nd} \) of two dwelling units \( d \) and \( \hat{d} \) that are situated in different zones is as follows:

\[ H(\varepsilon_{nd}, \varepsilon_{nd}) = \exp \left( -\sum_{\nu \in \mathcal{N}_{ij}} \alpha_{d,ij} \left( e^{-\varepsilon_{nd}} \frac{1}{\mu} + \varepsilon_{nd} \right) \right) \] (4-16)

Substituting \( \sum_{\nu \in \mathcal{N}_{ij}} \alpha_{d,ij} \) with:

\[ \sum_{\nu \in \mathcal{N}_{ij}} \sum_{\nu \in \mathcal{N}_{ij}} + \sum_{\nu \in \mathcal{N}_{ij} \text{ and } \nu \in \mathcal{N}_{ij}} + \sum_{\nu \in \mathcal{N}_{ij} \text{ and } \nu \in \mathcal{N}_{ij}} \] (4-17)

we can write:

\(^{21}\) The bivariate marginal CDF for any pair of stochastic elements is obtained by setting all other stochastic elements in the multivariate cumulative distribution to infinity.
\begin{equation}
H(\varepsilon_{nd}, \varepsilon_{\hat{n}d}) = \exp\left(-\sum_{v_{ij} | d \in N_{ij} \text{ and } \hat{d} \notin N_{ij}} \alpha_{d,ij} e^{-\varepsilon_{nd}} - \sum_{v_{ij} | \hat{d} \in N_{ij} \text{ and } d \notin N_{ij}} \alpha_{d,ij} e^{-\varepsilon_{\hat{n}d}} - \sum_{v_{ij} | d \text{ and } \hat{d} \in N_{ij}} \left(\alpha_{d,ij} e^{-\varepsilon_{nd}}\right)^{\frac{1}{\mu}} + \left(\alpha_{d,\hat{i}j} e^{-\varepsilon_{\hat{n}d}}\right)^{\frac{1}{\mu}} \right) \right)
\tag{4-18}
\end{equation}

Since \(\sum_{v_{ij} | d \in N_{ij}} \alpha_{d,ij} = 1\), we can write:

\begin{equation}
H(\varepsilon_{nd}, \varepsilon_{\hat{n}d}) = \left(-\left(1 - \alpha_{d,ij}\right) e^{-\varepsilon_{nd}} - \left(1 - \alpha_{d,\hat{i}j}\right) e^{-\varepsilon_{\hat{n}d}} - \left(\alpha_{d,ij} e^{-\varepsilon_{nd}}\right)^{\frac{1}{\mu}} + \left(\alpha_{d,\hat{i}j} e^{-\varepsilon_{\hat{n}d}}\right)^{\frac{1}{\mu}} \right) \right) \tag{4-19}
\end{equation}

where, \(d\) and \(\hat{d}\) \(\in N_{ij}\).

Again the bivariate CDF collapse to the independent bivariate CDF, when the dissimilarity parameter \(\mu = 1\) (i.e., no correlations among the dwellings that are in the different zones).

The choice probability of the proposed model is derived based on the GEV probability expression discussed in chapter 2 as:\textsuperscript{22}

\begin{equation}
P_{nd} = \frac{\sum_{i=1}^{\mathcal{Y}} \sum_{j=i+1}^{\mathcal{Y}} \left(\alpha_{d,ij} e^{\varepsilon_{nd}}\right)^{1/\mu} \left(\sum_{d \in N_{ij}} \left(\alpha_{d,ij} e^{\varepsilon_{nd}}\right)^{1/\mu}\right)^{n-1}}{\sum_{i=1}^{\mathcal{Y}} \sum_{j=i+1}^{\mathcal{Y}} \left(\sum_{d \in N_{ij}} \left(\alpha_{d,ij} e^{\varepsilon_{nd}}\right)^{1/\mu}\right)^{n-1}} \tag{4-20}
\end{equation}

Consequently, the likelihood function for the proposed GEV model can be written as:

\begin{equation}
L(\beta, \mu, \phi) = \prod_{n=1}^{N} \prod_{d} P_{nd} Y_{nd} \tag{4-21}
\end{equation}

where, \(Y_{nd}\) takes the value 1 if household \(n\) chose dwelling \(d\), and 0 otherwise.

As discussed in chapter 2, the main advantage of the GEV formulation is the closed-form expression for the choice probabilities, which is crucial in estimation of models with a large number of alternatives. The spatial correlations among alternatives are captured

\textsuperscript{22} It should be noted that although the proposed GEV model can accommodate the spatial correlation among dwelling units within a closed-form formulation, it cannot accommodate for the unobserved taste heterogeneity. The unobserved taste heterogeneity across decision makers can also be accommodated by superimposing a mixing distribution into the proposed GEV model which leads to an MGEV model structure (see Guo and Bhat, 2004a; Hess et al., 2005).
in the GEV class of models in a very limited way, however, based on a pre-assumed correlation structure. Previously developed GEV models (including the model proposed in this section) which accommodate spatial correlation among alternatives do not address spatial heteroscedasticity. The next section therefore proposes a novel Mixed Logit formulation to capture the spatial correlation in a more flexible way in zonal-based dwelling-level models. The proposed model structure also allows for capturing spatial heteroscedasticity in zonal-based dwelling–level models at the additional cost of estimating the heteroscedastic standard deviations parameters.

4.2.3 The Mixed Logit Formulation

As discussed in chapter 2, some authors have proposed Mixed Logit formulations to provide a more general correlation structure and to capture heteroscedasticity in zonal-based residential location choice models (see Vichiensan et al., 2005). This section extends the zonal-based Mixed Logit models to capture the spatial correlation and heteroscedasticity in zonal-based dwelling-level models.

Since implementing a fully unrestricted error correlation structure using a Mixed Logit model is problematic in choice situations with a large number of alternatives, similar to Vichiensan et al. (2005), the model proposed here adopts a Spatial Auto-Regressive (SAR) framework since this allows the capture of the general error correlation using a parsimonious parametric specification. In a fully unrestricted error correlation structure model, the number of parameters in the error structure grows quadratically with the number of alternatives; while in a SAR framework, the number of parameters in the error structure grows linearly with the number of alternatives. This is the main advantage of SAR frameworks (see Ben-Akiva et al., 2001).

Similar to section 4.2.2, in the development of the Mixed Logit model it is assumed that: (i) dwellings that are in the same zones have some common unobserved attributes, and (ii) dwellings that are in different zones also have some common unobserved attributes which decrease as the distances between the zones increase. Here, an extended SAR framework is proposed which allows for such a complex error structure. For the sake of simplicity we adopted the factor analytic form to present the model.
The utility function of the Mixed Logit model in a vector form can be written as:

\[ U_n = V_n + \xi_n + \nu_n \]  

(4-22)

where,

\( \nu_n \) is the idiosyncratic error term,

\( \xi_n = [\xi_{n1}, ..., \xi_{nD}] \) is a vector of disturbances at the dwelling-level and defined as:

\[ \xi_n = M\eta_n \]  

(4-23)

where, \( M \) is a \((D \times Z)\) membership matrix, and \( \eta_n \) is a vector of the zonal-level disturbances \((\eta_n = [\eta_{n1}, ..., \eta_{nZ}] )\).

The membership matrix assigns the same error term for those dwellings belonging to a particular zone and a different error term for dwellings that do not belong to the same zone, as:

\[ M = [m_{dz}], \quad m_{dz} = \begin{cases} 1 & d \in z \\ 0 & d \notin z \end{cases} \]  

(4-24)

Therefore, the dwellings that are in the same zone have the same zonal disturbance \( \eta_{nz} \) (but have different idiosyncratic error terms), and dwellings that are not in the same zone have different zonal disturbances \( \eta_{nz} \) (and have different idiosyncratic error terms). Table 4-3 illustrates the zonal-level and dwelling-level error terms for the residential choice example in the previous section.

Table 4-3- An Example of Dwelling-level and Zonal-level Disturbances

<table>
<thead>
<tr>
<th>Dwellings</th>
<th>Zones</th>
<th>Error Terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>( \nu_{n1} + \eta_{n1} )</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>( \nu_{n21} + \eta_{n2} )</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>( \nu_{n3} + \eta_{n2} )</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>( \nu_{n4} + \eta_{n2} )</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>( \nu_{n5} + \eta_{n3} )</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>( \nu_{n6} + \eta_{n3} )</td>
</tr>
<tr>
<td>7</td>
<td>4</td>
<td>( \nu_{n7} + \eta_{n4} )</td>
</tr>
<tr>
<td>8</td>
<td>4</td>
<td>( \nu_{n8} + \eta_{n4} )</td>
</tr>
</tbody>
</table>

The heterogeneity across decision makers can be also accommodated by specifying the parameters in the utility function as random variables.
We define the zonal-level disturbances $\eta_n = [\eta_{n1}, ..., \eta_{nz}]$ as a Spatial Auto-Regressive process (SAR) as:

\[ \eta_n = \rho W \eta_n + T \zeta_n, \quad \zeta_n \sim N(0, I_Z) \]  
\[ \eta_n = (I_Z - \rho W)^{-1} T \zeta_n \]  

where,

$W$ is a $(Z \times Z)$ weight matrix.

$\rho$ is an unknown parameter to be estimated (usually referred to as the correlation coefficient).

$\zeta_n = [\zeta_{n1}, ..., \zeta_{nz}]$ is a vector of independent standard normal distribution across zones.

$T$ is a $(Z \times Z)$ diagonal matrix containing the standard deviations of each zonal-level disturbance. Under the homoscedasticity assumption: $T = \sigma I_Z$

The weight matrix is defined as:

\[ W = [w_{ij}], \quad w_{ij} = \frac{w'_{ij}}{\sum_{k=1}^{Z} w'_{ij}} \]  

where, $w_{ij}$ describes the influence of each $\eta_{zn}$ on the others and can be defined based on the contiguity or proximity of zones.

Assuming that the spatial correlation exists only among contiguous zones, $w'_{ij}$ can be defined as:

\[ w'_{ij} = \begin{cases} 1 & \text{if } i \text{ and } j \text{ are contiguous} \\ 0 & \text{otherwise} \end{cases} \]  

$w_{ij}$ can also be defined based on the proximity measure between pairs of alternatives as:

\[ w'_{ij} = d_{ij}^\phi \]  

where, $d_{ij}$ is the distance between the centroids of zone $i$ and $j$, and $\phi$ is an unknown parameter to be estimated.
In summary, the utility function in a factor analytic form can be written as:

\[ w_{ij}' = d_{ij} \phi \]  

\[ U_n = V_n + M(I_Z - \rho W)^{-1}T \zeta_n + v_n \]  

Denoting \( \Gamma = M(I_Z - \rho W)^{-1}T \), and \( \Gamma_d \) as dth row matrix of \( \Gamma \), the conditional choice probability can be written as:

\[ P_{nd|\zeta_n} = \frac{\exp(V_{nd} + \Gamma_d' \zeta_n)}{\sum_{d'} \exp(V_{nd'} + \Gamma_d' \zeta_n)} \]  

The unconditional choice probability can therefore be evaluated as:

\[ P_{nd} = \int \frac{\exp(V_{nd} + \Gamma_d' \zeta_n)}{\sum_{d'} \exp(V_{nd'} + \Gamma_d' \zeta_n)} f(\zeta_n) d\zeta_n \]  

where, \( f(\zeta_n) \) is a multivariate standard independent normal density function. It should be noted that the dimension of the above integration is in the order of the number of zones.

Finally, the likelihood function for heteroscedastic disturbances and homoscedastic disturbances is:

Heteroscedastic:

\[ L(\beta, \rho, \phi, vec(T)) = \prod_{n=1}^{N} \prod_{d} P_{nd}^{y_{nd}} \]  

Homoscedastic:

\[ L(\beta, \rho, \phi, \sigma) = \prod_{n=1}^{N} \prod_{d} P_{nd}^{y_{nd}} \]  

where,

\( vec(T) \) vectorises the unique elements of \( T \) which are the heteroscedastic standard deviations (i.e., \( \sigma_1, \ldots, \sigma_Z \)).

\( Y_{nd} \) takes the value 1 if household \( n \) chose dwelling \( d \), and 0 otherwise.
4.2.4 Estimation Challenges

As discussed in chapter 2, the estimation of models that have closed-form choice probabilities (such as the proposed MNL and GEV models) is conducted using the Maximum Likelihood (ML) method. McFadden (1978) and Guevara and Ben-Akiva (2013) have also proved that the parameters of MNL models and GEV models can be estimated consistently based on a random sample of alternatives. Therefore, the estimation of the proposed dwelling-level MNL model and the GEV model can be conducted based on a ML approach applying a random sampling of alternatives strategy in order to reduce the estimation time.

In the case of the GEV model, the alternative-specific correction terms and expansion factors should be considered in order to achieve consistent parameter estimates based on the sampling of alternatives strategy as discussed in Guevara and Ben-Akiva (2013). Here, it is worth noting that the proposed GEV model structure can be estimated consistently based on a random sample of alternatives since only the sum of exponentials (i.e., \( \sum_{d \in N_l} (\alpha_{d,ij} e^{\nu_{nd}})^{1/\mu} \)) will be affected in the case of sampling of alternatives. From this perspective, the proposed GEV model is similar to the Nested Logit and Cross-Nested Logit examples given in Guevara and Ben-Akiva (2013). Hence, a similar approach can be followed to estimate the proposed model with a large number of dwellings.

Estimation of models in which the evaluation of choice probabilities involve intractable integrals (such as the proposed Mixed Logit model) are usually conducted using the Maximum Simulated Likelihood (MSL) approach (Bhat, 2001), based on the numerical approximation of multidimensional integrals, or using the Maximum Approximate Composite Marginal Likelihood (MACML) approach (Bhat, 2011b), based on the analytical approximation of integrals. Additionally, as discussed in chapter 2, numerical experiments show that the sampling of alternatives approach for Mixed Logit models may not affect the parameter estimates (see Nerella and Bhat, 2004; Azaiez, 2010; Guevara and Ben-Akiva, 2012; Lemp and Kockelman, 2012), even though this is not proved formally in the literature.
Although the dimensionality of integration in the likelihood function is to the number of zones rather than the number of dwellings in the proposed Mixed Logit model, the high computation cost of the MSL approach means that it becomes infeasible in urban areas with a large number of zones. In addition, the simulation noise in the MSL approach increases at medium to high dimensional integration which may result in problems in convergence of the optimisation algorithm as well as errors in the estimated covariance matrix. The MACML approach, on the other hand, is claimed to overcome the difficulties of the MSL approach and can potentially be used to estimate high dimensional models such as the proposed Mixed Logit model (Bhat, 2011b)

The estimation of MNP models using the MACML approach is straightforward and computationally very efficient. Estimation of Mixed Logit models using the MACML approach, however, can also be conducted by using a normal scale mixture representation for the extreme value error terms, although this requires an additional computational layer (Bhat, 2011a). Here, it seems that the benefit of Mixed Logit models using sampling of alternatives might justify the additional computational effort of estimating Mixed Logit models compared to MNP models. This is an interesting issue in the estimation of complex spatial choice models which invites future work in this area.

4.3 Constructing the Universal Choice Set in Dwelling-level Models:
A Dwelling Synthesising Approach

As discussed earlier, disaggregated dwelling supply data, which provides information on non-chosen alternatives, is required to construct the universal choice set in dwelling-level models. Unfortunately disaggregated dwelling supply data is not available for many metropolitan regions and, in the absence of this type of data, some authors have proposed to construct the choice set by taking random samples from the pool of all chosen dwellings in the entire household survey (see Zhou and Kockelman, 2008). This will result in biased parameter estimates, however, because the set of chosen dwellings in the household survey is not, in general, a random sample of all dwellings in the area. In fact, the household survey can be viewed as a choice-based sampling protocol because we have only observed the chosen alternatives. Since we do not know the choice probabilities, the sampling correction method for choice-based sampling cannot be applied here (see Manski and Lerman, 1977).
In order to avoid this limitation, some studies have used independent dwelling supply data in order to acquire the attributes of the non-chosen alternatives, as discussed in chapter 2. The combined use of household survey data and independent dwelling supply data requires matching the observations of two data sources. Guevara (2010), for example, used a nearest-neighbour approach for matching the two data sources. Given that the two data sources might be collected at different times and they might have different structures for categorical variables, matching will inevitably be error-prone. This might also result in biased parameter estimates and should be further explored.

This study proposes to impute the attributes of the non-chosen alternatives when dwelling supply data is unavailable. In fact, the required information about the non-chosen alternatives is imputed in other choice situations such as route choice and mode choice as well (e.g., travel time of non-chosen alternatives are imputed from network skims). Imputing the attributes of non-chosen alternatives will result in the measurement error bias in discrete choice models. It is well-known in the econometric literature that the measurement error in continuous models will result in biased inference and inconsistent estimates (see Carroll et al., 2010). Disaggregate discrete choice models are frequently estimated using variables with substantial measurement errors, however (Brownstone, 2001; Hellerstein, 2005).

Here we propose a synthesising approach to impute the attributes of non-chosen dwelling units in the residential location choice context and we numerically illustrate the abilities of synthesising algorithms to reduce the bias of choice set construction in dwelling-level models. Further theoretical work is required to support our simulation results, however.

In the synthesising approach, the population of dwelling units are estimated by combining aggregated dwelling data and a sample of disaggregated dwelling data. The choice set in the proposed approach is constructed by taking a sample from the synthesised dwellings for the study area and adding the chosen alternatives.

Synthesiser algorithms were originally developed to construct microdata representing the characteristics of the decision makers as the baseline households and persons population for microsimulation travel demand models (see Beckman et al., 1996; Ryan
et al., 2009). Following Beckman et al. (1996), several population synthesising approaches have been proposed in the literature. Although enhancement of population synthesising algorithms is still an active research area for transport modellers, the fundamental issues relating to the overall population synthesising approach have been addressed in the literature (see Müller and Axhausen, 2010; Pritchard and Miller, 2012). This study adopts the algorithm of Guo and Bhat (2007b) to create the synthesised population of dwelling units (see Appendix B for a comprehensive discussion on the synthesising algorithms).

The dwelling synthesising process begins with identifying the set of controlled variables for which aggregated dwelling supply data (i.e., marginal tables) are available, such as dwelling prices (as a categorical variable), dwelling sizes (i.e., number of beds), and dwelling types. Aggregated dwelling supply data is available for many metropolitan regions and can be acquired from population census data or housing census data.

In addition to aggregated dwelling supply data, a sample of disaggregated data is also required (i.e., microdata). This disaggregated data provides the seed joint distribution across the controlled variables (i.e., the correlation structure among dwelling attributes) and also provides a set of micro records from which dwellings will be drawn to form the synthetic population of dwellings. The disaggregated dwelling data used in this study is derived from the chosen dwellings in the household survey.

The synthesising algorithm first estimates the zonal joint distributions based on the seed joint distribution and marginal tables using the Iterative Proportional Fitting (IPF) procedure. The synthesised population of dwellings is then created by probabilistically drawing dwellings from the microdata according to the zonal joint distributions.24 Having synthesised the population of dwelling units, the choice set can be constructed by taking a random sample of non-chosen alternatives from the synthesised dwelling data and adding the chosen alternative. The next section presents the construction of the Monte Carlo experiment used in this study to validate the dwelling synthesising approach.

---

24 We do not incorporate the household weights (available in the household survey) in the probabilistic drawing process since these weights are associated to the attributes of households rather than the attributes of dwellings. The drawing process, therefore, only considers the zonal joint distributions computed in the IPF procedure.
4.4 Validation of the Choice Set Synthesising Approach

4.4.1 Monte Carlo Experiment

As part of this study a Monte Carlo experiment was conducted to validate the proposed synthesising approach of choice set construction and to illustrate the bias of constructing the choice set based only on the pool of chosen alternatives in the entire household survey. The true model in this experiment is a zonal-based dwelling-level MNL model based on the following utility function:

\[ V_{nd} = \beta_1 P_d + \beta_2 DS_d + \beta_3 DD_d + \beta_4 SQ_{z(d)} + \beta_5 A_{z(d)} + \beta_6 TD_{n,x(d)} + \beta_7 I/P_{nd} + \beta_8 HS/DS_{nd} \] (4.36)

The choice set of the true model is constructed by taking a sample of all available dwellings in the study area. Table 4-4 describes the notation of variables and the true parameters. To evaluate the impact of different choice set construction scenarios, we have estimated the following models (including the true model) based on the simulated data:

- **Model 1**: Choice set is constructed by taking a sample from independent dwelling supply data and adding the chosen alternatives (i.e., the true model).
- **Model 2**: Choice set is constructed by taking a sample from the synthesised dwellings for the study area and adding the chosen alternatives.
- **Model 3**: Choice set is constructed by taking a sample from the pool of chosen alternatives in household survey.

<table>
<thead>
<tr>
<th>Notation</th>
<th>Definition</th>
<th>Beta</th>
</tr>
</thead>
<tbody>
<tr>
<td>P</td>
<td>Dwelling price</td>
<td>-1</td>
</tr>
<tr>
<td>DS</td>
<td>Number of beds (dwelling size)</td>
<td>+1</td>
</tr>
<tr>
<td>DD</td>
<td>Dummy for detached dwellings (dwelling type)</td>
<td>+1</td>
</tr>
<tr>
<td>SQ</td>
<td>School quality</td>
<td>+1</td>
</tr>
<tr>
<td>A</td>
<td>Accessibility</td>
<td>+1</td>
</tr>
<tr>
<td>TD</td>
<td>Travel distance to households’ workplaces</td>
<td>-1</td>
</tr>
<tr>
<td>I/P</td>
<td>Interaction of household income and dwelling price (ratio)</td>
<td>-1</td>
</tr>
<tr>
<td>HS/DS</td>
<td>Interaction of dwelling size and household size (ratio)</td>
<td>-1</td>
</tr>
</tbody>
</table>

The performances of the different choice set construction scenarios defined above are evaluated based on their ability to retrieve the true parameters.
Simulation of the synthetic city has six stages: (i) simulation of cross-correlated random fields for spatial distribution of dwellings, employment, and average dwelling attributes within gridcells (see Figure 4-4), (ii) simulation of disaggregated dwellings’ attributes based on average gridcell attributes, (iii) simulation of the OD matrix, school quality, and accessibility measures at the zonal-level, (iv) simulation of households’ attributes, (v) simulation of dwelling prices, and (vi) assigning dwellings to households according to the true model (i.e., Model 1).

![Images of various statistical distributions and attributes for dwellings and employment](image)

**Figure 4-4: An Example of Cross-correlated Random Fields Used in the Simulation of the Synthetic City**

The simulation of the synthetic city is based on the following assumptions:

- The number of dwellings and the numbers of households are equal in the city

---

25 The simulation of random fields has been conducted based on the LU triangular decomposition of the covariance matrix (Davis, 1987).
Households have perfect information on all dwellings in the city
Households’ preferences on dwelling attributes and locations are represented by an MNL model (IIA assumption)
Analyst knows the true utility function of households
Households are price takers and dwelling prices can be determined based on the disequilibrium approach and based on the equilibrium approach

Dwelling prices can be simply simulated exogenously assuming the prices are a hedonic function of dwelling and zonal attributes (i.e., the disequilibrium approach). The variables and parameters used in the hedonic price model are tabulated in Table 4-5. Since the prices in this approach are not the market clearing prices, some dwellings will be assigned to more than one household, while some others will remain unassigned during the Monte Carlo sampling process.

<table>
<thead>
<tr>
<th>Variables</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of bedrooms (dwelling size)</td>
<td>+1</td>
</tr>
<tr>
<td>Dummy for detached dwellings (dwelling type)</td>
<td>+1</td>
</tr>
<tr>
<td>School quality</td>
<td>+1</td>
</tr>
<tr>
<td>Accessibility</td>
<td>+1</td>
</tr>
</tbody>
</table>

We have also simulated the prices endogenously in order to match the demand and supply following the stochastic dwelling equilibrium approach (see Anas, 1982). In this approach, equilibrium dwelling prices are determined by setting the expected excess demand for each dwelling to zero in an iterative process described in Appendix A. It should be noted that assignment of dwellings to households based on the equilibrium prices will not be a one to one match because of the stochastic nature of the equilibrium. However, using equilibrium prices in a demand model will result in a larger number of one to one matches (roughly 60% in our settings) compared to exogenous prices in the disequilibrium approach (roughly 30% in our settings). It should also be noted that the expectation of the assignment (repeating the assignment multiple times) in the stochastic equilibrium approach will result in a one to one match.

---

26 Hence, there is no endogeneity due to omitted variables as all of the variables that are considered by decision makers are observed by the analyst (see Bayer et al., 2005; Guevara and Ben-Akiva, 2006).
Having simulated the synthetic city, household survey data (i.e., households’ chosen dwelling attributes, dwelling location and households’ attributes), aggregated dwelling supply data, and zonal attributes can be simply derived from the synthetic city. It should be noted that the simulated dwelling prices based on both the disequilibrium and equilibrium approach.
equilibrium approach are continuous variables. We have therefore converted these into categorical variables in order to be able to derive the aggregated price data and use the dwelling synthesiser approach. The next section presents the comparison of the three models described earlier for different household survey sample sizes for both disequilibrium and equilibrium dwelling prices.

4.4.2 Monte Carlo Results

To assess the impact of choice set construction approaches in the estimation of model parameters and to evaluate the performance of the synthesising approach, three models were estimated (i.e., Model 1, Model 2 and Model 3) for different household survey sizes (i.e., NH=500, NH=1000, and NH=2000) and for 10 repetitions. For each model, the average of Rho squares of different runs, as well as the Mean Absolute Deviation Error (MADE), and the Root Mean Square Error (RMSE) against the true value of parameters have been calculated and reported in Table 4-6 and Figure 4-7.

The results clearly show the superiority of the proposed approach (Model 2) against the conventional approach (Model 3) for constructing the choice set in dwelling-level models since the Rho Square, MADE and RMSE measures are very similar between Model 1 (i.e., the true model) and Model 2 (i.e., the proposed approach) in the disequilibrium case. All of the parameters of Model 2 are found to be highly significant, while this is not the case for Model 3. It should also be noted that the signs of estimated parameters for Model 2 are completely compatible with the true signs of parameters, while this is not true for Model 3.

Surprisingly, the goodness of fit measures (i.e., Rho Square) of models with equilibrium dwelling prices are very low in Model 1 and Model 2 compared to the disequilibrium case. The low goodness of fit of equilibrium models is accompanied by instability in estimation of parameters which results in higher MADE and RMSE for Model 1 and Model 2. Although the RMSE and MADE measures of Model 2 are smaller than for Model 3 in the equilibrium case, these measure are much higher than the corresponding measures in the disequilibrium case. In the equilibrium case, all parameters of Model 2 are also completely compatible with the true signs of parameters, while this is not true for Model 3.
The reason for achieving very small goodness of fit for models in the equilibrium case is rooted in a multicollinearity problem that may arise in the equilibrium market clearing approach. Since the prices are determined endogenously based on the demand model in the equilibrium approach, we suspect the equilibrium prices to be highly correlated with other variables of the demand model which results in a multicollinearity problem. This issue should be further explored in the literature.

The simulation results of this section, in general, provide empirical evidence for the validity of the proposed dwelling synthesising approach. Further theoretical work is required, however, to support the Monte Carlo evidence. The next section applies the proposed dwelling synthesising approach to estimate a dwelling-level residential location choice for London.

### Table 4-6: Validation of the Choice Set Synthesising Approach for Disequilibrium and Equilibrium Market Clearing

<table>
<thead>
<tr>
<th>Disequilibrium, NH=500</th>
<th>Equilibrium, NH=500</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Model 1</strong></td>
<td><strong>Model 2</strong></td>
</tr>
<tr>
<td>MADE</td>
<td>0.0344</td>
</tr>
<tr>
<td>RMSE</td>
<td>0.0446</td>
</tr>
<tr>
<td>RHO2</td>
<td>0.3651</td>
</tr>
<tr>
<td>Disequilibrium, NH=1000</td>
<td>Equilibrium, NH=1000</td>
</tr>
<tr>
<td><strong>Model 1</strong></td>
<td><strong>Model 2</strong></td>
</tr>
<tr>
<td>MADE</td>
<td>0.0481</td>
</tr>
<tr>
<td>RMSE</td>
<td>0.0649</td>
</tr>
<tr>
<td>RHO2</td>
<td>0.3497</td>
</tr>
<tr>
<td>Disequilibrium, NH=2000</td>
<td>Equilibrium, NH=2000</td>
</tr>
<tr>
<td><strong>Model 1</strong></td>
<td><strong>Model 2</strong></td>
</tr>
<tr>
<td>MADE</td>
<td>0.0319</td>
</tr>
<tr>
<td>RMSE</td>
<td>0.0392</td>
</tr>
<tr>
<td>RHO2</td>
<td>0.3513</td>
</tr>
</tbody>
</table>
This section describes the application of the proposed dwelling synthesising approach to estimating an MNL dwelling-level residential location choice model for the Greater London area. Greater London is divided into 32 London boroughs and the City of London, and has 7,172,091 inhabitants and 3,015,997 households according to the 2001...
census. London covers 1,594,720 square meters (thousands) and occupies 1.2 per cent of England.

4.5.1 Data

The primary data source used in this analysis is the 2002 London Household Survey (LHS) data. The survey was conducted for the Greater London Authority (GLA) and provided the GLA with the essential data for development of policies across a wide range of policy areas, including quantifying the linkages between different aspects of needs such as housing affordability, poverty, employment, health, etc. (Greater London Authority, 2003).

The LHS 2002 is a large sample survey (i.e., 8,158 households) which is a representative sample of the population of London. This study focused on the residential location choice of the owner's sub-market, which includes households that own outright and those owned through mortgage/shared ownership arrangements. 4,479 out of the total of 8,158 households fall into this category. Unfortunately, there are lots of missing variables in our dataset and the estimation sample only includes 1,263 households (see Table 4-7).

| Table 4-7: Counts of Remaining Cases in Data Cleaning |
|---------------------------------|-----------|
| Households                      | Remaining |
| Owner’s sub-market              | 4479      |
| Omitting Cases Missing Dwelling Attributes | 2852      |
| Omitting Cases Missing Household Income | 2228      |
| Omitting Cases Missing Household Employment Location | 1263      |

The population of London based on the 2011 census is 8.2 million, an increase of 12 per cent from 2001. The results of the 2011 census were only recently published, however, and therefore this study uses the 2001 census in order to be consistent with the other data sources used.
Households’ residential locations in the LHS 2002 are available at the ward level. In order to be consistent with other used variables, we matched the locations of households to Travel Analysis Zones (TAZs). The LHS 2002 included a wide range of dwelling attributes, but the dwelling attributes used in the final model specification were: council tax band, dwelling type (i.e., detached, semi-detached, terraced, and flat), dwelling size (i.e., number of bedrooms). The household characteristics (available in LHS 2002) used in the model estimation are: annual household income, household size, and the employment location of household members. Household characteristics

28 A ward is an electoral district and the primary unit of British administrative organisation.
29 We matched wards and TAZs based on the centroids of wards. Some inaccuracies might be involved in this approach as the level of granularity of wards and TAZs in the Greater London Area are similar and their boundaries overlap.
30 Council tax is a form of local taxation which is used to help pay for the services that the Local Council provides. The Valuation Office Agency (VOA), which is an executive agency of the UK government values the capital value of properties and divides them into bands which are then used to calculate the council tax. Previous findings show that the council tax band proxy is a good indicator of dwelling prices (see Chiaradia et al., 2009).
are included in the utility function using interaction variables. Employment locations are available at either postal sectors (e.g., SW7 2) or postal districts (e.g., SW7). We matched the employment locations to TAZs in order to be consistent with other variables used. The employment locations were required to calculate the commute variable, as discussed later.

There are also some locational variables included in the estimation of the zonal-based dwelling-level residential location choice model. The locational variables used in this study were compiled from various spatial data sources. Here, we describe briefly the different locational variables that were used in the final model specification; Appendix C describes the complete set of spatial data sources and locational variables as well as data preparation stages.

The 2001 census tables provide a rich set of variables for consideration in the model specification. The census variables included in the final model specification include number of residents per hectare and average household size.

Using the Annual Business Inquiry (ABI) data (see Appendix C), we computed accessibility to employment based on Hansen’s (1959) formulation as:

\[ A_{i}^{Emp} = \frac{1}{N} \sum_{j=1}^{N} E_{j} T_{ij} \]

where, \( E_{j} \) is total number of employees in zone \( j \), and \( T_{ij} \) is the travel time from zone \( i \) to zone \( j \).

The commute variable (i.e., travel time from households’ workplace zone(s) to residential zones) is computed from the zone-to-zone travel time matrix in auto mode (i.e., acquired from the London Transportation Studies (LTS) model), assuming that the households’ employment location choices are predetermined and exogenous to residential location choices. For households with more than one worker, the aggregate values across all workers in the household are calculated.

In addition to LHS 2002 and locational variables (acquired from various data sources), we have also used 2001 census dwelling data to create the synthetic population of households.

---

31 Some inaccuracies might be involved in matching employment locations to TAZs since the level of granularity of postal sectors/postal districts are similar to the TAZs, and their boundaries overlap.
dwelling units for the study area. The aggregated data for tenure type, council tax band, dwelling type and dwelling size variables tables are available in the 2001 census and have been used in this study. The census dwelling data has been aggregated to match the categories of the LHS 2002 data and spatially aggregated to TAZs. Since we do not need households’ income and employment locations to synthesise the dwellings (and to construct the choice set), the microdata used for the synthesising algorithm contains 4,494 cases including all tenure types.

4.5.2 Results

The LTS 2002 survey provides information on the attributes of the chosen dwellings as well as the characteristics of households. The attributes of non-chosen dwellings should ideally be derived from independent dwelling supply data. In the absence of such data, here we apply the proposed dwelling synthesising approach in order to construct the choice set.

The dwelling synthesiser combines census aggregated dwelling supply and the dwelling attributes available in the LHS 2002 to create the synthetic population of dwelling units for the Greater London area. The set of controlled variables include: tenure type, council tax band, dwelling type and dwelling size. The dwelling synthesiser was performed at the TAZ level and a total of 3,107,443 dwellings were synthesised for the 879 TAZs in Greater London. The synthesised dwellings were filtered based on the tenure type variable and reduced to 1,747,868 dwellings since we focused on the owner’s sub-market. The choice set was then constructed by taking a random sample from the synthesised dwellings and adding the chosen alternatives.

We estimated a zonal-based dwelling-level model (i.e., Model 1) based on the proposed choice set synthesising approach. The estimation is conducted for 10 repetitions of the synthesising algorithm in order to reflect the uncertainty involved in the choice set synthesising approach and show the stability of the results. The results of the average of 10 repetitions are reported in Table 4-8. There are two sources of errors involved in the estimation of the parameters of the Model 1, one concerns the error of the maximum

---

32 The empirical application of this study was limited to the conventional MNL model for the sake of simplicity since the objective this section was to illustrate the practicality of the proposed choice set synthesising approach. We did not also address the price endogeneity problem that might arise in dwelling-level residential location choice models (see Bayer et al., 2005; Guevara and Ben-Akiva, 2006). Estimation of the proposed GEV model using real data remains for future work of this thesis.
likelihood estimator which is used to calculate the t-statistics and another concerns the synthesising and the sampling of alternatives error. The standard deviations of the parameter estimates across different repetition of the synthesising algorithm and the sampling of alternatives are also reported in in Table 4-8. For comparison purposes, we have also estimated a model (i.e., Model 2) based on the conventional choice set construction approach (i.e., taking a sample from a pool of chosen alternatives in LHS 2002). Table 4-8 also reports the estimation results from Model 2.

### Table 4-8: Dwelling-level Residential Location Choice Model Estimation Results

<table>
<thead>
<tr>
<th>No</th>
<th>Parameter</th>
<th>Choice Set Sampled from Synthesised Dwellings (Model 1)*</th>
<th>Choice Set Sampled from Chosen Dwellings (Model 2)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Estimate</td>
<td>t-stat</td>
<td>Estimate</td>
</tr>
<tr>
<td>1</td>
<td>Dwelling price (council tax band)</td>
<td>-2.296 [0.030]</td>
<td>-28.035</td>
</tr>
<tr>
<td>2</td>
<td>Dummy for detached dwellings</td>
<td>0.964 [0.078]</td>
<td>6.815</td>
</tr>
<tr>
<td>3</td>
<td>Dwelling size (number of bedrooms)</td>
<td>0.379 [0.024]</td>
<td>13.631</td>
</tr>
<tr>
<td>4</td>
<td>Number of residents per hectare</td>
<td>-0.010 [0.001]</td>
<td>-9.117</td>
</tr>
<tr>
<td>5</td>
<td>Accessibility to employment</td>
<td>-0.019 [0.001]</td>
<td>-10.183</td>
</tr>
<tr>
<td>6</td>
<td>Interaction of household size with average household size (absolute differences)</td>
<td>-0.628 [0.058]</td>
<td>-4.616</td>
</tr>
<tr>
<td>7</td>
<td>Interaction of household income with dwelling price(^\text{34}) (ratio)</td>
<td>-0.295 [0.102]</td>
<td>-3.887</td>
</tr>
<tr>
<td>8</td>
<td>Interaction of household size with dwelling size (ratio)</td>
<td>-2.951 [0.071]</td>
<td>-14.433</td>
</tr>
<tr>
<td>9</td>
<td>Travel time from households’ workplace zone(s) to residential zones</td>
<td>-0.053 [0.001]</td>
<td>-31.415</td>
</tr>
</tbody>
</table>

**Summary statistics**

<table>
<thead>
<tr>
<th></th>
<th>1263</th>
<th>1263</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Observations</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Universal Choice Set</td>
<td>1747868</td>
<td>3033</td>
</tr>
<tr>
<td>Sample Size</td>
<td>50</td>
<td>50</td>
</tr>
<tr>
<td>Likelihood Ratio Index ((\rho)^2)</td>
<td>0.508 [0.003]</td>
<td>0.242</td>
</tr>
</tbody>
</table>

*The estimated parameters of Model 1 are the average of 10 repetitions. The STDs are reported in the brackets.

33 This is the microdata used in the synthesising algorithm filtered based on the tenure type variable since we focused on the owner's sub-market (i.e., 3,033 cases).

34 In the calculation of this variable, continuous dwelling prices and household incomes were created based on council tax bands and income bands by generating random numbers corresponding to each category.
In general, the estimated models provide significant evidence for choice set construction bias in dwelling-level residential location choice models. Similar to the results of the Monte Carlo experiment that were presented in the previous section, Model 1 has a much higher goodness of fit compared to Model 2. The parameters and t-statistics of dwelling attributes of Model 1 and Model 2 are significantly different.

Both models estimated the expected negative sign for the dwelling price parameters which are also statistically different from zero. The estimated parameters are very different, however, and the estimated t-statistic of the dwelling price parameter in Model 1 is much higher than the t-statistic of the parameter in Model 2. Model 1 found a positive and statistically significant parameter for dwelling type variable (i.e., dummy for detached dwellings) as expected, but this parameter was found to be insignificant in Model 2. The dwelling size parameter has the expected positive sign in Model 1 while Model 2 estimated a negative parameter for the dwelling size variable. Model 1 also estimated a much higher t-statistic (in absolute value) for the dwelling size parameter.

All locational variables and interaction variables have the expected signs and are statistically significant in both models. Although the parameters and t-statistics for locational variables and interaction variables of Model 1 and Model 2 are more similar compared to the ones for dwelling attributes, they are still significantly different.

It should be noted that the choice set construction bias in the estimation of dwelling-level models might be misleading for researchers. A reasonably high goodness of fit and highly significant parameter estimates might be achieved if the choice set is constructed based on the pool of chosen dwellings in the household survey (as in the case of Model 2 and Zhou and Kockelman (2008)), however, the results could be significantly biased.

Finally, we should acknowledge the limitations of the dataset used for estimation of our models. First, the sample size of the microdata used to synthesise dwelling units is very small (less than 1 per cent) which can potentially introduce biases in the synthesising process. Secondly, the dataset contains many missing values, especially for council tax band and employment location variables. Although one could potentially impute these missing values, we decided to estimate the models based on complete cases since there are already many uncertainties involved in the proposed choice set synthesising approach. Since almost 70% of observations were omitted, the estimation sample might
not be a representative sample of the owner’s sub-market of London. Consequently, the estimated models based on this data should be seen as preliminary results. Our empirical results based on the real data, however, can be seen as a proof of concept when taken in addition to the Monte Carlo results.

### 4.6 Conclusion

This chapter set out to model residential location choices at the level of dwelling units in order to circumvent the aggregation bias of spatially aggregated models. From the methodological perspective, the IIA assumption of MNL models is not valid in residential location choice models because of the association of dwelling units to space. We have developed a GEV and a Mixed Logit formulation to relax the IIA assumption and capture the spatial correlation and spatial heteroscedasticity in the proposed zonal-based dwelling-level residential location choice model. Estimation of such model structures remains as future work for this thesis.

From the empirical perspective, a lack of dwelling supply data results in biases in dwelling-level residential location choice models as illustrated numerically in this chapter. We proposed a dwelling synthesising approach to reduce the bias of choice set construction in the absence of dwelling supply data. Our simulation results confirm the superiority of the proposed dwelling synthesising approach compared to the conventional approach. The proposed dwelling synthesising approach has also been applied to estimate a zonal-based dwelling-level residential location choice model for Greater London. The empirical analysis of this chapter can be improved by incorporating other dwelling and locational attributes in the model specification such as the locational crime rate.
CHAPTER 5

SPATIAL SEARCH AND CHOICE SET FORMATION IN RESIDENTIAL LOCATION CHOICE MODELLING

5.1 Overview

Households in residential location choice models are conventionally assumed to possess perfect information about the housing market. This leads to the development of models (including the model presented in the previous chapter) that assume that households consider all available alternatives in the market when they are making residential location decisions. As discussed in chapter 2, the random sampling of alternatives approach is typically applied in order to deal with the computational problem of the estimation of such models with a large number of alternatives. The perfect information assumption, however, contradicts the psychological concept of bounded rationality which results in the choice set formation problem, as discussed in chapter 2.

Conceptually, residential location choice is a dynamic spatial search process (see chapter 3), in which households are exposed to a dynamically changing set of residential alternatives from which, at any point in time, they assemble and evaluate a choice set of credible alternatives, before ultimately, at some point in time, making a selection. This underlying dynamic spatial search and choice process is typically unobserved, however,\textsuperscript{35} and the econometrical frameworks that can represent such behaviour are complex (see Lerman and Mahmassani, 1985; Caplin and Dean, 2011). This is why most researchers stick to the perfect information assumption in order to develop practical residential location choice models.

\textsuperscript{35}Taking advantage of novel data, some authors have formulated search-based models of the housing market which go beyond utility maximisation choice theory by taking into account the role of imperfect information and high search costs in the econometric modelling of housing demand (see Read, 1991; Anas, 1997; Van der Vlist et al., 2002). The application of such frameworks in applied land use modelling, however, is not well-established due to the requirement for novel data and the complexity of model structures.
An alternative approach is to acknowledge the existence of an underlying dynamic search process by formulating behaviourally plausible rules describing households’ search behaviour in a residential location choice context, such as an anchor-based search strategy (see chapter 2, 2.4.4). These behavioural rules can then be used to construct households’ choice sets exogenously using the importance sampling approach (see Elgar et al., 2009; Rashidi et al., 2012), or the deterministic constraint approach (see Parsons and Hauber, 1998).

An important research question is how these different approaches impact the predictive ability of a residential location choice model. This chapter addresses this gap by examining the performance of several exogenous choice set formation methods within the context of residential location choice in London. In particular, it compares the simple random sampling of alternatives against the deterministic constraint approach and the importance sampling approach (with and without considering alternative-specific correction terms). The alternative choice set formation approaches are assessed in terms of their prediction performances on a hold-out validation sub-sample and also in terms of estimated model parameters.

The rest of this chapter is organised as follows: section 5.2 presents the modelling framework for this research, including the details of the choice set formation models that were implemented. Section 5.3 presents the details of the empirical analysis. Section 5.4 presents the estimation and validation results for the different models. Section 5.5 concludes the chapter.

5.2 Modelling Framework

5.2.1 Spatial Search Strategies and Choice Set Formation

Huff (1986) investigated the spatial search behaviour of a sample of prospective home buyers based on their visited vacancies data in the San Fernando Valley of Los Angeles. He evaluated three complementary models for different spatial search strategies: (i) supply-constraint model, (ii) area-based model, and (iii) anchor-points model.

The supply-constraint model is motivated by the fact that households are more likely to search in areas with a greater supply of their desired housing. Assuming that the tendency to search in an area is proportional to the number of vacancies that meet the
requirement of households, the supply-constraint model relates the observed search behaviour to the underlying distribution of vacancies (meeting the household’s basic requirements) using a linear regression model. Although the supply-constraint model provides a good description of search behaviour, the model does not adequately capture the search patterns of individual households.

In order to acknowledge the effects of individuals' search strategies in observed search patterns, Huff (1986) also presented area-based and anchor-point models. These are refinements of the supply-constraint model in that both assume that the observed search pattern reflects the underlying distribution of vacancies as well as spatial biases in the household’s search strategy.

The area-based model acknowledges the existence of geographical sub-markets and assumes that households tend to continue searching in a particular community area once they have begun searching there. Consequently, areas with a large number of vacancies that meet households’ requirements are expected to have a greater chance of being visited (or being in their choice sets) than areas with fewer vacancies. The area-based model represents a major improvement over the performance of the supply-constraint model. Huff’s results also provide strong behavioural evidence for the existence of geographically defined sub-markets.

The anchor-points model assumes that a household’s intensity of search in an area is a function of that household’s attachment to critical nodes of interaction or anchor-points in the activity space (e.g., prior residence and workplace). Unlike supply-constraint model and area-based models, locational attributes (i.e., distances to the anchor-points) are introduced into the anchor-points model in an effort to explain observed differences among households’ search patterns. The results confirm that the anchor-points in a household’s search space do play an important role in structuring the spatial aspects of the household’s search strategy.

Huff (1986) described and empirically validated different search strategies that households adopt in the housing market. These different search strategies focus on different aspects of the residential search process and can potentially be used to construct the choice sets exogenously in residential location choice models. As discussed in chapter 2, the deterministic constraint approach and the importance
sampling approach (without correction terms) can be used to model choice sets exogenously in such spatial choice problems (see Figure 5-1).

In the context of office location choice, Elgar et al. (2009) generated the choice sets by oversampling from zones that have some office buildings in them compared with zones without any office buildings in order to generate choice sets while incorporating a supply-constraint search strategy (i.e., an importance sampling approach).

In practice, however, the main focus has been on the anchor-points search strategy for the generation of choice sets in spatial choice models. In fact, the deterministic choice set formation approach, which was discussed in chapter 2 in the context of destination choice, can be viewed essentially as an application of the anchor-points search strategy,

36 The distance to anchor-points can also be included in the utility function as an additional variable. In fact, the impedance variables in destination choice (e.g., distance to residential locations) and in residential location choice (e.g., distance to workplace locations) are the anchor-point variables. In the context of office location choice, Elgar et al. (2009) also considered two anchor-point variables in the specification of the utility function (i.e., distance to current location of the firm and distance to the owner’s residential location) and reported a highly significant parameter for these variables as well as significant improvements in the goodness of fit measures of models incorporating these variables. Implementing anchor-points in the utility function as additional variables instead of choice set generation is also in accordance with the Horowitz and Louviere (1995) interpretation of choice sets. They argued that the choice set formation stage is only a reflection of households’ preferences, and modelling choice behaviour as a two-stage process (choice set formation stage followed by actual choice from the smaller choice set) provides no information beyond that contained in the utility function.
where the origins of individuals’ trips are the anchor-points (see Landau et al., 1982; Thill and Horowitz, 1997; Scott, 2006).

Elgar et al. (2009) also implemented the anchor-points search strategy for choice set generation using the importance sampling approach. In this approach, choice sets are generated by oversampling alternatives around the anchor-points. Choice set generation using two anchor-points is also examined in the literature, which in general assumes that the two anchor-points are the foci of an ellipse (see Cirillo et al., 2003; Elgar et al., 2009).

### 5.2.2 Screening Models

An anchor-points search strategy provides a framework to capture the heterogeneity that exists in the spatial searching based on distances to households’ anchor-points. There are other sources of heterogeneity that can potentially explain the underlying searching behaviour, however. Some authors have attempted to develop explanatory models incorporating different explanatory variables to predict the probability that a particular alternative belongs to a decision maker’s choice set. Here, we refer to these models as screening models.

Screening models incorporate explanatory variables other than distance to anchor-points in order to determine the thresholds, in the deterministic constraint approach, and sampling weights in the importance sampling approach. Since screening models enable the incorporation of different sources of spatial heterogeneity (i.e., distance to anchor-points) and intra-personal heterogeneity (i.e., socioeconomic variables) into the searching process, they can potentially predict the choice sets more realistically.

In the context of shopping destination choice, van der Heijden and Timmermans (1984) developed a screening model based on a logistic regression formulation to predict the probability that an individual will possess information about a shopping centre based on the centre’s distance to the individual’s home (i.e., anchor-points), the size of the shopping centre, and a binary variable denoting the presence of intervening opportunities. The estimation of this model requires data on individual familiarity with shopping destinations.
Rashidi et al. (2012) applied the proportional hazard formulation in order to develop a screening model in a residential location choice context. In their proposed model, households do not consider an alternative if it is further than some distance threshold from the household’s workplace locations. The next section describes in more detail the hazard-based screening model that has been implemented in this study.

5.2.3 Hazard-based Screening Model

Rashidi et al. (2012) assumed that households will not consider an alternative if it is more expensive than threshold $P$, or if it is further from the household’s workplace than threshold $T$. This interpretation of acceptable housing price and acceptable commute time is similar to survival analysis, where the time it takes for events to occur is examined (e.g., time before a failure). Using this analogy, Rashidi et al. developed an innovative proportional hazard-based screening model where the dependent variable in the hazard model formulation is not time, but distance (i.e., travel distance to workplace locations).

In conventional survival analysis, the independent variable is time and the survival function is the probability that the time taken before the occurrence of an event (usually failure) is later than some specified threshold. In order to incorporate the effects of covariates associated with the event including in principle time varying covariates, a number of formulations are available including proportional hazard models and accelerated failure time models (see Hougaard, 2000).

This study uses the Weibull and the Log-logistic distributions for average work commute times and housing prices respectively. The results presented in the empirical analysis section justify the use of these distributions. The baseline hazard function $\lambda(t)$, and its corresponding survival function $S(t)$, for the Weibull model are as follows:

$$
\lambda(t) = \alpha \beta t^{\beta-1} \quad S(t) = \exp\left(-\int_0^t \lambda(t) \, dt\right) = \exp(-\alpha t^\beta)
$$

(5-1)

where, $\alpha$ is the scale parameter and $\beta$ is the shape parameter. In order to incorporate covariates into the model, the scale parameter in the baseline hazard function is reparameterised in terms of explanatory variables and regression parameters while the shape parameters are held fixed ($\alpha = \exp(\theta_0 - \theta X)$). Therefore, the proportional
hazard function of work commute time and its corresponding survival function are as follows:

\[
\lambda(t) = \beta t^{\beta - 1} \exp\left(\theta_0 - \hat{\theta}^\top \bar{X}\right) \quad S(t) = e^{-\int_0^t \lambda(s) \, ds} = \frac{1}{1 + \alpha t^\beta} \quad (5-2)
\]

The baseline hazard function and its corresponding survival function for the log-logistic model are as follows:

\[
\lambda(t) = \frac{\alpha \beta t^{\beta - 1}}{1 + \alpha t^\beta} \quad S(t) = \exp\left(- \int_0^t \lambda(s) \, ds\right) = \frac{1}{1 + \alpha t^\beta} \quad (5-3)
\]

where, \(\alpha\) is the scale parameter and \(\beta\) is the shape parameter. Similar to the Weibull model, the scale parameter in the baseline hazard function is reparameterised in terms of explanatory variables and regression parameters and the shape parameters are held fixed in order to incorporate covariates into the model. Therefore, the proportional hazard function of housing prices and its corresponding survival function are as follows:

\[
\lambda(p) = \frac{\exp\left(\theta_0 - \hat{\theta}^\top \bar{X}\right) \beta p^{\beta - 1}}{1 + \exp\left(\theta_0 - \hat{\theta}^\top \bar{X}\right) p^\beta} \quad S(p) = \frac{1}{1 + \exp\left(\theta_0 - \hat{\theta}^\top \bar{X}\right) p^\beta} \quad (5-4)
\]

Based on the definition of hazard function and survival function, the probability density functions of accepting a housing price and accepting a work commute time are:

\[
f(p) = \lambda(p) S(p) \quad f(t) = \lambda(t) S(t) \quad (5-5)
\]

Parameters of Weibull and Log-logistic models can be estimated using the maximum likelihood approach. The outputs of the screening model are the distribution of acceptable housing prices and acceptable work commute times conditional on the households’ socioeconomic attributes, as follows:

\[
f(t) = \beta t^{\beta - 1} \exp\left(\theta_0 - \hat{\theta}^\top \bar{X}\right) e^{-t^\beta \exp\left(\theta_0 - \hat{\theta}^\top \bar{X}\right)} \quad (5-6)
\]

\[
f(p) = \frac{\exp\left(\theta_0 - \hat{\theta}^\top \bar{X}\right) \beta p^{\beta - 1}}{1 + \exp\left(\theta_0 - \hat{\theta}^\top \bar{X}\right) p^\beta} \frac{1}{1 + \exp\left(\theta_0 - \hat{\theta}^\top \bar{X}\right) p^\beta} \quad (5-7)
\]

The structure of the hazard-based screening model enables us to model systematic inter-individual heterogeneity in choice set formation. In addition, Rashidi and Mohammadian (2012) included the unobserved heterogeneity into the screening model by introducing a gamma distribution; they concluded that incorporating the unobserved

110
heterogeneity in the parametric hazard formulation with the Weibull baseline hazard function provided a better fit with the commute distance data.

5.2.4 Different Choice Set Formation Approaches

In this chapter, we implement two different procedures to generate choice sets based on the hazard-based screening model: (a) the importance sampling approach, and (b) the deterministic constraint approach.

5.2.4.1 Importance Sampling Approach

Assuming that decision makers follow the anchor-points search strategy, Elgar et al. (2009) proposed to generate choice sets by oversampling of alternatives around the anchor-points. Similarly, the household’s choice set in a residential location choice context can be generated based on the distributions of acceptable commute times and housing prices derived from the hazard-based screening model.

It should be noted that the alternative-specific correction terms are not included in this approach since the estimated model is expected to incorporate households’ searching behaviour in the choice set formation stage. In fact, we can argue that the importance sampling approach (without correction terms) is equivalent to the deterministic choice set formation approach, as discussed in the chapter 2. Since the deterministic approach is hypothesised to be a behavioural approach for choice set formation, the alternative-specific correction terms are not considered in estimation to correct for the bias of estimating over a subset of alternatives. Based on the same argument, in order to incorporate the behavioural aspect of choice set formation in the importance sampling approach the alternative-specific correction terms should not be considered in estimation.

Rashidi et al. (2012) applied the hazard-based screening model (based on commute distance criteria) in order to generate households’ choice sets using an importance sampling approach. They used the acceptable commute distance probability distribution to compute the most desirable commute distances, which is defined as the commute distances corresponding to the maximum probability. Further, they computed the sampling weight of each alternative using an exponential distribution, as follows:
where, $W_{j,n}$ is the sampling weight of alternative $j$ for household $n$, $d_{j,n}$ is the commute distance from the workplace location of household $n$ to alternative $j$, and $\hat{d}_n$ is the most desirable commute distance computed from the acceptable commute distance distribution.

Rashidi et al. (2012) estimated a MNL residential location choice model incorporating alternative-specific correction terms in order to correct for the bias of importance sampling. They argued that an importance sampling approach including the correction factor provides a way of including behavioural choice set formation in the discrete choice model. As mentioned in chapter 2, however, no matter how the sampling weights are computed, as long as the alternative specific correction terms are included in the utility function in the estimation process, the results will be equivalent to the universal choice set. In the discrete choice modelling of residential location choices, therefore, Rashidi et al.’s (2012) approach does not incorporate any behavioural realism in the choice set formation stage because it is an approach that is asymptotically equivalent to the universal choice set approach and thus does not incorporate any behavioural aspects in the choice set formation stage. We have also illustrated this empirically in the next section.

In this study, we have applied both commute time and housing price distributions in order to compute the sampling weights. Since the probability of accepting price and accepting commute time are assumed to be independent, the joint probability distribution of accepting price and accepting commute time for a household is the product of the two probabilities. Hence, the sampling weight of an alternative for each household can be directly computed from this joint probability function as follows:

$$W_{j,n} = e^{\frac{|d_{j,n} - \hat{d}_n|}{\hat{d}_n}}$$  \hspace{1cm} (5-8)

where, $f_n(t_{j,n})$ is the probability of acceptable commute time and $f_n(p_n)$ is the probability of acceptable price.
Finally, the choice sets can be generated using these weights based on the importance sampling approach (without correction terms). In order to show empirically that the Rashidi et al. (2012) approach (i.e., importance sampling with correction terms) is equivalent to the universal approach, we have also implemented the importance sampling approach with alternative-specific correction terms. These are calculated similar to Frejinger et al. (2009) and Rashidi et al. (2012).

5.2.4.2 Deterministic Constraint Approach

As discussed earlier, the deterministic constraint approach is a commonly used approach in spatial choice models and is also equivalent to the anchor-points search strategy. In the deterministic constraint approach, the choice set for each decision maker is generated by restricting the choice set to include only the alternatives within a pre-specified threshold to the anchor-points, based on the analyst’s judgement. In destination choice, for example, the thresholds on the distances to trip origins (i.e., anchor-points) are used to generate the choice sets. These thresholds are conventionally determined by assuming that individuals consider only alternatives with distances less than 90 per cent of the observed commute distances to the individual’s home location (see Parsons and Hauber, 1998).

In order to derive households’ thresholds on commute times and housing prices (which are assumed to determine the searching behaviour) and to incorporate the intra-personal heterogeneity of households into the choice set formation stage, this study used the distribution of acceptable housing prices and work commute times by assuming that the thresholds are the values from the distributions corresponding to the 90th percentiles of those distributions. Since the choice of this percentile depends on the analyst’s judgement, the sensitivity of the parameter estimates to different threshold values is examined in section 5.4. Finally, the choice set for a household can be generated by filtering the universal choice set using the household-specific thresholds on commute times and housing prices generated from the screening model.

5.3 Empirical Analysis

There are some subtle issues to be taken into consideration in the comparison of different choice set formation approaches. On the one hand, if we assume that the
universal choice set is the true choice set, then any pruning of the universal choice set will potentially introduce bias. McFadden (1978) demonstrated that the MNL model produces consistent estimates when the choice set is generated using a simple random sampling approach and he also proposed a procedure for correcting the biases that result from non-simple random sampling methods such as importance sampling. On the other hand, if the true choice set is a pruned version of the universal choice set (where the pruning reflects the details of a spatial search process, as discussed in the previous section) then using the universal choice set, or randomly sampled subsets of it, will likewise introduce bias. Since the true choice set is unobserved, this issue is not directly resolvable empirically. The principal basis for assessing the alternative approaches to choice set generation, therefore, is in terms of their prediction performance on a hold-out validation sub-sample (although we do also report on estimated model parameters). The following five zonal-based residential location choice models, differentiated based on their choice set formation approaches, have been considered for empirical comparison:

1. **UCS (Universal Choice Set):** Households' choice sets are the universal choice set
2. **SRS (Simple Random Sampling):** Households' choice sets are randomly sampled (uniform sampling) from the universal choice set
3. **ISC (Importance Sampling with Correction terms):** Households' choice sets are generated by oversampling of alternatives based on the hazard-based screening model and adding alternative-specific correction terms in order to correct for bias introduced by importance sampling.
4. **ISNC (Importance Sampling without Correction terms):** Households' choice sets are generated by oversampling of alternatives based on the hazard-based screening model but **NOT** including the alternative-specific bias correction terms.
5. **DC (Deterministic Constraint):** Households' choice sets are generated using deterministic constraints on commute time and housing prices derived from the hazard-based screening model

**5.3.1 Data Preparation for the Residential Choice Model**

The empirical analysis of this study uses a specially constructed dataset of residential location choice, developed for the Greater London area, which draws on information from a number of separate data sources. The primary data source is the London Area Transport Survey (LATS), 2001. This survey provided essential data for understanding Londoners' travel, and informed the development of strategy and policies for transport in London (Transport for London, 2001). The LATS study area covered the area within
the M25 motorway. For the purposes of this study, however, the results are restricted to household residents of the Greater London area.

The estimation dataset used for this study is 12,836 cases from the LATS 2001 household file for which households’ employment locations are known. The spatial choice alternatives for residential location are LTS (London Transport Studies) model zones which are referred to as Transport Analysis Zones (TAZs). Households’ residential locations in LATS 2001 are determined by Easting/Northing coordinates; hence, they can be matched to any level of aggregation, such as TAZs. LATS 2001 also provides a long list of household characteristics including: annual household income, number of households’ members, number of children aged 1-5, number of employed, number of vehicles that households have access to, and the employment location of members of the household. Households’ socioeconomic attributes are included in the model using interaction variables. Figure 5-2 depicts the study area and the number of households by TAZs from LATS 2001.

![Figure 5-2- LATS 2001 Households' Spatial Distribution Zoning by Travel Analysis Zones (TAZs)](image-url)
In addition to the LATS data, several other data sources were used to obtain spatial and locational variables for the study area. The different locational variables that were used in the final model specification are described briefly here; Appendix C describes the complete set of spatial data sources and locational variables, as well as the data preparation stages. In the final model specification, the following variables have been considered:

- Log of zonal area\(^37\) (hectare)
- Number of residents per hectare
- Average household size
- Percentage of zonal area occupied by domestic buildings
- Zonal average housing price (1/1000)
- Accessibility to employment
- Absolute difference of household income and annualised rent in the zone\(^38\)
- Absolute difference of household size and average household size in the zone\(^39\)
- Travel time from households’ work zone(s) to candidate residential zones

Most of the locational variables were acquired from the 2001 census as discussed in chapter 4. The census variables incorporated into the final model specification include: number of residents per hectare, average household size, percentage of zonal area occupied by domestic buildings.

The housing price variable (acquired from the Land Registry) includes average paid prices for the year 2001 at the postal sector level (e.g., SW7 2). Unlike census variables, simple GIS aggregation cannot be applied to aggregated prices data at the TAZ level because the level of granularity of TAZs and postal sectors are more or less the same (i.e., 879 TAZs and 889 postal sectors). A model-based approach is adopted here, therefore, using the kriging interpolation to obtain average housing prices at the TAZ level (see Appendix C for a comprehensive discussion of this approach).

The computation of accessibility and commute variables were discussed in chapter 4. We also included interaction variables in the utility function in order to incorporate households’ observed taste heterogeneity in residential location choice decisions.

\(^37\) This is the size variable and is included in the model in order to avoid the level of aggregation sensitivity (see chapter 2).

\(^38\) This variable represents household’s disposable income (i.e., the amount of money that remains for living costs by choosing to live in a specific zone).

\(^39\) This variable indicates the clustering of households based on household size.
5.3.2 Data Preparation for the Hazard-based Screening Model

As mentioned earlier, two dependent variables were used to model the screening process of residential locations, i.e., average work commute time and average housing price. Average work commute time can be computed from LATS 2001 as aggregate values across all workers in the household. Average housing price data can also be computed from the Land Registry disaggregated dataset. The Kolmogorov-Smirnov test (K-S test) of different distributions confirms that the average work commute time follows a Weibull distribution, and average housing price follows a Log-logistic distribution (see Figure 5-3, Figure 5-4, and Table 5-1).

The explanatory variables of the hazard-based screening model include household socioeconomic variables since they are behaviourally considered as influential factors in determining the probability distribution of acceptable housing price and work commute time for a household. Many households’ socioeconomic variables that are available in LATS 2001 survey, such as household income, household size, number of workers in the household, number of children in the household, and households’ vehicle ownership, are used in the model specification of the screening model. Despite the fact that these variables are ordinal in nature, conventionally they are treated as continuous variables in the estimation of the statistical models (see Ben-Akiva and Lerman, 1985). The difference between introducing these variables as continuous and acknowledging their ordinal nature is effectively that the continuous variable imposes a linear relationship between the variable and the dependent variables, whereas the ordinal variable (introduced through dummies) allows for a non-linear relationship.

<table>
<thead>
<tr>
<th>Distribution</th>
<th>Average Work Commute Time</th>
<th>Average Housing Price</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>K-S Statistic</td>
<td>Rank</td>
</tr>
<tr>
<td>Exponential</td>
<td>0.0862</td>
<td>4</td>
</tr>
<tr>
<td>Logistic</td>
<td>0.1096</td>
<td>5</td>
</tr>
<tr>
<td><strong>Log-logistic</strong></td>
<td>0.0559</td>
<td>2</td>
</tr>
<tr>
<td>Lognormal</td>
<td>0.0823</td>
<td>3</td>
</tr>
<tr>
<td>Normal</td>
<td>0.1699</td>
<td>6</td>
</tr>
<tr>
<td><strong>Weibull</strong></td>
<td>0.0323</td>
<td>1</td>
</tr>
</tbody>
</table>
5.4 Model Estimation Results

5.4.1 Hazard-based Screening Model

As discussed earlier, the covariates of the hazard-based screening model are households’ socioeconomic variables. We estimated models with non-linear specifications for the number of vehicles and workers, and the results indicate that the non-linearity is not statistically significant. It should be noted that the effect of covariates in a hazard model is facilitated by incorporating a negative sign for the
parameters in the formulation. In other words, if a covariate gets a negative sign, the chance of failure or, equivalently, the probability of accepting a price or a work distance, is increased. The results of parameter estimation of the hazard-based screening model are tabulated in Table 5-2.

**Table 5-2- Hazard-based Screening Model Estimation Results**

<table>
<thead>
<tr>
<th>No</th>
<th>Parameter</th>
<th>Acceptable Housing Price (Log-logistic distribution)</th>
<th>Acceptable Commute Time to Work (Weibull distribution)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Estimate</td>
<td>STD</td>
</tr>
<tr>
<td>1</td>
<td>Theta (constant)</td>
<td>-26.7808</td>
<td>0.2032</td>
</tr>
<tr>
<td>2</td>
<td>Beta (shape parameter)</td>
<td>5.1389</td>
<td>0.0385</td>
</tr>
<tr>
<td>3</td>
<td>Number of household members</td>
<td>-0.2316</td>
<td>0.0126</td>
</tr>
<tr>
<td>4</td>
<td>Number of children age 1-5</td>
<td>0.1469</td>
<td>0.0353</td>
</tr>
<tr>
<td>5</td>
<td>Average household income (x1000)</td>
<td>0.0162</td>
<td>0.0007</td>
</tr>
</tbody>
</table>

**Summary statistics**

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Observations</td>
<td>12836</td>
</tr>
<tr>
<td>Log-likelihood at Zero Covariates (constant only): LL(C)</td>
<td>130139</td>
</tr>
<tr>
<td>Log-likelihood at Convergence: LL(β)</td>
<td>129485</td>
</tr>
<tr>
<td>Likelihood Ratio Test Statistic$^{40}$ ($\lambda^2$)</td>
<td>1308</td>
</tr>
</tbody>
</table>

The negative sign for the number of household members in the model suggests that in the searching process large households are looking for less expensive areas. This can be interpreted as reflecting the fact that large households need more space. Therefore, all else being equal, they tend to obtain more housing space for their limited budget in less expensive neighbourhoods. On the other hand, the positive sign of number of children aged 1-5 suggests that households with small children prefer more expensive neighbourhoods, which could be a proxy for safer neighbourhoods. Furthermore, the positive sign of the average household income shows the obvious fact that wealthier households tend to live in more expensive neighbourhoods.

$^{40}$ $\lambda^2 = -2[\text{LL(C)} - \text{LL(β)}]$, this test is used for evaluating the overall significance of the model. This statistic is asymptotically Chi-squared distributed with 6 degree of freedom which is highly significant.
Household income, which is positively correlated with the number of vehicles, is also an important factor in the household decision about the average work commute time. The positive signs of household income and number of vehicles show that the wealthier households are also more likely to live in suburban areas and commute farther to their workplaces. On the other hand, the negative sign on number of workers suggests that households with more workers tend to live closer to their workplaces. Since the total travel cost of households with more workers is more than households with fewer workers, we can argue that households with more workers prefer the overall travel burden to be less, which can explain the negative sign of the number of workers variable.

5.4.2 Residential Location Choice Model

The locational variables, the commute variable and interaction variables introduced in the previous section were identified from the literature as likely determinants of a household’s choice of residential location. Several specifications of the residential choice model were tested and statistically insignificant variables were systematically eliminated in order to find the preferred model specification for residential location choice in Greater London.

As discussed earlier, the expected choice sets can be generated for each household conditional on the price and distance thresholds of each household derived from the hazard-based screening model. Assuming that the thresholds correspond to the 90th percentile of acceptable commute time and acceptable price distributions, the mean size of the expected choice sets is 389 (the universal choice set contains 861 alternatives). This indicates that applying the deterministic constraints derived from the housing screening model to generate the choice sets can reduce the size of the choice set by over 50% (see Table 5-3 for summary statistics of expected choice sets).

The estimation results of five residential choice models (as described in section 5.2) are presented in Table 5-4. These are the results of the model estimation using the full dataset (100 per cent). In general, the coefficient of the log of the zonal area has a positive sign, as expected, indicating that households are more likely to locate themselves in zones with a large number of housing opportunities. This parameter is significant in all the models. The coefficient of population density (number of residents
per hectare) is also positive and significant in all models. The zonal average household size and its interaction with household size are negative and significant in all models. This confirms the clustering of households based on the zonal household size that has also been observed in previous studies.

<table>
<thead>
<tr>
<th>Statistics</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min of Choice Set Sizes</td>
<td>1</td>
</tr>
<tr>
<td>Max of Choice Set Sizes</td>
<td>756</td>
</tr>
<tr>
<td>Mean of Choice Set Sizes</td>
<td>388.75</td>
</tr>
<tr>
<td>STD of Choice Set Sizes</td>
<td>149.89</td>
</tr>
</tbody>
</table>

Intuitively, accessibility to employment has a positive effect on the utility but the coefficient of the accessibility to employment variable is found to be negative here. Guo and Bhat (2004a) also found a negative utility of accessibility to work for African-American households in the Dallas county area. The residential choice models estimated in this study likewise suggest that most people in London live in neighbourhoods that are less accessible to employment opportunities. This is a reflection of the housing market in London, the most desirable neighbourhoods typically being located far from the major employment centres. For example, places like the City of London and Heathrow airport which have very high accessibility to employment are not very desirable residential areas. We also believe that people have very heterogeneous taste regarding to the accessibility of a location. In the case of accessibility to employment, for example, characteristics of households’ members such as their employment sector, their age, etc. are very important and different households might have very different tastes regarding the accessibility of a location. Hence, formulating accessibility using a simple Hansen-type measure would not reflect the attractiveness of a location very well. Further research is required to explore this issue in the context of residential location choice modelling.
Table 5-4 - Residential Location Choice Models Estimation Results

<table>
<thead>
<tr>
<th>No</th>
<th>Parameter</th>
<th>1. UCS</th>
<th>2. SRS</th>
<th>3. ISC</th>
<th>4. ISNC</th>
<th>5. DC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Estimate</td>
<td>t-stat</td>
<td>Estimate</td>
<td>t-stat</td>
<td>Estimate</td>
</tr>
<tr>
<td>1</td>
<td>Log of zonal area (hectare)</td>
<td>2.3614</td>
<td>49.2990</td>
<td>2.3935</td>
<td>47.4931</td>
<td>2.4674</td>
</tr>
<tr>
<td>2</td>
<td>Number of residents per hectare</td>
<td>0.0098</td>
<td>29.3184</td>
<td>0.0099</td>
<td>28.6386</td>
<td>0.0102</td>
</tr>
<tr>
<td>3</td>
<td>Average household size</td>
<td>-0.3366</td>
<td>-8.2541</td>
<td>-0.3354</td>
<td>-7.8050</td>
<td>-0.3174</td>
</tr>
<tr>
<td>4</td>
<td>Percentage of zonal area occupied by domestic buildings</td>
<td>0.8271</td>
<td>5.9886</td>
<td>0.7385</td>
<td>5.1373</td>
<td>0.6403</td>
</tr>
<tr>
<td>5</td>
<td>Zonal average housing price (1/1000)</td>
<td>-0.0028</td>
<td>-16.7385</td>
<td>-0.0028</td>
<td>-16.1052</td>
<td>-0.0026</td>
</tr>
<tr>
<td>6</td>
<td>Accessibility to employment</td>
<td>-0.0141</td>
<td>-23.3048</td>
<td>-0.0121</td>
<td>-19.9343</td>
<td>-0.0127</td>
</tr>
<tr>
<td>7</td>
<td>Absolute difference of household income and annualized rent</td>
<td>-0.0046</td>
<td>-2.0255</td>
<td>-0.0056</td>
<td>-2.4317</td>
<td>-0.0034</td>
</tr>
<tr>
<td>8</td>
<td>Absolute difference of household size and average household size</td>
<td>-0.6798</td>
<td>-18.2415</td>
<td>-0.6939</td>
<td>-17.7371</td>
<td>-0.6941</td>
</tr>
<tr>
<td>9</td>
<td>Travel time from households’ workplace zone(s) to residential zones</td>
<td>-0.0598</td>
<td>-118.9103</td>
<td>-0.0575</td>
<td>-113.7098</td>
<td>-0.0602</td>
</tr>
</tbody>
</table>

Summary statistics

<table>
<thead>
<tr>
<th></th>
<th>12836</th>
<th>12836</th>
<th>12836</th>
<th>12836</th>
<th>12836</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Observations</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sample Size</td>
<td>861</td>
<td>50</td>
<td>50</td>
<td>50</td>
<td>50</td>
</tr>
<tr>
<td>Log-likelihood at Zero: LL(0)</td>
<td>86740.1</td>
<td>50210.8</td>
<td>55117.5</td>
<td>49679.1</td>
<td>50126.2</td>
</tr>
<tr>
<td>Log-likelihood at Convergence: LL(β)</td>
<td>73159.9</td>
<td>37401.08</td>
<td>42315.3</td>
<td>39810.5</td>
<td>42366.9</td>
</tr>
<tr>
<td>Likelihood Ratio Index (ρ²)</td>
<td>0.157</td>
<td>0.255</td>
<td>0.232</td>
<td>0.199</td>
<td>0.155</td>
</tr>
</tbody>
</table>
The commute-related variables are important determinants of residential choice. Households, in general locate to reduce their commute time. Therefore, commute time would be expected to have a negative sign which is in fact the case in all the models. Zonal housing price and its interaction with household income would also be expected to have a negative effect on the utility of a residential choice alternative, as has also been addressed in previous studies. This parameter was indeed found to be negative in UCS, SRS and ISC. But, interestingly, the zonal housing price parameter and its interaction with households’ average income are positive in DC and ISNC. This result is unexpected and initially somewhat counter-intuitive. It may, however, reflect a selectivity effect in the operation of the screening model. In particular, it could be argued that the use of a price threshold in the screening model has the effect of essentially filtering unaffordable locations out of the choice set, leaving behind those that are affordable. Amongst these affordable locations, price is likely to be positively correlated with unobserved quality attributes, implying that households in effect choose the best location that they can afford amongst the screened alternatives. This possibility of the existence of positive price gradients has also previously been examined by Richardson (1977b).

5.4.3 Sensitivity Analysis of Parameter Estimates

In the estimation of the DC model we assumed that an alternative is considered by a household if, and only if, the average housing price of that alternative is less than the household’s price threshold and the commute time to the household’s permanent workplace is less than the household’s commute time threshold. Further we assumed that the household’s price and commute time thresholds are the values from probability distributions of acceptable housing price and acceptable commute time corresponding to the probability of 0.9 (90th percentile). As mentioned earlier, the choice of this percentile depends on the analyst’s judgement. In this section, we examine the sensitivity of the parameter estimates of the residential location choice model to different threshold values.

With the exception of commute time and housing price coefficients (as well as interaction of housing price with annualised rent), all the coefficients in the residential location choice model show little change as the threshold expands. The housing price
and commute time coefficients, however, change dramatically as the value of the percentile (and therefore housing price and commute time thresholds) changes. Figure 5-5 and Figure 5-6 illustrate this for the housing price and the commute time parameter.

![Figure 5-5: Sensitivity of Housing Price Parameter and Rho Square to the Choice of Percentile](image)

The commute time coefficient has the expected negative sign in all the models, however this coefficient becomes more negative (implying a greater disutility attributed to compute time) when the boundary expands. This shows that when the effect of commute time is modelled as a non-compensatory process in the choice set formation stage (by assuming commute time thresholds), the commute time coefficient is estimated to have less impact in the compensatory stage when compared to a choice model that considers the universal choice set. In other words, incorporating a constraint on the commute time (non-compensatory process) in the choice set formation stage reduces the effect of commute time in the compensatory stage and vice versa.
In our empirical setting, as illustrated in Figure 5-5 and Figure 5-6, the rho square (goodness of fit) of the models increases as the boundary expands for both commute time and price thresholds. This effectively means that a model with the universal choice set is a better fit for the data in terms of the value of likelihood ratio index. However, Elgar et al. (2009) interpreted the drop in the goodness of fit in the model with the expected choice sets differently and claimed that drop of goodness of fit in the model with the expected choice sets (i.e., DC and ISNC) indicates that much of the location choice behaviour was captured in the choice set generation stage. This will be discussed in more detail in the next section.

5.5 Model Validations and Discussion

The model estimation results indicate that whether one assumes the universal choice set (UCS) for households or whether one applies sampling strategies, i.e., uniform random sampling (SRS) and importance sampling with correction (ISC), the outcomes are equivalent. This agrees with the theory, which states that sampling of alternatives is
a statistical solution to cope with large numbers of alternatives and is asymptotically equivalent to using the universal choice set. In fact, UCS, SRS and ISC represent the same assumed behaviour, which is that the universal choice set is the actual choice set considered by households.

On the other hand, the deterministic constraint (DC) approach and the importance sampling approach without correction (ISNC) attempt to incorporate the underlying search process of residential location choices based on simplifying assumptions regarding the spatial search behaviour of households (e.g., an anchor-based search strategy), and households’ characteristics (based on the hazard-based screening model). As noted earlier, the DC model and the ISNC model are in fact equivalent as the deterministic approach can be viewed as an importance sampling protocol where alternatives are assigned a weight of zero if they are outside the thresholds and a weight of 1 if they are inside the thresholds. The similar estimated parameters of the DC model and the ISNC model also show that the deterministic constraint approach and the importance sampling without correction are equivalent.

The SRS model has the highest goodness of fit ($\rho^2=0.255$) which is also close to the goodness of fit of the ISC model ($\rho^2=0.232$). The goodness of fit of ISNC and DC dropped by 22% and 39% respectively from the highest goodness of fit.\footnote{We did not compare the goodness of fit models against the UCS model as the goodness of fit in the UCS model has been computed over a much larger choice set than the other models which have similar choice set sizes.} Models with the expected choice sets have poorer goodness of fit, therefore. From this result, we can conclude that the SRS model (which is asymptotically equivalent to the universal choice set approach) is the superior model and any attempts to incorporate the underlying search process and the behavioural realism of the choice set formation process reduces the goodness of fit of the residential location choice model. As discussed earlier, Elgar et al. (2009) claimed that a drop in the goodness of fit in the model with the expected choice sets indicates that much of the location choice behaviour was captured in the choice set generation stage. They argued that the higher goodness of fit of the SRS model, compared to the model with the expected choice sets, does not necessarily show the strong explanatory ability of the SRS model since most of the randomly selected alternatives are very poor substitutes to the chosen alternative in the SRS model. From this perspective, therefore, the lower goodness of fit in models with the expected choice
sets implies that a smaller part of the variability in location choice decisions were captured by explanatory variables as the unfeasible alternatives were filtered out from the choice sets in the choice set formation stage. Hence, comparison of different choice set generation approaches based on the goodness of fit criteria is debatable since households’ true choice sets are usually not observable. Consequently, the principal basis for assessing the alternative approaches to choice set generation in this chapter is founded on the hold-out validation approach.

In the validation of the models, model estimations were achieved using a randomly drawn 75% of the data with the remaining 25% being kept aside as a hold-out sample. This 25% hold-out sample was used to compare the performance of the choice set formation models. The performance of the models was compared using different measures of disaggregate and aggregate validation as reported in Table 5-5.

<table>
<thead>
<tr>
<th>Statistics</th>
<th>UCS</th>
<th>SRS</th>
<th>ISC</th>
<th>ISNC</th>
<th>DC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Households (Validation Sample)</td>
<td>3209</td>
<td>3209</td>
<td>3209</td>
<td>3209</td>
<td>3209</td>
</tr>
<tr>
<td>Chosen Alternatives Truly Included in the Choice Sets¹</td>
<td>3209</td>
<td>188.7</td>
<td>194.8</td>
<td>196.6</td>
<td>363.5</td>
</tr>
<tr>
<td>Percent Truly Included (PTI)²</td>
<td>100%</td>
<td>5.88%</td>
<td>6.07%</td>
<td>6.13%</td>
<td>11.3%</td>
</tr>
<tr>
<td>Sample Size (Choice Set Size)</td>
<td>861</td>
<td>50</td>
<td>50</td>
<td>50</td>
<td>50</td>
</tr>
<tr>
<td>Choice Set Size/Universal Choice Set (Percentage)</td>
<td>100%</td>
<td>6%</td>
<td>6%</td>
<td>6%</td>
<td>6%</td>
</tr>
<tr>
<td>Percent Correctly Predicted (PCP)³</td>
<td>4.43%</td>
<td>1.58%</td>
<td>1.12%</td>
<td>1.19%</td>
<td>1.67%</td>
</tr>
<tr>
<td>Average Probability of Correct Prediction (APCP)⁴</td>
<td>0.0078</td>
<td>0.0066</td>
<td>0.0061</td>
<td>0.0054</td>
<td>0.0064</td>
</tr>
<tr>
<td>RMSE ⁵</td>
<td>2.9</td>
<td>2.94</td>
<td>3.13</td>
<td>3.73</td>
<td>3.82</td>
</tr>
<tr>
<td>MADE ⁶</td>
<td>2.13</td>
<td>2.16</td>
<td>2.29</td>
<td>2.63</td>
<td>2.67</td>
</tr>
</tbody>
</table>

¹The results of this table are the average of 10 runs for SRS, ISC, ISNC and DC.
²PTI represents the percent of households (in the validation sample) in which the chosen alternatives included in the generated choice sets.
³PCP represents the percent of households (in the validation sample) in which the chosen alternatives correspond to the highest probability alternative.
⁴APCP is computed as: \( N^{-1} \sum \sum y_{ni} p_{ni} \), where \( N \) is the number of households in the validation sample and \( y_{ni} \) takes the value 1 if household \( n \) chose alternative \( i \), and 0 otherwise.
⁵RMSE represents the Root Mean Square Error between predicted and actual shares.
⁶MADE represents the Mean Absolute Deviation Error between predicted and actual shares.

It should be noted that the chosen alternatives are not included in the households’ choice sets in the choice set formation stage in order to test the actual prediction performance of different choice set formation approaches. The Percent Truly Included (PTI) measure is calculated to assess the performance of each choice set formation approach in including the chosen alternatives in the generated choice sets. As the PTI
measure suggests the improvement in capturing the chosen alternative in the importance sampling approach is marginal compared to the simple random sampling of alternatives. The improvement in capturing the chosen alternative in the deterministic constraint approach, however, is significant.

All prediction test measures, including: RMSE and MADE statistics, Percent Correctly Predicted (PCP) and Average Probability of Correction Prediction (APCP), suggest that UCS is the superior approach. Hence, the validation results seem to be in favour of the universal choice set approach in the predication of residential location choices. It is also interesting to note that the DC model has the poorest performance based on prediction test measures although the Percent Truly Included (PTI) measure shows a significant improvement compared to the SRS model. This results show that any systematic mis-specification in the choice set formation (i.e., ISNC and DC models) will be propagated into the resulting choice model and reflected in a degraded prediction performance (as well as impacting on the estimation of key model parameters, such as the importance of housing prices).

The RMSE and MADE statistics, as well as the Percent Correctly Predicted (PCP) and the Average Probability of Correct Prediction (APCP) criterion, have also been computed for different percentile values (thresholds) of the DC model, as described in the previous section. The prediction results confirm that as the choice set boundary expands the RMSE and MADE decreases and the PCP and APCP increases, which essentially shows that the model with the universal choice set is the superior model.

5.6 Conclusion

In this chapter we have compared the empirical performance of alternative exogenous choice set formation approaches in London's residential location choice context with more conventional statistical methods of choice set pruning (i.e., simple random sampling of alternatives). While a comparison of the model parameters of the residential choice model is a key element in identifying the preferred choice set formation model, it is also important to compare models for goodness of fit and prediction on a hold-out validation sub-sample. The principal basis for the comparison in this study, therefore, is in terms of their prediction performances on a hold-out validation sub-sample. Despite its intuitive appeal and behavioural plausibility, the
results of aggregate and disaggregate validation of the model suggest that the models with the expected choice sets resulting from the anchor-based search strategy and the hazard-based screening model in fact perform worse than the universal choice set approach in the residential location choice context.

The finding of this study emphasises both the importance of modelling choice set formation and the high level of challenge involved in doing so effectively in the context of residential location choice. It must be remembered that any choice set formation model is attempting to characterise what is an underlying highly complex and dynamic process of housing market search, which will depend on the spatial dynamics of labour markets and housing supply, market facilitators (e.g., real estate agents, mortgage lenders) activity, government policies (e.g., central bank interest rate) as well as many perceptual and cognitive factors. The empirical results of this chapter indicate that generating choice sets based on the deterministic constraint approach or the importance sampling approach (without correction) is unlikely to capture well this complex set of influences.

The empirical residential location model developed in this study can be improved by incorporating other determining variables in location choices, such as the locational crime rate. It should be noted that incorporating new variables will result in a better behavioural residential location choice model and is an important step towards making the model practice-ready, however early results and analyses indicate that improving the specification of the residential choice model does not change the findings of this research in terms of the comparative performance of the choice set formation approaches. Other future research tasks include testing the robustness of the findings with better functional forms and in other empirical contexts.
CHAPTER 6

SIMPLIFIED PROBABILISTIC CHOICE SET FORMATION MODELS IN A RESIDENTIAL LOCATION CHOICE CONTEXT

6.1 Overview

The feasibility and practicality of applying deterministic choice set formation approaches for capturing the underlying search process of households in a residential location choice context has been investigated in chapter 5. There are some uncertainties inherent in the choice set formation process due to the limited knowledge of the underlying search process and heterogeneity among decision makers regarding the search strategies. This chapter investigates the uncertainty of choice sets in residential location choice modelling and proposes a simplified probabilistic choice set formation approach to model choice sets and choices simultaneously.

The choice context with uncertain choice sets can be modelled using the two-stage discrete-choice modelling paradigm introduced by Manski (1977). The general Manski model considers all potential choice sets and requires summation over the power set \( G \) (i.e., choice set space) which is a set of all non-empty subsets of the universal choice set (i.e., a set of all available alternatives). The size of the power set \( G \), however, increases exponentially with the increase in the number of alternatives. Therefore, implementation of the Manski model is impractical when the number of alternatives is large, which is a typical case in most spatial choice contexts.

As mentioned in chapter 2, some authors have tried to overcome the computation burden of the Manski approach by imposing a priori restrictions on the composition of choice sets based on some exogenous evidence from the choice set formation process. In the context of residential location choice, Kaplan, Bekhor and Shiftan (2009, 2011, 2012) (KBS, hereafter) developed a probabilistic choice set model incorporating data of
individuals searching for dwellings observed using a customised real estate agency website. This secondary data was used to compute the probability of considering a choice set that takes the form of an Ordered Probit model. In this chapter, we illustrate that the simplicity of the KBS model arises because of an unrealistic assumption that individuals’ choice sets only contain alternatives that derive from their observed combination of thresholds. By relaxing this assumption we introduce a new probabilistic choice set formation model that allows the choice set space to include all potential choice sets derived from variations in the combinations of thresholds.

In addition to extending the KBS model, our proposed model asymptotically approaches the classical Manski model, provided a suitable structure is used to categorise alternatives. The proposed model can also be viewed as a variation of Swait's (2001) GenL model where, instead of latent grouping of alternatives, alternatives are categorised based on observed thresholds. Similar to the GenL model, the proposed model assumes that the choice set space is part of the model specification. In order to illustrate the biases inherent in the original KBS approach, we compare it with our proposed model and the MNL model using a Monte Carlo experiment. The results of this experiment show that the KBS model causes biases in predicted market share if individuals are free to choose from any potential choice sets derived from combinations of thresholds.

**6.2 Simplified Probabilistic Choice Set Formation Models**

Experimental research suggests that in a complex choice situation (e.g., choice among a large number of alternatives) decision makers adopt non-compensatory screening strategies (e.g., elimination-by-aspects, see Tversky, 1972) to reduce the number of alternatives to a smaller number before using a compensatory decision rule to make a final decision (see Manrai and Andrews, 1998). This has led to the view that decisions may be made in two stages: (a) non-compensatory stage (b) compensatory stage. This two-stage decision making paradigm can be modelled based on Manski’s (1977) formulation as follows:

\[
P_{i,n} = \sum_{C_n \in G} P(i|C_n) \cdot Q_n(C_n)
\]

where,
\( P_{i,n} \) is the unconditional probability of alternative \( i \) being chosen,

\( P(i|C_n) \) is the conditional probability of alternative \( i \) being chosen given choice set \( C_n \),

\( Q_n(C_n) \) is the probability of individual \( n \)'s choice set being \( C_n \), and

\( G \) is a set of all non-empty subsets of the universal choice set \( M \).

The major drawback of the Manski model is that it is computationally intensive since the number of theoretically possible choice sets increases exponentially as the size of \( M \) increases (i.e., \(|G| = 2^{|M|} - 1\)). This approach is therefore impractical for spatial choice models, which are characterised by large number of alternatives.

Some authors have tried to overcome the computation burden of the Manski model, yet still accommodate the probabilistic choice set formation models, by restricting the composition of set \( G \) based on some simplifying assumptions regarding the behavioural realism of the choice set formation process. By restricting \( G \), we in fact assume that an aggregated set of alternatives can be in or out of the choice set together.

Assuming that an individual either chooses from the choice set that contains the feasible/available alternatives (considering the constraints of the individual) or is free to choose from the universal choice set, Swait and Ben-Akiva (1987) proposed a model with the restricted \( G \) as:

\[
P_{i,n} = K \cdot [ P(i|M_{n,r(i)}) \cdot Q_n(M_{n,r(i)}) + P(i|M_n) \cdot Q_n(M_n) ]
\]

\[
G_n = \{ M_{n,1}, ..., M_{n,r}, M_n \}
\]

where, \( G_n \) is the restricted choice set space and \( K \) is a normalisation constant to account for the restrictions imposed on the composition of the choice sets.

In this approach, the universal choice set is subdivided into \( R \) non-empty, mutually exclusive and collectively exhaustive subsets \( M_{n,r} \) (i.e., \( M_n = \bigcup_{r=1}^{R} M_{n,r} \) and \( M_{n,r} \cap M_{n,r'} = \emptyset, \forall r \neq r' \)). Hence, an individual is either free to choose from the universal choice set \( M_n \) or he is limited to one of its subsets which meets the individual's constraints. The captivity model (Dogit model) developed by Gaudry and Dagenais (1979) can be seen as a special case of this model. In the Dogit model, individuals are either captive to an alternative (e.g., low-income workers may be captive to public transportation) or free to
choose among all the alternatives, which can be interpreted as special case of the model presented in Equation 6-2 if we assume that $|M_{n,r}| = 1, r = 1..R$. The model proposed by Siddarth et al. (1995) based on secondary data on consumer purchase history is also equivalent to Swait and Ben-Akiva’s (1987) model with restricted $G$ as described in Equation 6-2. The GenL model proposed by Swait (2001) is also developed by restricting the choice set space based on some exogenous evidence from the choice process, even though the behavioural motivation of the GenL model is different from the two-stage models.

In the marketing literature, Andrews and Manrai (1998a) developed a feature-based elimination model that assumed that decision makers apply a sequence of non-compensatory rules to form their choice sets based on the features of the alternatives. In this model, alternatives in the universal choice set are categorised based on their features in a hierarchical structure. This model can also be considered as a simplified two-stage model because the $G$ set is restricted to include the choice sets matching the hierarchical structure, not all of the subsets of the universal choice set.

In the destination choice context, Zheng and Guo (2008) developed a random constraint model for destination choice with 27 alternatives. The restricted set $G$ is constructed by assuming that a destination alternative is included in an individual’s potential choice set if, and only if, all other alternatives that are closer to the individual’s trip origin are also included in the choice set. By restricting the composition of set $G$ based on the abovementioned spatial contiguity assumption, Zheng and Guo reduced the total number of possible choice sets from $2^{|M|} - 1$ to $L \leq |M|$ for each trip origin.

Hicks and Schnier (2010) developed a simplified probabilistic choice set formation model using different macro-definitions of spatial regions in order to focus on the micro-level spatial decision making and to investigate the sensitivity of the results to alternative macro-level spatial definitions. Again, they demonstrated that by assuming some structure on the composition of choice sets (i.e., restricting $G$); the dimensionality problems associated with general Manski model can be reduced.

In a residential location choice context, Kaplan, Bekhor and Shiftan (2009; 2011; 2012) developed a probabilistic choice set formation model incorporating data of individuals searching for dwellings observed using a real estate agency website. The KBS model
uses observed choice set data as well as chosen alternative data in order to reduce the computational complexity of probabilistic choice set models. Estimation of the KBS model, therefore, requires observation of individual’s thresholds, such as number of bedrooms and price thresholds, etc. as discrete categorical variables during the search process.

The KBS model assumes that the choice sets derived by observed individuals’ thresholds is the “true choice set”. Kaplan et al. (2009) argued that by observing the “true choice set” using a customised real estate agency website, the unconditional choice probability of an alternative does not require enumeration over all potential choice sets (unlike the Manski model) and can be written as:

\[ P_{i,n} = P(i|C_n).Q_n(C_n) \]  \hspace{1cm} (6-4)

where,

\[ P(i|C_n) \] is the conditional probability of alternative \( i \) being chosen given the “true choice set” \( C_n \),

\[ Q_n(C_n) \] is the probability that the “true choice set” is \( C_n \).

In the original KBS model the conditional probability is assumed to have a MNL structural form. It also assumes that the probability of selecting the choice set \( C_n \) is equal to the probability of selecting a combination of thresholds. Further assuming that choices of different thresholds are independent, the following choice set probability is proposed:

\[ Q_n(C_n) = P(t_{1,n}^* \cap t_{2,n}^* \cap ... \cap t_{K,n}^*) = P(t_{1,n}^*) \times P(t_{2,n}^*) \times ... \times P(t_{K,n}^*) \]  \hspace{1cm} (6-5)

As mentioned earlier, in the KBS model individuals’ thresholds on different criteria, such as number of beds, prices, etc., as well as their characteristics such as income, age, etc., are observed during the search-by-criteria stage using a real estate agency website. Since the thresholds are categorical variables (pre-determined values) and are naturally ordered, the probability of individual \( n \) to select the thresholds \( t^* \) of the \( k \)th criterion (i.e., \( P(t_{k,n}^*) \)) is modelled using an Ordered Probit formulation in the KBS model. Assuming that the probability of selecting a threshold is related to characteristics of an
individual, the dependent variable $t^*_{k,n}$ is characterised as a function of individual's characteristics and an error term as:

$$t^*_{k,n} = \alpha_k Z_{k,n} + \varepsilon_{k,n}$$  \hspace{1cm} (6-6)

where, $Z_{k,n}$ is the vector of explanatory variables (i.e., individual's characteristics), $\alpha_k$ is the model parameters and $\varepsilon_{k,n}$ is the normal error term.

The further extensions of the KBS model also incorporate the correlations among different criteria in the non-compensatory choice set formation stage using the Multinomial Ordered Probit model developed by Bhat and Srinivasan (2005) as well as incorporating a flexible error structure in the compensatory choice stage based on Mixed Logit formulation (see Kaplan et al., 2012). The overall approach in all of these extensions is the same, however, and the unconditional choice probability is still represented by Equation 6-4. The non-compensatory choice set formation model based on the Multinomial Ordered Probit formulation and the compensatory choice model based on the Multinomial Logit or the Mixed Logit formulation are estimated jointly using a maximum likelihood approach in the KBS model.

### 6.3 The Shortcomings of the KBS Model

Here, we argue that the simplicity of the KBS model arises not because of observing the thresholds and adding more information, rather by restricting $G$ to include only specific choice sets similar to the other models presented earlier.

In the KBS model, it is assumed that each combination of thresholds corresponds to a choice set. This effectively means that in this approach the choice set space $G$ is limited to include $R$ (i.e., $R = l_1 \times l_2 \times \ldots l_k$) mutually exclusive and collectively exhaustive subsets of the universal choice set corresponding to each combination of thresholds ($l_k$ is the number of categories for $k$th criterion). Mathematically speaking, the choice set space $G$ corresponding to different combinations of thresholds ($R$) is given as:

$$G = \{ C_1, C_2, \ldots, C_R \}$$  \hspace{1cm} (6-7)

where, $C_r$ is the choice set corresponding to $r$th combination of thresholds. Since $P(i|C_n) = 0$ if $i \notin C_n$, the general Manski model with this restricted $G$ reduces to Equation 6-4 in the KBS model.
It should be noted that a two-stage model is probabilistic when the composition of the choice set is not known with certainty and non-zero occurrence probabilities are attached to two or more sets (Manrai and Andrews, 1998). With this in mind, the KBS model cannot be considered as a probabilistic model, because the non-zero occurrence probabilities are only attached to one set (the observed choice set). This is an important drawback of the KBS model because the summation of choice probabilities would not equal to one. Moreover, in the KBS model, it is assumed that by observing the combination of thresholds for each individual during the search process, we are in fact observing the “true choice set” of that individual. It should be borne in mind that observing the choice sets with certainty is impossible (Shocker et al., 1991). All points considered, we believe that the KBS model is in fact a deterministic model and it should not be considered as a probabilistic choice set formation model.

It should be also noted that the joint estimation of the choice set formation stage and the choice stage, as suggested by Kaplan et al. (2009), is not necessary since the probability of choosing an alternative given a choice set and the probability of selecting a combination of thresholds are independent in the KBS model.

In summary, there are both methodological and conceptual shortcomings in the KBS model. From the conceptual perspective, the assumption that the “true choice set” can be observed by observing individuals’ thresholds is not empirically supported. From the methodological perspective, meanwhile, the summation of choice probabilities over the alternatives in the KBS model presented in Equation 6-4 does not add up to one which results in incorrect predictions of the market shares when the model is applied. The KBS model is interesting, however, in that it introduces an innovative approach to estimate choice models based on observed individuals’ thresholds on different criteria in the web-based store environment.

6.4 The Proposed Model

Punj and Moore (2009) proposed a conceptual model of information search and choice set formation in a web-based store environment based on constructs that are known to affect consumer behaviour in online settings such as elimination-type screening strategies (e.g., thresholds on different attributes), electronic decision aids, etc. They conducted a survey of 120 undergraduate students choosing an apartment to rent near
a hypothetical university in order to test different hypothesis about student's searching behavior and choice set formation in an online setting. They concluded that (1) when more time is available, the number of search iterations conducted increases, as well as the number of alternatives in the individuals' choice sets, (2) when many alternatives are available consumers conduct fewer search iterations but they do not actually examine fewer alternatives (they have a larger choice set), and (3) consumers form larger choice sets when many alternatives are available but form smaller choice sets when more time is available.

Thresholds for different criteria in real estate websites, as used in the KBS model, therefore, are actually tools to simplify the information search process and it would be behaviourally unrealistic to assume that an individual's choice set can only contain alternatives corresponding to a combination of thresholds. Here, we assume that the choice sets can be the union of two or more sets of alternatives derived from a combination of thresholds and propose the following model:

\[
P_{i,n} = \sum_{C_n \in G} P(i|C_n) Q_n(C_n) \tag{6-8}
\]

\[
G = \{ C_1, C_2, \ldots, C_{R}, \{C_1, C_2\}, \{C_1, C_3\}, \ldots, \{C_1, C_2, C_3\}, \ldots, \{C_1, C_2, \ldots, C_R\} \} \tag{6-9}
\]

where,

\( C_1, C_2, \ldots, C_R \) are an aggregated set of alternatives based on the combination of thresholds,

\( R = I_1 \times I_2 \times \ldots \times I_K \) is the total combination of \( K \) attributes' thresholds,

\( I_1, I_2, \ldots, I_K \) are the number of ordered categories in each attribute's threshold.

Each criterion in the choice set formation stage of the KBS model (and our proposed model) is represented by different threshold values as discrete categorical variables (e.g., three price categories and three categories). The universal choice set is partitioned by different combinations of thresholds of different criteria (e.g., \( R = 3 \times 3 = 9 \)). Each of these partitions is corresponded to an aggregated set of alternatives. The original KBS model assumes that the choice set space can only include the choice sets corresponding to these partitions. In our proposed model, however, the choice set space includes different variations of these partitions, as depicted in Figure 6-1.
Similar to Hicks and Schnier (2010), in our proposed model, alternatives are categorised to larger sets and the choice set space is constructed based on these sets of aggregated alternatives. The aggregated set of alternatives, similar to the KBS model, is derived from a combination of thresholds on different attributes implemented as searching tools in real estate agency websites.

The size of set $G$ grows exponentially as the number of thresholds and number of criteria increases (i.e., $|G| = 2^R - 1$). Hence, in addition to extending the KBS model, our proposed model asymptotically approaches the classical Manski model when the universal choice set is partitioned in such a way that each alternative belongs to a combination of thresholds. It should be noted that the computation burden of this model is much less than the original Manski (1977) model since alternatives are

\[
\sum_{G \in G} \prod_{i\in G} \binom{P_i}{Q_i} \cdot Q_i \cdot (C_{n})
\]
grouped into $R$ categories and the choice set space thus grows exponentially with the number of categories rather than the number of alternatives, as in the original Manski model.

The proposed model assumes that an individual choice set can include the alternatives corresponding to one or more combination of thresholds (i.e., one or more categories) and cannot include, for example, half of the alternatives from one category and another half from another category. Hence, both the KBS model and our proposed model assume some predetermined restrictions on the composition of the choice set space, and the simplification, compared to original Manski model, happens because of these restrictions. This is also analogous to the model developed by Hicks and Schnier (2010), except that, in Hicks and Schnier’s (2010) model, observed spatial regulations are used to categorise the alternatives rather than observed thresholds. We observe the combination of thresholds on different attributes as well as the decision maker’s socioeconomics during the information search process, in a similar way to the KBS model. The probabilities of selecting aggregated sets of alternatives, therefore, similar to the KBS model, can be modelled as a Multinomial Ordered Probit model:

$$P_n(C_r) = P(t_{1,r}^*) \times P(t_{2,r}^*) \times \ldots \times P(t_{k,r}^*)$$

(6-10)

$$\forall C_r \in C = \{ C_1, C_2, \ldots, C_R \}$$

(6-11)

The probability of selecting an aggregated set of alternatives should be updated during the searching stage. Hence, the more general model for probability of selecting aggregated sets of alternatives for $J$ iterations takes the following form:

$$P_n(C_r) = f(P_n(C_{r_1}, P_n(C_{r_2}), \ldots, P_n(C_{r_J}))$$

(6-12)

The form of function $f$ is not known and further empirical investigation is required to identify the form of function $f$ in this general model. In the absence of this empirical work, the probability of an aggregated set of alternatives is calculated based on the selected thresholds of the final search iteration, again similar to the KBS approach. In order to compute choice set probability based on probabilities of selecting aggregated sets of alternatives in our model, we need to assume that probabilities of selecting

---

42 For the sake of simplicity, we do not incorporate the correlations between the thresholds in our formulation (see Bhat and Srinivasan, 2005).
aggregated sets of alternatives are independent, similar to the independent availability assumption of random constraint models (see Swait and Ben-Akiva, 1987). Further research is required in order to understand the implication of this assumption in the calculation of the choice set probabilities. All points considered, the probability of selecting choice set $C_n$ is defined as following:

$$Q_n(C_n) = K \prod_{C_r \in C} P_n(C_r \in C) \prod_{C_r \notin C_n} (1 - P_n(C_r \in C))$$

(6-13)

where, $K$ is a normalisation constant in order to allow the choice set probabilities over $G$ to sum to one.

The overall estimation approach has the following stages: First, the Multinomial Ordered Probit model presented in Equation 6-10 is estimated using the Maximum Likelihood approach (ML). Then, the parameters of choice model of Equation 6-8 is estimated using the ML approach, while the choice set probabilities are calculated following Equation 6-13 based on the estimated parameters of the Multinomial Ordered Probit model. In the next section, we conduct a Monte Carlo experiment to compare the proposed model against the original Manski model and to show the bias inherent in the KBS model using simulated data.

### 6.5 Monte Carlo Simulation

In order to illustrate the performance of our approach against the original Manski model and to show the bias inherent in the KBS model, we have conducted two Monte Carlo case studies. The simulation in general has three stages: (i) simulation of alternatives (dwelling units), (ii) simulation of decision makers (households), (iii) assigning alternatives to decision makers based on the proposed choice model.

The generation of alternatives was conducted by randomly scattering dwellings units across a region which comprised five Travel Analysis Zones (TAZ). For each alternative, we generated two attributes: (i) price, and (ii) location (TAZ). Dwelling prices and dwelling locations (TAZs) were generated independently. This might not be true in general but for the purposes of our analysis this independency assumption is not limiting. The number of alternatives in each case study is different and is described in the following sub-sections.
For each household we generated three characteristics: (i) income, (ii) TAZ of the workplace location, and (iii) price threshold. Here again, the generation of incomes and workplace locations are independent although it might not be true in reality. We adopted quite a large sample size (NH=1000) in order to avoid biases associated with small sample sizes.

The price threshold was generated based on the following latent threshold as a function of household income:

\[ t^*_n = \alpha I_n + \varepsilon_n \tag{6-14} \]

where, \( I_n \) is the income of household \( n \), and \( \varepsilon_n \) is a standard normal error term. The postulated parameter of the latent threshold is assumed to be one (i.e., \( \alpha = 1 \)).

We assumed that alternatives are classified into NCAT categories (different for each case study as discussed later) based on their prices in a hypothetical real estate agency website. Without loss of generality, it can also be assumed that the number of alternatives in each category is the same.

The alternatives were assigned to decision makers based on Equation 6-8 following a Monte Carlo sampling process. This procedure generates a random number between 0 and 1, and compares it to the cumulative choice probability of alternatives. The alternative that has a cumulative probability interval which contains the random number is assigned to the households. The following utility function is assumed for the compensatory stage:

\[ U_{in} = \beta_1 TT_{in} + \beta_2 P_i + \varepsilon_{in} \tag{6-15} \]

where,

\( TT_{in} \) is the travel time from the TAZ of dwelling \( i \) to TAZ of the workplace of household \( n \),

\( P_i \) is price of alternative \( i \) divided by 1000, and

\( \varepsilon_{in} \) is the error term, which is independent and identically distributed (IIA) across alternatives and decision makers. Finally, the postulated parameters for compensatory stage are assumed as: \( \beta_1 = \beta_2 = -1 \).
6.5.1 Case Study 1

The first case study concerns hypothetical choice situations with 10, 8 and 6 alternatives in order to be able to implement the original Manski model. The simulation had two stages. First, we assumed that the Manski model is the true model; therefore, we used Manski's choice probabilities to generate the chosen alternatives for all households in the sample. Secondly, we estimated three models (M1, M2 and M3) varying in number of categories for each choice situation. Model estimation was achieved using a randomly drawn 75% from the sample with the remaining 25% being kept aside as a hold-out sample. M1 corresponded to the original Manski model where the number of alternatives is equal to the number of categories. M2 and M3 were estimated based on grouping of alternatives into different categories. For simplicity, we kept the number of alternatives within each category the same as discussed before. Table 6-1 summarises the specifications of the different models.

<table>
<thead>
<tr>
<th>Choice Situation 1</th>
<th>Number of Alternatives</th>
<th>Number of Categories</th>
</tr>
</thead>
<tbody>
<tr>
<td>M1</td>
<td>NALT = 10</td>
<td>NCAT = 10</td>
</tr>
<tr>
<td>M2</td>
<td>NALT = 10</td>
<td>NCAT = 5</td>
</tr>
<tr>
<td>M3</td>
<td>NALT = 10</td>
<td>NCAT = 2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Choice Situation 2</th>
<th>Number of Alternatives</th>
<th>Number of Categories</th>
</tr>
</thead>
<tbody>
<tr>
<td>M1</td>
<td>NALT = 8</td>
<td>NCAT = 8</td>
</tr>
<tr>
<td>M2</td>
<td>NALT = 8</td>
<td>NCAT = 4</td>
</tr>
<tr>
<td>M3</td>
<td>NALT = 8</td>
<td>NCAT = 2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Choice Situation 3</th>
<th>Number of Alternatives</th>
<th>Number of Categories</th>
</tr>
</thead>
<tbody>
<tr>
<td>M1</td>
<td>NALT = 6</td>
<td>NCAT = 6</td>
</tr>
<tr>
<td>M2</td>
<td>NALT = 6</td>
<td>NCAT = 3</td>
</tr>
<tr>
<td>M3</td>
<td>NALT = 6</td>
<td>NCAT = 2</td>
</tr>
</tbody>
</table>

The 25% hold-out sample was used to compare the predicted market share of the proposed model (with a different number of categories) against the true market share (market share of the hold-out sample). Two performance measures: (a) Root Mean Square Error (RMSE), and (b) Mean Absolute Percentage Error (MAPE) were used for the evaluation of the proximity of the predicted and true market share of the different models. We have also reported the estimated parameters for different models; however, we could not compare the estimated parameters against the postulated true parameters, except for M1, because of the different structure of the models. The average results of 10 Monte Carlo simulation runs are tabulated in Table 6-2.
Table 6-2 – Estimated Parameters and Prediction Tests of Case Study 1  
(Average of 10 runs)

<table>
<thead>
<tr>
<th></th>
<th>NALT=6</th>
<th></th>
<th>NALT=8</th>
<th></th>
<th>NALT=10</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NCAT = 6</td>
<td>Estimate</td>
<td>t-stat</td>
<td>NCAT = 3</td>
<td>Estimate</td>
<td>t-stat</td>
</tr>
<tr>
<td>β1</td>
<td>-0.95</td>
<td>-3.43</td>
<td>β1</td>
<td>-0.44</td>
<td>-13.93</td>
<td>β1</td>
</tr>
<tr>
<td>β2</td>
<td>-0.92</td>
<td>-3.86</td>
<td>β2</td>
<td>-0.28</td>
<td>-13.78</td>
<td>β2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prediction test on the hold-out sample</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RMSE</td>
<td>1.83</td>
<td></td>
<td>RMSE</td>
<td>15.53</td>
<td></td>
<td>RMSE</td>
</tr>
<tr>
<td>MADE</td>
<td>1.68</td>
<td></td>
<td>MADE</td>
<td>15</td>
<td></td>
<td>MADE</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NCAT = 8</td>
<td>Estimate</td>
<td>t-stat</td>
<td>NCAT = 4</td>
<td>Estimate</td>
<td>t-stat</td>
<td>NCAT = 2</td>
</tr>
<tr>
<td>β1</td>
<td>-0.95</td>
<td>-2.51</td>
<td>β1</td>
<td>-0.16</td>
<td>-7.21</td>
<td>β1</td>
</tr>
<tr>
<td>β2</td>
<td>-1.02</td>
<td>-5.86</td>
<td>β2</td>
<td>-0.16</td>
<td>-8.88</td>
<td>β2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prediction test on the hold-out sample</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RMSE</td>
<td>5.54</td>
<td></td>
<td>RMSE</td>
<td>18.58</td>
<td></td>
<td>RMSE</td>
</tr>
<tr>
<td>MADE</td>
<td>3.76</td>
<td></td>
<td>MADE</td>
<td>14.94</td>
<td></td>
<td>MADE</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NCAT = 10</td>
<td>Estimate</td>
<td>t-stat</td>
<td>NCAT = 5</td>
<td>Estimate</td>
<td>t-stat</td>
<td>NCAT = 2</td>
</tr>
<tr>
<td>β1</td>
<td>-0.9</td>
<td>-2.8</td>
<td>β1</td>
<td>-0.06</td>
<td>-4.21</td>
<td>β1</td>
</tr>
<tr>
<td>β2</td>
<td>-0.95</td>
<td>-3.25</td>
<td>β2</td>
<td>-0.06</td>
<td>-3.88</td>
<td>β2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prediction test on the hold-out sample</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RMSE</td>
<td>7.28</td>
<td></td>
<td>RMSE</td>
<td>24.06</td>
<td></td>
<td>RMSE</td>
</tr>
<tr>
<td>MADE</td>
<td>6.1</td>
<td></td>
<td>MADE</td>
<td>20.45</td>
<td></td>
<td>MADE</td>
</tr>
</tbody>
</table>

The results also show that performance of the proposed model is very sensitive to the imposed structure on the choice set space. Hence, similar to the GenL model (Swait, 2001), the choice set space should be considered as a part of the model specification. As expected, however, the RMSE and MAPE measures of proximity suggest that the performance of M1 is better than M2 and the performance of M2 is better than M3.

The results show that the proposed model with restricted choice set space cannot approximate the model with the full choice set space (i.e., the Manski model) and it should be considered as a model in its own right for searching and choice set formation. Further empirical studies are required to evaluate the performance of the proposed approach and to determine a suitable structure (i.e., number of thresholds and number of categories) before applying the proposed model in practice. This shapes the future work of this study.
It should be noted that this case study assumes that the true behaviour is represented by the Manski model. Although all empirical studies have confirmed the superiority of the Manski model against the MNL model (see Swait and Ben-Akiva, 1987), the performance of the proposed simplified probabilistic choice set formation model against the MNL should also be investigated using real data. This remains a task for future research.

6.5.2 Case Study 2

The second case study concerns a hypothetical choice situation among 50 alternatives. Clearly, estimation of the original Manski model with 50 alternatives is not feasible given current computing power. In the first stage of simulation of this case study, therefore, the proposed model was used with five categories to generate the chosen alternatives for all households in the sample. Here again, for the sake of simplicity, we kept the number of alternatives within in each category the same (i.e., ten alternatives in each category).

In the second stage, we fitted the KBS model and the MNL model to the simulated data. For comparison purposes, we also fitted the proposed model with five categories (assumed to be a true model in this case study). Here again, model estimation was achieved using 75% the simulated data with the remainder kept as a hold-out sample. The hold-out sample was used to compare the predicted market share of the proposed model against the KBS model and the MNL model using RMSE and MADE measures. The estimated parameters are also reported but, as discussed earlier, we cannot draw any conclusions from the estimated parameter because of the different structures of the models. Table 6-3 tabulates the results of the averages of 10 Monte Carlo simulation runs.

The prediction tests using the hold-out sample method show that the KBS model cannot reproduce the market shares with good approximation whereas the MNL model retrieves the true market shares more accurately. The KBS model is therefore not suitable for choice situations in which the chosen alternative is not a member of the choice set derived from observed thresholds. In fact, where the chosen alternatives for a household are not a member of the choice set derived from their observed thresholds, that household should be eliminated in order to be able to estimate the KBS model. Our
simulation confirms that, on average, only 19% of households will end up with alternatives that belong to the choice set derived from their observed thresholds of the final search iteration. The results of this case study illustrate that the KBS model causes biases in predicted market share if we take into account the probabilistic nature of choice sets and assume that individuals are free to choose from any potential choice sets derived from combinations of thresholds.

Table 6-3 – Estimated Parameters and Prediction Tests of Case Study 2
(Average of 10 runs)

<table>
<thead>
<tr>
<th>Proposed Model</th>
<th>Estimate</th>
<th>t-stat</th>
<th>KBS</th>
<th>Estimate</th>
<th>t-stat</th>
<th>MNL</th>
<th>Estimate</th>
<th>t-stat</th>
</tr>
</thead>
<tbody>
<tr>
<td>β1</td>
<td>-1.01</td>
<td>-10.19</td>
<td>β1</td>
<td>-0.83</td>
<td>-6.99</td>
<td>β1</td>
<td>-0.24</td>
<td>-17.86</td>
</tr>
<tr>
<td>β2</td>
<td>-0.97</td>
<td>-9.96</td>
<td>β2</td>
<td>-0.62</td>
<td>-6.81</td>
<td>β2</td>
<td>-0.21</td>
<td>-12.77</td>
</tr>
</tbody>
</table>

Prediction test on the hold-out sample

<table>
<thead>
<tr>
<th></th>
<th>RMSE</th>
<th>MADE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proposed Model</td>
<td>1.98</td>
<td>1.34</td>
</tr>
<tr>
<td>KBS</td>
<td>6.49</td>
<td>4.16</td>
</tr>
<tr>
<td>MNL</td>
<td>3.93</td>
<td>2.75</td>
</tr>
</tbody>
</table>

6.6 Conclusion

Estimation of probabilistic choice set formation models based on Manski (1977) formulation is impractical even for only a moderately large number of alternatives, as is typically the case in most spatial choice contexts. This chapter has described and numerically illustrated that the simplicity of the simplified choice set formation model developed by Kaplan, Bekhor and Shiftan (2009; 2011; 2012) arises because of unrealistic assumptions that they made regarding the choice set space. It has also proposed a novel simplified probabilistic choice set formation approach to model simultaneously both choice set formation and choices of households in online real estate websites. Using this simulated dataset, the performance of the proposed model against the Manski model in predicting the true market share was evaluated using the hold-out sample technique. The results show that the proposed model cannot approximate the model with the full choice set space and should be considered as a model in its own right for searching and choice set formation. The bias inherent in the KBS model was also illustrated by comparing it against the proposed model and the MNL model.
Future research will need to assess the performance of the modelling framework developed in this study using empirical data. Further empirical studies are also required to determine a suitable model structure for updating the probabilities of selecting aggregated sets of alternatives when the number of search iterations is more than one. The validity and the implications of assuming that probabilities of selecting aggregated sets of alternatives are independent should also be empirically examined.
CHAPTER 7

CONCLUSION

7.1 Overview

In the past five decades, urban economists, transportation engineers, and land use planners have been actively involved in the development of planning and policy simulation tools, applying principles of urban economics, econometrics and simulation techniques in order to predict the spatial distribution of households (and employment) in an urban area and to analyse land use and transportation polices. This thesis has contributed to the growing field of urban modelling in several methodological and empirical aspects as summarised in section 7.2. The thesis also raises some questions for future research, as discussed in section 7.3.

7.2 Summary of Contributions

The development of more accurate urban simulation models is crucial for the analysis of urban and transportation policies and projects. Residential location choice models are one of the most important components of such urban simulation models. Although the development of empirical residential location choice models dates back to the works of Lerman (1976) and McFadden (1978), significant empirical and methodological challenges remain. This study aimed to address these challenges by improving the representation of the housing market in residential location choice modelling.

In order to achieve this aim, an extensive survey of literature was conducted to identify the key challenges involved in the modelling of residential location choices from both methodological and empirical perspectives. Subsequently, a comprehensive conceptual framework for the housing market was developed that incorporates the processes involved in both the demand-side and the supply-side of the market. This conceptual framework sheds light on how the market works in reality and is used as the basis for the development of improved residential location choice models capturing the wider aspects of the housing market as described in chapters 4, 5 and 6.
Chapter 4 proposed to model residential location choices at the level of dwelling units in order to acknowledge the heterogeneity of housing market products and to circumvent the aggregation bias inherent in spatially aggregated models. An important challenge in the development of dwelling-level models is the relaxation of the IIA assumption of conventional MNL models. Due to the locational fixity of housing market products (i.e., the association of dwelling units with space), the IIA assumption of the conventional MNL models is no longer valid. The model proposed in chapter 4, therefore, develops a GEV and a Mixed Logit formulation to relax the IIA assumption and to capture the spatial correlation and spatial heteroscedasticity in the proposed zonal-based dwelling-level residential location choice model.

Another critical challenge in the development of dwelling-level residential location choice models is the lack of independent dwelling supply data for constructing the universal choice set. Chapter 4 also proposed, therefore, an innovative approach to synthesise dwelling units based on aggregated dwelling unit data in order to reduce the bias in the choice set construction evident in models previously developed at the level of the dwelling unit. The performance of this dwelling synthesising approach was evaluated numerically using a Monte Carlo simulation. The results of the empirical application of the dwelling synthesising approach for London also confirmed the validity of the proposed approach to choice set construction in terms of both achieving high goodness of fit and retrieving the true parameters with reasonably small errors. The dwelling synthesiser approach set out in chapter 4 is practice-ready and can be used to estimate residential location choice models at the level of dwelling units without requiring independent and disaggregated dwelling supply data.

Chapter 5 acknowledged the role of information and the underlying search process in residential location choice modelling. It has been a common belief in the literature that the underlying search process can be accommodated in discrete choice models of residential location choices by formulating behaviourally plausible rules describing households’ search behaviour (e.g., anchor-based search strategies), and then using these rules to generate the expected choice sets. Chapter 5 investigates the feasibility and practicality of applying exogenous choice set formation approaches in the residential location choice context for capturing the underlying search process of households. In particular, the empirical performance of the deterministic constraint
approach and the importance sampling approach were compared with more conventional statistical methods of choice set pruning (i.e., Simple Random Sampling). The results of this chapter showed that exogenous choice set formation approaches based on the simplified rules and screening models were unlikely to capture accurately the underlying search process of households in the housing market. In fact, it was found that residential location choice models with the expected choice sets resulting from the anchor-based search strategy and the hazard-based screening model performed less well than the alternative statistical pruning approaches. The empirical results of chapter 5 also indicate that any systematic mis-specification in the choice set formation will be propagated into the resulting choice model, and reflected in a degraded prediction performance. The empirical comparison of alternative exogenous choice set formation approaches provides a guideline for modellers and land use planners to avoid inappropriate choice set formation approaches in practice.

Chapter 6 investigated the uncertainty of choice sets in the residential location choice context and proposed a simplified probabilistic choice set formation model. Chapter 6 also illustrated that the simplicity of the simplified choice set formation model developed by Kaplan, Bekhor and Shiftan (2009; 2011; 2012) arises because of unrealistic assumptions that they made regarding the choice set space. The proposed simplified two-stage choice set formation model can be applied to model simultaneously both the choice set formation and residential choices of households in online real estate websites.

7.3 Future Research and Potential Extensions

This thesis has focused mainly on the modelling of residential location choices. As described in the conceptual framework of chapter 3, however, there are other choice dimensions involved in the housing market, such as households’ residential mobility/refurbishment/tenure decisions on the demand-side of the market, and real estate developers’ decisions on location/type/quantity of new real estate developments on the supply-side. The development of improved empirical models of these various choices made by different agents shapes the future goals of this research project. Obviously, this is crucial for the development of a fully market-based urban simulation model and can only be achieved by further data collection and empirical analysis.
Additionally, this study has revealed the need for a range of further work concerning residential location choice modelling which are summarised in this section.

Chapter 4 set out the theoretical frameworks of complex discrete choice models which are able to capture spatial correlation in dwelling-level residential location choice models. It also proposed an innovative approach to tackle the choice set construction bias of dwelling-level models in the absence of disaggregated dwelling supply data. The empirical application of chapter 4 is limited, however, to the conventional MNL model, with the objective of illustrating the practicality of the proposed dwelling synthesiser approach. An important potential extension to this work, therefore, would be to estimate the proposed dwelling-level GEV model based on our real data. As discussed in chapter 4, such estimation could be achieved based on the recent work Guevara and Ben-Akiva (2013) which provides a consistent estimator for GEV class models on a subset of alternatives.

Unlike MNL and GEV models, a consistent estimator for Mixed Logit models using only a subset of alternatives has not yet been developed. Numerical experiments, however, have shown that the sampling of alternatives approach for random parameter Mixed Logit models may not significantly affect the parameter estimates based on the MSL approach, as discussed in chapter 2 and chapter 4. Further research is required to study numerically the effect of sampling of alternatives in the estimation of Mixed Logit models with spatial correlation and/or heteroscedasticity across alternatives using the MSL and MACML approach. This would be a major step towards the estimation of the proposed Mixed Logit model of chapter 4.

Another important extension to this work would be to apply the proposed dwelling-level GEV model to the forecasting of dwelling choices in order to assess its methodological improvements compared to the conventional MNL model. Forecasting of non-MNL models based on a subset of alternatives is still an unexplored area of research and requires further work (see Kikuchi et al., 2003).

The dwelling synthesiser approach proposed in chapter 4 to reduce the choice set construction bias in dwelling-level models has been validated numerically using the Monte Carlo simulation approach. Further theoretical work is required, however, to reinforce our simulation results. More importantly, the proposed dwelling synthesiser
approach should be validated by comparing the results with actual disaggregated dwelling supply data in other empirical contexts where this type of data is available.

The analysis of chapter 5 was limited to choice set formation approaches based on the anchor-based search strategy. There are other spatial search strategies (e.g., supply-constraint search strategies), however, that could be implemented using the importance sampling approach or the deterministic constraint approach to construct the choice sets. Implementation of alternative search strategy approaches for choice set construction of residential location choice models is a task that remains for future work in this area. It would be interesting to test the robustness of our results based on different spatial search strategies and based on different specifications of the screening model.

Chapter 6 adopted a Monte Carlo simulation approach in order to evaluate the performance of the proposed simplified two-stage choice set formation model. Further data collection and empirical work is required to estimate and evaluate the proposed model in an empirical context. It would also be interesting to compare the proposed model (i.e., where choice set probabilities are estimated based on observed data) with Swait’s (2001) GenL model (where choice set probabilities are estimated latently) in a similar choice set space. An empirical analysis on a suitable choice set space for different choice situations is also crucial in this area.

Residential location choice models can be further improved by extending the standard random utility framework to reflect other behavioural aspects of location choices. For example, households’ attitudes and perceptions towards residential locations can be incorporated in discrete choice modelling using the latent class modelling paradigm (see Greene and Hensher, 2013).

Another important extension for residential location choice modelling, and for urban simulation modelling in general, would be to incorporate the investment features of the housing market in applied modelling. While urban economists are more interested in the consumption features of the housing market and thus tend to be involved in the development of spatially explicit models of the housing market, real estate economists are more interested in the investment aspects of the housing market and treat real estates as financial assets. Clearly, there is a need for a great deal of research in this area.
before these different streams of research can be combined in urban simulation models incorporating both the consumption and investment features of the housing market.

The most state-of-the-art urban simulation models predict prices by matching the supply and demand based on the consumption preference of households in a market clearing process (see Hurtubia et al., 2011; Farooq and Miller, 2012), but do not consider the investment preference of households. Obviously, it is necessary to incorporate the investment features of the housing market in urban simulation models in order to forecast the future evolution of housing prices and to evaluate policy scenarios related to housing affordability in a more realistic way. Some authors have already started combining the consumption and investment features of housing market based on the dynamic discrete choice modelling framework (see Bayer et al., 2007; Murphy, 2013).
BIBLIOGRAPHY


ANAS, A. 1983a. The Chicago area transportation-land use analysis system: a dynamic methodology for forecasting the impact of multimodal metropolitan transportation changes on travel mode choices, residential location, housing stock adjustment, housing values and value capture potential, U.S. Dept. of Transportation, University Research Program.


LERMAN, S. 1976. Location, housing, automobile ownership, and mode to work: a joint choice model. Transportation Research Record, 610.


MARTINEZ, C., JAVIER, F. & DONOSO, P. P. MUSSA II: a land use equilibrium model based on constrained idiosyncratic behavior of all agents in an auction market. 2007.


MURPHY, A. 2013. A dynamic model of housing supply. *Available at SSRN 2200459.*


QUIGLEY, J. M. 1979. What have we learned about urban housing markets? Current issues in urban economics, 391-429.


WADDELL, P. Accessibility and residential location: the interaction of workplace, residential mobility, tenure, and location choices. 1996.


WILSON, A. G. 1970. Entropy in urban and regional modelling, Pion Ltd.


162


APPENDIX A

COMPUTATION OF WALRASIAN EQUILIBRIUM

This appendix examines determination of equilibrium dwelling prices in the housing market where demand for dwellings are represented by probabilistic choice model. The notion of equilibrium provides a simple and powerful market clearing approach for simulation of housing market. Here, we assume that the utility that household $n$ associates with dwelling $d$ is specified as:

$$V_{nd} = \alpha R_d + \hat{V}_{nd}$$  \hspace{1cm} (A-1)

where,

$R_d$ is the price of dwelling $d$, $\hat{V}_{nd}$ is the remaining part of the systematic utility.

Assuming that the unobserved part of the utility is distributed according to type I extreme value (Gumbel) distribution, the probability that household $n$ chooses dwelling $d$ is given as:

$$P_{nd} = \frac{\exp(V_{nd})}{\sum_{d'} \exp(V_{nd'})}$$  \hspace{1cm} (A-2)

A probabilistic utility-maximising distribution of these $N$ households among $D$ dwellings based on the simulated prices will result in a nonmarket clearing assignment. Some dwellings will be chosen by more than one household, while some others will remain vacant. In order to have one to one assignment of dwellings and households, we need to compute the vector of equilibrium dwelling prices ($i.e., \bar{R}_e = [R_1, R_2, ..., R_j, ..., R_D]$) that clears the market.

The expected demand for the $d$th dwelling for given price vector $\bar{R}$ is:

$$demand_d(\bar{R}) = \sum_{n=1}^{N} P_{nd}(\bar{R}), \hspace{1cm} d = 1, ..., D$$  \hspace{1cm} (A-3)

We define the expected excess demand function as:

$$e_d(\bar{R}) = demand_d(\bar{R}) - 1, \hspace{1cm} d = 1, ..., D$$  \hspace{1cm} (A-4)
We define stochastic equilibrium as:

\[ e_d(\bar{R}_e) = \text{demand}_d(\bar{R}_e) - 1 = 0, \quad d = 1, \ldots, D \]  
(A-5)

We can also say that expected excess demand for each dwelling vanishes when \( \bar{R} = \bar{R}_e \).

\[ \sum_{n=1}^{N} P_{nd}(\bar{R}_e) - 1 = 0, \quad d = 1, \ldots, D \]  
(A-6)

The equilibrium vector \( \bar{R}_e \) can be proven to be unique based on Brouwer fixed-point theorem (Anas, 1982).

In order to compute equilibrium prices, these \( D \) equations must be solved simultaneously. Here, we adopt an iterative price adjusting procedure similar to Anas (1982). We begin with an arbitrary initial price vector \( \bar{R}_1 \) and adjust the price vector until, at the final iteration; the adjusted price vector is close to the equilibrium vector \( \bar{R}_e \).

Let the iterations be numbered as \( k = 1, \ldots, K \) such that \( \bar{R}_K = \bar{R}_e \). The adjusted rent vector for the \((k+1)\)th iteration is computed from the vector of the \( k \)th iteration as:

\[
\bar{R}_{k+1} = \bar{R}_k - \begin{bmatrix}
    e_1(\bar{R}_k) \\
    \vdots \\
    e_D(\bar{R}_k)
\end{bmatrix} \begin{bmatrix}
    \frac{\partial e_1(\bar{R}_k)}{\partial R_1} & \cdots & \frac{\partial e_D(\bar{R}_k)}{\partial R_1} \\
    \vdots & \ddots & \vdots \\
    \frac{\partial e_1(\bar{R}_k)}{\partial R_D} & \cdots & \frac{\partial e_D(\bar{R}_k)}{\partial R_D}
\end{bmatrix}
\]

(A-7)

The new prices are calculated until the absolute value of the excess demands at convergence is less than 0.1% for each dwelling.

\[ |e_d(\bar{R}_K)| \leq 0.001, \quad d = 1, \ldots, D \]  
(A-8)

The derivative of expected excess demand function should be calculated analytically in order to obtain the equilibrium prices more efficiently. The derivative of expected excess demand function is calculated as:

\[
\frac{\partial e_d(\bar{R}_k)}{\partial R_d} = \sum_{n=1}^{N} \alpha P_{nd}(1 - P_{nd}), \quad d = 1, \ldots, D
\]  
(A-9)
APPENDIX B

DWELLING SYNTHESISER ALGORITHM

B.1. Overview

In recent years, combing survey data with census or administrative data based on statistical models has received wide attention by researchers and practitioners in order to produce finer-level statistics for variables that are collected only by sample survey. Spatial microsimulation (including synthetic reconstruction and combinatorial optimisation methods), mass imputation, and small area estimation are three different approaches that are used in the literature to combine survey and census data in order to produce finer-level statistics (Haslett et al., 2010). While there are important differences between these three methods that affect their applications, they are fundamentally similar as discussed by Haslett et al. (2010). In transport and urban planning application, the synthetic reconstruction approach (i.e., spatial microsimulation) have been widely used to construct microdata representing the characteristics of the decision makers as the baseline population for microsimulation travel demand models and land use models (see Beckman et al., 1996; Ryan et al., 2009).

Following Beckman et al. (1996), there have been different attempts in the literature to improve the synthetic reconstruction approach to deal with zero-cell and zero-marginal problems and to control household and person level attributes simultaneously (see Guo and Bhat, 2007b; Auld et al., 2009; Ye et al., 2009; Auld and Mohammadian, 2010; Müller et al., 2011). Reviewing all of these different approaches is not the scope of this appendix, next section presents the synthetic reconstruction algorithm used to create the synthetic population of dwellings in chapter 4. This approach is adopted from Guo and Bhat (2007b) and modified accordingly to synthesise the population of dwellings. The synthesising algorithm is implemented in Matlab.

B.2. Dwelling Synthesiser Algorithm

The first stage in the synthetic reconstruction approach is to calculate zonal joint distributions using the Iterative Proportional Fitting (IPF) procedure. For each zone in the area, the IPF algorithm is used to estimate the joint distribution of variables that
their marginal distributions are available in the census data (i.e., controlled variables). In addition to aggregated census data, the IPF procedure takes joint distribution of controlled variables (i.e., seed) constructed from disaggregated data (e.g., household survey) as inputs. Here, we are using an example to describe the IPF algorithm for a two-dimensional table in Figure B-1. The readers are referred to Beckman et al. (1996) for formal description of the IPF algorithm. It should be noted that for K controlled variables, a K-dimensional table should be estimated for each zone.

In the second stage, for each zone, synthetic dwellings are generated by probabilistic sampling of dwellings from the disaggregated data based on the estimated joint distribution of the zone. Similar to Guo and Bhat (2007b), we denote the joint distribution of dwellings within a zone (i.e., output of the IPF procedure) by $D_z(\text{vec})$, where vec is the vector of values of controlled variables. $D_z(\text{vec})$ gives the expected number of dwellings with the attributes values of $\text{vec} = [v_1, v_1, ..., v_K]$ in zone $z$.

In the selection procedure, we randomly draw a dwelling unit (with its all attributes, controlled or not controlled) from the disaggregated data for each zone, and assign it to the zone, until we reach the total number of dwellings in the zone. Similar to Guo and Bhat (2007b), we compute the selection probability as:

![Figure B-1 - Illustration of the IPF Algorithm for a Two-dimensional Table](image)
\( P_{dz} = \frac{D_z(\text{vec}) - DI_z(\text{vec})}{\sum_{\text{vec}} (D_z(\text{vec}) - DI_z(\text{vec}))} \) \hspace{1cm} (B-1)

where,

\( DI_z(\text{vec}) \) is the number of dwellings with the attributes values of \( \text{vec} \) that are already selected in the zone. \( DI_z(\text{vec}) \) gets updated during iterative selection procedure and implies that the selection probability of a dwelling with the attributes values of \( \text{vec} \) decreases as more dwellings with similar attributes are selected into the zone.

The synthesising of dwelling units has one-level, hence, it is much simpler than synthesising of persons and households in which household and person level attributes should be controlled simultaneously. It should also be noted that we do not incorporate the household weights (available in the household survey) in probabilistic drawing process as these weights are associated to the attributes of households rather than the attributes of dwellings.

**B.3. Validation of the Dwelling Synthesiser**

Validation of population synthesisers, in general, includes two set of tests: (i) tests on marginal distributions, and (ii) tests on joint distributions. Marginal distribution tests evaluate the performance of the synthesising algorithm in retrieving the true aggregated values of variables across all zones. Joint distribution tests evaluate the performance of the population synthesiser in preserving the correlation structures of variables across all zones.

In marginal distribution tests, we compare the actual marginal distributions of controlled variables (i.e., observed distributions) to the synthesised ones (i.e., expected distributions) for the whole study area. For all controlled variables\(^{43}\), the Mean Absolute Percent Error (MAPE) of the synthesised values of each category of the variables against the true values across all zones, and the average of MAPEs across all categories are calculated as:

\[
MAPE_j = \frac{1}{N} \sum_{i=1}^{N} \left| \frac{O_{ij} - E_{ij}}{O_{ij}} \right|
\] \hspace{1cm} (B-2)

\(^{43}\) It should be noted that marginal distribution tests can be also used for uncontrolled variables if their true marginal distributions are available.
\[ \text{Average(MAPE}_j) = \frac{1}{A} \sum_{j=1}^{A} \text{MAPE}_j \]  

where,

\( j \) refers to the category index of the variable,

\( i \) refers to the zonal index,

\( N \) is the number of zones in the study area,

\( O_{ij} \) is the actual marginal total for zone \( i \), category \( j \),

\( E_{ij} \) is the synthesised marginal total for zone \( i \) category \( j \), and

\( A \) is the number of categories of the variable.

There are different tests proposed in the literature for joint distributions. These tests are in fact goodness of fit tests for contingency tables and compare the observed frequencies with the expected frequencies derived from the synthesised population. However, in practice, the observed frequencies are not available for the actual population and the frequencies resulting from the IPF procedure are used to compute the statistics. Therefore, tests on joint distributions assume that the joint distributions from the IPF procedure are the true joint distributions. Hence, they can be at best used to validate the second stage of population synthesisers (i.e., the selection procedure). However, if the actual population do exist (e.g., in simulation studies), one can use joint distribution tests to validate both stages of population synthesising.

There are two types of joint distribution validation tests have been used in the literature: (i) distance-based measures, and (ii) chi-squared statistics. Distance-based measures calculate the distances between the actual joint distributions (i.e., resulting from the IPF procedure) and the expected joint distributions (i.e., resulting from the synthesised population). Different distance-based measures have been used in the literature such as the Frobenius norm and the Standardised Root Mean Square Error (SRMSE) which are in fact very similar. Here we use SRMSE measure which is defined as the RMSE normalised to the mean of the observed values as:
$SRMSE = \sqrt{\frac{1}{A_1 \times A_2 \times \ldots \times A_p} \sum_{a_1 \ldots a_p}^{A_1 A_2 \ldots A_p} \left( \frac{O_{a_1 \ldots a_p} - E_{a_1 \ldots a_p}}{E_{a_1 \ldots a_p}} \right)^2}$ (B-4)

where,

$A_i$ is the number of categories for the $i$th-dimension of the contingency table, and $p$ is the dimension of the contingency table.

Chi-square statistics are conventionally used to measure the goodness of fit of contingency tables (i.e., joint distributions) by comparing it against the chi-square distribution for a significance level of 5%. All of chi-square statistics for contingency tables are approximate; however, some of these statistics are more accurate (see Read, 1993). Some of the most important chi-square statistics for contingency tables are tabulated in Table B-1. These statistics follow a chi-square distribution with the degree of freedom of:

$$k = \prod_{i=1}^{P} A_i - \sum_{i=1}^{P} A_i + (p - 1)$$ (B-5)

Table B-1- Different Chi-squared Statistics Used for Contingency Tables

<table>
<thead>
<tr>
<th>Pearson's chi-squared</th>
<th>$\chi^2 = \sum_{a_1 \ldots a_p}^{A_1 A_2 \ldots A_p} \left( \frac{O_{a_1 \ldots a_p} - E_{a_1 \ldots a_p}}{E_{a_1 \ldots a_p}} \right)^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Likelihood ratio chi-squared (G-square)</td>
<td>$G^2 = 2 \sum_{a_1 \ldots a_p}^{A_1 A_2 \ldots A_p} O_{a_1 \ldots a_p} \ln \left( \frac{O_{a_1 \ldots a_p}}{E_{a_1 \ldots a_p}} \right)$</td>
</tr>
<tr>
<td>Freeman and Tukey chi-squared (variation 1)</td>
<td>$FT^2_0 = 4 \sum_{a_1 \ldots a_p}^{A_1 A_2 \ldots A_p} \left( \sqrt{O_{a_1 \ldots a_p}} - \sqrt{E_{a_1 \ldots a_p}} \right)^2$</td>
</tr>
<tr>
<td>Freeman and Tukey chi-squared (variation 2)</td>
<td>$FT^2_1 = \frac{4}{3} \sum_{a_1 \ldots a_p}^{A_1 A_2 \ldots A_p} \left( 1 + 2 \frac{O_{a_1 \ldots a_p}}{E_{a_1 \ldots a_p}} \right) \left( \sqrt{O_{a_1 \ldots a_p}} - \sqrt{E_{a_1 \ldots a_p}} \right)^2$</td>
</tr>
</tbody>
</table>
Both distance-based measures and chi-square statistics can be calculated for the whole study area and for each zone separately. For example, assuming that there are 3 controlled variables ($A= 5, 5, \text{ and } 4$), and 100 zones, two scenarios of joint distribution tests can be defined as:

- **Scenario 1**: One contingency table for 3 controlled variables and considering zones as an additional dimension
- **Scenario 2**: 100 contingency tables for 3 controlled variables (i.e., one contingency table for each zone)

In the case of scenario 2, the average of SRMSE values across all zones and the number of zones that their joint distributions were fitted based on the chi-square test can be used to evaluate the performance of the synthesiser. Table B-2 exemplifies the parameters of chi-square joint distribution tests for the described scenarios.

<table>
<thead>
<tr>
<th>Description</th>
<th>Dimension</th>
<th>Degree of Freedom</th>
<th>5% Critical Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scenario 1</td>
<td>$p=4$</td>
<td>$5 \times 5 \times 4 \times 100 - (5 + 5 + 4 + 100) - (4 - 1) = 9889$</td>
<td>10121.45</td>
</tr>
<tr>
<td>Scenario 2</td>
<td>$p=3$</td>
<td>$5 \times 5 \times 4 - (5 + 5 + 4) + (3 - 1) = 88$</td>
<td>110.89</td>
</tr>
</tbody>
</table>

**B.4. Validation Results for the Simulated Data**

This section presents the results of validation of the dwelling synthesiser for the simulated data discussed in chapter 4. As mentioned in chapter 4, six separate populations varying in terms of the household survey sizes (i.e., NH=500, NH=1000, and NH=2000), and the market clearing approaches (i.e., equilibrium and disequilibrium) were synthesised. For each case, we have conducted 10 repetitions of the population synthesiser in order to check the robustness of results. Since the validation results were similar for different sizes of the household survey, for equilibrium and disequilibrium cases, and for different repetitions, here we present the validation results of one of the synthesised population.

The marginal validation results of the dwelling synthesiser are presented in Figure B-2. Since in the simulation practice the actual dwelling population is available, we can evaluate the goodness of fit of joint distributions against the actual joint distributions. The results of joint distribution tests are tabulated in Table B-3, Table B-4, and Table B-5. The validation results show that the synthesising algorithm can effectively reproduce
the population of the dwelling units both in terms of marginal values and joint distributions.

![Figure B-2: MAPE and Average MAPE between True Marginal Distributions and Synthesised Marginal Distributions for the Simulated Data](image)

**Table B-3: Distance-based Measures for Joint Distributions (Scenario 1 & 2), Actual Population vs. Synthesised Population**

<table>
<thead>
<tr>
<th></th>
<th>SRMSE</th>
<th>Mean</th>
<th>STD</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2.19</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>2</td>
<td>2.15</td>
<td>0.302</td>
<td>-</td>
</tr>
</tbody>
</table>

**Table B-4: Chi-square Joint Distribution Tests (Scenario 1), Actual Population vs. Synthesised Population**

<table>
<thead>
<tr>
<th>Statistics</th>
<th>Value</th>
<th>DF</th>
<th>Critical Value</th>
<th>Fitted?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chi²</td>
<td>7752.04</td>
<td>9889</td>
<td>10121.45</td>
<td>Yes</td>
</tr>
<tr>
<td>G²</td>
<td>10311.11</td>
<td>9889</td>
<td>10121.45</td>
<td>No</td>
</tr>
<tr>
<td>FT2_0</td>
<td>14592.22</td>
<td>9889</td>
<td>10121.45</td>
<td>No</td>
</tr>
<tr>
<td>FT2_1</td>
<td>9025.367</td>
<td>9889</td>
<td>10121.45</td>
<td>Yes</td>
</tr>
</tbody>
</table>
Table B-5- Chi-square Joint Distribution Tests (Scenario 2), Actual Population vs. Synthesised Population

<table>
<thead>
<tr>
<th>Statistics</th>
<th>Mean</th>
<th>STD</th>
<th>Degree of Freedom</th>
<th>Critical Value</th>
<th>Number of Zones</th>
<th>Number of Fitted?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chi2</td>
<td>77.52</td>
<td>19.65</td>
<td>88</td>
<td>110.89</td>
<td>100</td>
<td>98</td>
</tr>
<tr>
<td>G2</td>
<td>103.11</td>
<td>31.10</td>
<td>88</td>
<td>110.89</td>
<td>100</td>
<td>55</td>
</tr>
<tr>
<td>FT2_0</td>
<td>145.92</td>
<td>37.22</td>
<td>88</td>
<td>110.89</td>
<td>100</td>
<td>18</td>
</tr>
<tr>
<td>FT2_1</td>
<td>90.25</td>
<td>22.78</td>
<td>88</td>
<td>110.89</td>
<td>100</td>
<td>81</td>
</tr>
</tbody>
</table>

B.5. Validation Results for the Real Data

This section presents the results of validation of the dwelling synthesiser for the real data discussed in chapter 4. The controlled variables used in the synthesiser were:

- Tenure type (i.e., own outright, own with mortgage, private rent, and social rent)
- Council tax band (i.e., band A to band H)
- Dwelling size (number of beds), and
- Dwelling type (i.e., detached, semi-detached, terraced and flat)

As mentioned in chapter 4, these variables were derived from census 2001. The disaggregated data includes 4,494 cases from the 2002 London Household Survey. The MAPEs for all categories of controlled variables and the average of MAPEs over all controlled variables are shown in Figure B-3. The largest MAPEs are for the bands F, G, and H of the council tax band variable. The reason for these large errors is that dwellings with these council tax bands are not very frequent in the data. A potential remedy for this problem is to reduce the number of categories of the council tax band variable by combining these categories (see Auld et al., 2009).
Figure B-3 - MAPE and Average MAPE between True Marginal Distributions and Synthesised Marginal Distributions for the Real Data
The results of joint distribution tests are presented in Table B-6, Table B-7, and Table B-8 which show that the synthesised joint distributions and the IPF joint distributions match very well. It should be noted that the joint distribution tests here do not provide any information for the goodness of fit of the dwelling synthesiser as whole because the actual disaggregated data do not exist for the study area (this is why we turn to synthesising approach in the first place). Therefore, these results only show that the selection procedure does not change the joint distributions resulting from the IPF procedure. These tests have been used to evaluate the performance of different selection procedures in the literature (see Auld and Mohammadian, 2010).

### Table B-6- Distance-based Measures for Joint Distributions (Scenario 1 & 2), IPF Result vs. Synthesised Population

<table>
<thead>
<tr>
<th></th>
<th>SRMSE</th>
<th>Mean</th>
<th>STD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scenario 1</td>
<td>0.287</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Scenario 2</td>
<td>0.616</td>
<td>0.978</td>
<td></td>
</tr>
</tbody>
</table>

### Table B-7- Chi-square Joint Distribution Tests (Scenario 1), IPF Result vs. Synthesised Population

<table>
<thead>
<tr>
<th>Statistics</th>
<th>Value</th>
<th>DF</th>
<th>Critical Value</th>
<th>Fitted?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chi2</td>
<td>46924.74</td>
<td>899197</td>
<td>901404</td>
<td>Yes</td>
</tr>
<tr>
<td>G2</td>
<td>57028.3</td>
<td>899197</td>
<td>901404</td>
<td>Yes</td>
</tr>
<tr>
<td>FT2_0</td>
<td>83386.05</td>
<td>899197</td>
<td>901404</td>
<td>Yes</td>
</tr>
<tr>
<td>FT2_1</td>
<td>53446.92</td>
<td>899197</td>
<td>901404</td>
<td>Yes</td>
</tr>
</tbody>
</table>

### Table B-8- Chi-square Joint Distribution Tests (Scenario 2), Actual Population vs. Synthesised Population

<table>
<thead>
<tr>
<th>Statistics</th>
<th>Mean</th>
<th>STD</th>
<th>Degree of Freedom</th>
<th>Critical Value</th>
<th>Number of Zones</th>
<th>Number of Fitted?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chi2</td>
<td>53.38</td>
<td>20.05</td>
<td>1003</td>
<td>1077.79</td>
<td>879</td>
<td>879</td>
</tr>
<tr>
<td>G2</td>
<td>65.26</td>
<td>24.89</td>
<td>1003</td>
<td>1077.79</td>
<td>879</td>
<td>879</td>
</tr>
<tr>
<td>FT2_0</td>
<td>94.86</td>
<td>35.31</td>
<td>1003</td>
<td>1077.79</td>
<td>879</td>
<td>879</td>
</tr>
<tr>
<td>FT2_1</td>
<td>60.80</td>
<td>22.07</td>
<td>1003</td>
<td>1077.79</td>
<td>879</td>
<td>879</td>
</tr>
</tbody>
</table>
APPENDIX C

DATA PREPARATION

C.1. Overview

This appendix details the sources of data and the construction procedure of the locational variables used in chapter 4 and 5. The data preparation of this study has been an exhaustive procedure involving many data manipulation and GIS tasks. The spatial data sources (i.e., in addition to household survey data) that have been used in this study are as follows:

- Generalised Land Use Database (GLUD), 2001 available at: www.neighbourhood.statistics.gov.uk
- Zone-to-zone travel time matrix of auto mode acquired from London Transportation Studies (LTS) Model
- Housing price paid data 2001 acquired from Land Registry: www.landreg.gov.uk
- Annual Business Inquiry (ABI) available at: www.nomisweb.co.uk
- OS MasterMap Integrated Transport Network (ITN) Layer available at: www.ordnancesurvey.co.uk
- School performance table available at: www.education.gov.uk/schools/performance
- Primary Schools, Point Location of Services (2005 Educational Establishments) available at: www.neighbourhood.statistics.gov.uk
- GIS files of zonal boundaries available at: edina.ac.uk/ukborders

These spatial data sources provide a rich set of locational variables for consideration in the model specification. The locational variables used in this study and the corresponding data sources are tabulated in Table C-1 and described in the following sections.
### Table C-1: Locational Variables and Data Sources

<table>
<thead>
<tr>
<th>Variable Group</th>
<th>Variables</th>
<th>Data Sources</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zonal Size and Density</td>
<td>Log of Zonal Area (hectare)</td>
<td>Census - Population (UV01), Population Density (UV02), Household Composition - Households (UV65), Dwellings (UV55)</td>
</tr>
<tr>
<td></td>
<td>Log of Number of Residents</td>
<td>Census - Household Composition - People (UV46), Household Composition - Households (UV65), Age (UV04), Income: Model-Based Estimates at Ward level, 2001/02, Census- Ethnic Group (UV09)</td>
</tr>
<tr>
<td></td>
<td>Log of Number of Households</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Log of Number of Dwellings</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Number of Residents per hectare</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Number of Households per hectare</td>
<td></td>
</tr>
<tr>
<td>Zonal Socioeconomic</td>
<td>Average Zonal Household Size</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Average Zonal Age of Residents</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Average Zonal Household Income</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ethnic Group: Percentage of White</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ethnic Group: Percentage of Black</td>
<td></td>
</tr>
<tr>
<td>Zonal Housing Type and Tenure Type Composition</td>
<td>Percentage of Detached Houses</td>
<td>Census - Accommodation Type - Household Spaces (UV56), Tenure - Households (UV63)</td>
</tr>
<tr>
<td></td>
<td>Percentage of Semi-detached Houses</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Percentage of Terraced Houses</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Percentage of Flats</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Percentage of Owner Households</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Percentage of Renter Households (Private Rent)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Percentage of Renter Households (Social Rent)</td>
<td></td>
</tr>
<tr>
<td>Zonal Land Use Structure</td>
<td>Percentage of Zonal Area Occupied by Domestic Buildings</td>
<td>Generalised Land Use Database (GLUD), 2001</td>
</tr>
<tr>
<td></td>
<td>Percentage of Zonal Area Occupied by Non-domestic Buildings</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Percentage of Zonal Area Occupied by Green Space</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Land Use Mix Diversity Index</td>
<td></td>
</tr>
<tr>
<td>Housing Prices</td>
<td>Average Zonal Housing Price</td>
<td>Land Registry</td>
</tr>
<tr>
<td>School Quality</td>
<td>Average Point Score</td>
<td>Primary Schools- Point Location of Services (2005), School Performance Table</td>
</tr>
<tr>
<td>Transportation Network</td>
<td>Proximity to Nearest Public Transport Point of Access</td>
<td>OS MasterMap Integrated Transport Network (ITN) Layer, National Public Transport Access Node (NaPTAN)</td>
</tr>
<tr>
<td></td>
<td>Dual Carriageway Density</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Single Carriageway Density</td>
<td></td>
</tr>
<tr>
<td>Accessibility</td>
<td>Employment Accessibility</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Shopping Accessibility</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Recreational Accessibility</td>
<td></td>
</tr>
<tr>
<td>Commute Variable</td>
<td>Commute Time Between Household Workplace Locations and Candidate Zone</td>
<td>Zone-to-zone travel time matrix of auto mode, Household Workplace Locations (From Household Survey)</td>
</tr>
</tbody>
</table>
C.2. Variable Descriptions

Zonal Size and Density

Size variables referred to the number or quantity of elementary alternatives (e.g., number of households, number of dwellings, etc.), and are specified inside a log operation in order to develop a model that is insensitive to the level of granularity (see chapter 2, 2.4.2). Density variables (e.g., number of Households per hectare) are also included as they might affect the attractiveness of zones.

Most of locational variables discussed here are acquired from census 2001 (see Table C-1). Census data are available at different level of aggregation (usually Census OAs) and have been aggregated to Travel Analysis Zones (TAZs). We have conducted two different approaches to aggregate census spatial variables to TAZ level. In the first approach, census data were associated with TAZs in accordance with the location of their centroids. Each Census OA was therefore associated with one and only one TAZ. However, this approach is prone to spatial aggregation bias especially where TAZs’ boundaries and Census OAs’ boundaries do not match. In the second approach, census data is aggregated by the proportion of the area that intersects the TAZs using the Proportional Overlap Tool of MapInfo.

Zonal Socioeconomic

The average socioeconomic attributes of households that live in a neighbourhood play an important role in attractiveness of a location, hence, in households’ residential location decisions. Variables such as zonal average household income, average household size, average household age, and percentage of different ethnic groups are likely to be influencing factors in households’ location preferences. This may lead to social or/and ethnic segregation in cities which has been studied by many authors (see Bayer et al., 2005).

Zonal Housing Type and Tenure Type Composition

The diversity of housing types and tenure types with in a zone might also determine the attractiveness of the zone. For example, zones with high percentage of detached housing might be more attractive for households with children, or high percentage of social housing with a zone might have a negative effect on the zonal attractiveness.
Zonal Land Use Structure

Certain type of land uses such as percentage of green spaces in a zone has a positive effect on the zonal attractiveness. Hence, diversity of land uses and the degree to which they are mixed might be a determinant factor in residential location choices. We have computed land use composition of zones (i.e., the percentage of each land use type) based on the Generalised Land Use Database (GLUD), 2001. Similar to Guo and Bhat (2007a), we have also calculated the land use mix diversity index as:

\[ I = 1 - \left( \frac{d^{1/3} + n^{1/3} + o^{1/3}}{L^{1/3}} \right) \]  

where,

- \( d \) is zonal area in domestic building use (residential),
- \( n \) is zonal area in non-domestic building (non-residential),
- \( o \) is zonal area in other land uses, and
- \( L = d + n + o \) is the zonal area.

Housing Prices

The average housing prices is definitely a determining factor in households’ location decisions. The housing prices data (acquired from Land Registry) include average paid prices for the year 2001 at the postal sector level (e.g., SW7 2). Unlike census variables, simple GIS aggregation cannot be applied to aggregated prices data to TAZ level because the level of granularity of TAZs and postal sectors are more or less the same (i.e., 879 TAZs and 889 postal sectors).

Due to the spatial nature of this problem, kriging method has been used to obtain average housing prices in the centroid of TAZs knowing housing prices in the centroid of postal sectors. Kriging is a method of spatial interpolation used in geostatistics in

---

44 Land use mix diversity index measures the diversity of land use types present within a zone and varies from 0 (homogenous land use) to 1 (most mixed land uses).
order to estimate the unobserved value of a location from observations of its values at nearby locations (Cressie, 1993).

Given observations \( z(x_1), \ldots, z(x_n) \) at locations \( x_1, \ldots, x_n \), (e.g., centroids of postal sectors) the kriging method can be used to estimate \( z(x) \) at some new location \( x \) (e.g., the centroid of a TAZ) based on the observed values. Kriging predicts the unobserved value using a leaner combination of \( z(x_1), \ldots, z(x_n) \), with weights chosen to minimise the variance of the prediction error. The kriging interpolation is conducted using mGstat geostatistical toolbox in Matlab (Hansen and Mosegaard, 2008). Figure C-1 and Figure C-2 depict average housing prices at centroids of postal sectors (observed) and TAZs (predicted).

![Figure C-1- Average Housing Prices at the Postal Sector Level (i.e., Observed Points)](image-url)

Figure C-1- Average Housing Prices at the Postal Sector Level (i.e., Observed Points)
School Quality

Availability of good schools in an area is an important factor for households’ location decisions, especially for households with children (Clark et al., 2006). This study quantifies zonal school quality as the average point score of primary school pupils within a zone. The average point scores for primary schools are published by the department for education. Average point scores have been geocoded based on the coordinates of primary schools derived from the 2005 educational establishment dataset. The zonal averages are then calculated in order to quantify the school quality of a zone (see Figure C-3). This approach might be prone to biases as the effects of nearby schools are not considered for the school quality of a zone. Also, the quality of secondary schools has not been incorporated in the zonal school quality due to the lack of spatial data on secondary schools. Exploring these issues remains for the future work of this study.
Transportation Network

The transport network of a city has both accessibility and environmental impacts which can affect households’ residential location choice decisions. The transportation network variables quantify the local measures of public transportation and auto service levels of location.

OS MasterMap ITN layer has been used to compute the density of dual carriageway and other roadway density variables. We have computed the distances of zonal centroids to the nearest public transport access points (i.e., Tube, DLR and Overground stations) in order to quantify the proximity of zones to public transit. The coordinates of public transport access points have been derived from the National Public Transport Access Node (NaPTAN) dataset. We have also quantified zonal transit proximity as the summation of inverse distances from all stations to the centroid of a zone as:

\[
A_{i}^{\text{Transit}} = \sum_{j=1}^{N} \frac{1}{d_{ij}}
\]  

(C-2)
where, $d_{ij}$ is the distance between the centroid of zone $i$ to station $j$, and $N$ is the total number of stations.

**Accessibility**

Accessibility measures quantify the potential of reaching spatially distributed opportunities (e.g., employment opportunity, shopping opportunity, etc.) which is an important factor in households’ location preferences. A number of different accessibility measures have been proposed in the literature (see Páez et al., 2012), here we adopt gravity-type accessibility measures initially proposed by Hansen (1959). The following accessibility variables have been computed and included in the model specification:

**Accessibility to Employment:**

$$A_{i}^{Emp} = \frac{1}{N} \sum_{j=1}^{N} \frac{E_j}{TT_{ij}}$$  \hspace{1cm} \text{(C-3)}

**Accessibility to Shopping:**

$$A_{i}^{Shop} = \frac{1}{N} \sum_{j=1}^{N} \frac{S_j}{TT_{ij}}$$  \hspace{1cm} \text{(C-4)}

**Accessibility to Recreational Facilities:**

$$A_{i}^{Rec} = \frac{1}{N} \sum_{j=1}^{N} \frac{R_j}{TT_{ij}}$$  \hspace{1cm} \text{(C-5)}

where,

$E_j$ is total number of employees,

$S_j$ is number of employees in retail trade sector,

$R_j$ is number of employees in recreational, cultural and sporting activities sector, and

$TT_{ij}$ is the travel time between zone $i$ and zone $j$.

Zone-to-zone travel time matrix of auto mode and Annual Business Inquiry (ABI) data have been used to compute these measures. ABI data are available at Census Super Output Areas and have been aggregated to the TAZ level.
**Commute Variable**

The commute variable (i.e., travel time from households’ workplace location zone(s) to residential zones) is the most important variable describing household locational preferences. This variable is computed using zone-to-zone travel time matrix of auto mode acquired from London Transportation Studies (LTS) Model assuming that the households’ employment location choices are predetermined and exogenous to residential location choices. Therefore, locations of households’ main activities (At least for primary workers) such as workplace or school locations are also needed to calculate this variable. Therefore, a household survey that provides information on both residential and workplace locations is required in order to estimate a residential location choice model incorporating the commute variable. For households with more than one worker, the aggregate values across all workers in the household can be calculated.
APPENDIX D

LIST OF ACRONYMS AND ABBREVIATIONS

CDF: Cumulative Distribution Function
CNL: Cross Nested Logit
GEV: Generalised Extreme Value
GNL: Generalised Nested Logit
IIA: Independence from Irrelevant Alternatives
IPF: Iterative Proportional Fitting
MACML: Maximum Approximate Composite Marginal Likelihood
MAUP: Modifiable Areal Unit Problem
ML: Maximum Likelihood
MMNL: Mixed Multinomial Logit
MGEV: Mixed GEV
MNL: Multinomial Logit
MNP: Multinomial Probit
MSL: Maximum Simulated Likelihood
NL: Nested Logit
PCL: Paired Combinatorial Logit
RP: Revealed Preference
RUM: Random Utility Model
SAR: Spatial Auto-Regressive
SP: Stated Preference
TAZ: Travel Analysis Zones
APPENDIX E

PUBLICATIONS

E.1. Journal Articles


E.2. Conference Papers

Zolfaghari A, Sivakumar A, Polak JW, Choice Set Pruning in Residential Location Choice Modelling, Annual UTSG meeting, Milton Keynes, UK, 2011


Zolfaghari A, Sivakumar A, Polak JW, Methodological and Empirical Challenges in Modelling Residential Location Choices at the Level of Dwelling Units, International Choice Modelling Conference (ICMC), Sydney, Australia, 2013