

A Longitudinal Study on the Linkage between Public Transport Demand and Land Use Characteristics: A Pseudo Panel Approach

Chi-Hong Tsai

A thesis submitted in partial fulfilment of the requirements for the degree of
Doctor of Philosophy in the Business School, University of Sydney, Australia

Institute of Transport and Logistics Studies
The University of Sydney Business School
The University of Sydney
NSW 2006 Australia

May 2013

Abstract

This study applies a pseudo panel approach to analyse public transport demand in the Sydney Greater Metropolitan Area (SGMA). A public transport demand model is constructed to incorporate two factors that have been highlighted in the literature of travel behaviour but still under-researched, which are: (i) the temporal effect of demand adjustment; and (ii) the land use characteristics of the built environment. The research gaps in previous applied pseudo panel data research including estimation techniques and issues involved with the applications to public transport are identified and addressed in this study.

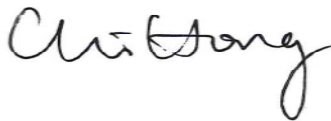
The pseudo panel approach allows for the identification of long-term demand changes using repeated cross-sectional data, which are collected at an individual level with detailed travel-related information and geographical information. This study constructs static and dynamic pseudo panel data models to analyse public transport demand in terms of its associations with price, socio-economic factors, level of public transport service, and land use factors. The research findings identify the significant determinants of public transport demand in the SGMA, with a distinction between short-run and long-run demand elasticities. This suggests a timeframe of 2.13 years is required to reach the long-run demand equilibrium. The estimated demand elasticities are used to forecast demand for the SGMA with validated results supporting the applicability of the public transport model based on the pseudo panel data.

The main contribution of this thesis is the identification of long-run public transport demand elasticities using a pseudo panel dataset created from existing repeated cross-sectional household travel survey data which uses more individual information than aggregate data. This approach enables a longitudinal analysis in the absence of genuine panel data, and this in turn provides important implications for urban public transport planning and policy formulation.

Statement of Originality

This is to certify that to the best of my knowledge, the content of this thesis is my own work. This thesis has not been submitted for any degree or other purposes.

I certify that the intellectual content of this thesis is the product of my own work and that all the assistance received in preparing this thesis and sources have been acknowledged.

A handwritten signature in cursive script, appearing to read "Chi-Hong Tsai".

Chi-Hong Tsai

Acknowledgements

This thesis could not be completed without enormous contribution from many people. The most important person I would like to thank is my supervisor Professor Corinne Mulley. Corinne has been very patient and supportive and providing high-quality supervision since the beginning of my PhD program. She has been guiding me to develop my independent research ability and has been giving very detailed and insightful comment on my work. I am deeply grateful for her input not only on this thesis but also on the development of my research career. I do realise how lucky I am to be Corinne's PhD student.

I would also like to thank my associate supervisor Dr Geoffrey Clifton and my former associate supervisor Dr Rhonda Daniels. Geoffrey's local knowledge of Sydney and strong statistical background made great contribution to the final stage of this study after Rhonda left the University of Sydney. Rhonda gave her professional experience in local transport planning and policy to build up the foundation of this study in the first two years, which was very helpful for me as a new comer of Sydney.

I also need to thank the academic staff from where I study: Institute of Transport and Logistics Studies. This is a great academic environment which encourages me to develop my research ideas and motivate me to pursue to the next level of research. Special thanks to Professor David Hensher, Professor Peter Stopher, Professor John Rose, Dr Stephen Greaves in giving valuable comments and suggestions for this thesis.

The PhD students from my institute undoubtedly have made great contribution to this work. I have to specially acknowledge Waiyan Leong who jointly conducted Monte Carlo experiment (Chapter 5) with me and shared his knowledge and research ideas when we were in the same office. I also need to thank Chinh, Alejandro, Adrian and Richard who have given their support on the analytical tools used in this study. It has been a great journey with all my PhD peers here.

I would also like to acknowledge Bureau of Transport Statistics in providing the Sydney Household Travel Survey data and land use data for this study. They have been processing my data requests very efficiently and professionally. Special thanks to Tim, Michael, Annette, Houshang, and Evelyn for their technical support.

Last but not least, I would not be here completing my PhD thesis without my family's support. They are always very supportive even though we do not live in the same country in these three years. Thank you for guiding to me to achieve this milestone. I am proud of being a member of my family and I hope you will be proud of me in the future.

Table of Contents

List of Figures	x
List of Tables	xii
List of Acronyms	xiv
Notational Glossary	xv
CHAPTER 1 INTRODUCTION	1
1.1 Background and research questions	1
1.2 Research approach.....	4
1.3 Thesis contributions	6
1.4 Thesis outline.....	8
CHAPTER 2 LITERATURE REVIEW	11
2.1 Public transport demand elasticity.....	11
2.1.1 Determinants of demand elasticity	14
2.1.2 Short-run and long-run elasticity.....	15
2.2 The relationship between land use and travel behaviour.....	17
2.2.1 Evidence from previous studies	17
2.2.2 Land use factors in public transport demand modelling.....	20
2.2.3 Self-selection of location choice.....	21
2.3 Pseudo panel data.....	22
2.3.1 Background of pseudo panel data.....	22
2.3.2 Principles of pseudo panel data construction.....	24
2.3.3 Pseudo panel data in transport literature	25
2.3.4 Issues of application to public transport.....	28
2.4 Panel data model estimation.....	29
2.4.1 Static genuine panel data model and estimation	30
2.4.2 Dynamic panel data model and estimation.....	33

2.4.3 Pseudo panel data model and estimation	39
2.5 Diagnostics and correction of panel data model assumptions	43
2.6 Research gaps and summary	48
CHAPTER 3 DESCRIPTION OF CASE STUDY	51
3.1 Introduction	51
3.2 The Sydney Greater Metropolitan Area	51
3.2.1 Demographics and geography.....	51
3.2.2 Public transport in the Sydney Greater Metropolitan Area	53
3.3 Data description.....	56
3.3.1 Data sources	56
3.3.2 Definitions of variables	57
3.4 Exploratory analysis.....	64
3.4.1 Introduction of Geographically Weighted Regression	66
3.4.2 Global model estimation	68
3.4.3 Local model estimation	73
3.5 Summary.....	79
CHAPTER 4 PSEUDO PANEL DATA APPROACH	82
4.1 Introduction	82
4.2 Grouping criteria for pseudo panel data.....	82
4.3 Pseudo panel data construction	87
4.3.1 Forming the cohorts	87
4.3.2 Variables in the pseudo panel dataset	90
4.4 Preliminary analysis.....	93
4.4.1 Group-specific effects	93
4.4.2 Historical trends of variables by groups	98
4.5 Pseudo panel data model.....	101
4.6 Summary.....	104

CHAPTER 5	MONTE CARLO SIMULATION.....	106
5.1	Introduction	106
5.2	Experiment design.....	108
5.3	Estimators and performance measurements.....	111
5.4	Analysis of Monte Carlo simulation.....	112
5.4.1	Simulation results for static models.....	112
5.4.2	Simulation results for dynamic models.....	114
5.5	Investigation of correlation	120
5.6	Summary.....	122
CHAPTER 6	STATIC MODEL ESTIMATION	125
6.1	Introduction	125
6.2	Static public transport demand model.....	125
6.2.1	Functional forms	125
6.2.2	Descriptive statistics.....	129
6.2.3	Estimation techniques	132
6.3	Estimation results	133
6.3.1	Base model.....	133
6.3.2	Test of functional forms	136
6.3.3	Model Diagnostics	138
6.3.4	Comparison of estimation techniques	144
6.4	Summary.....	147
CHAPTER 7	DYNAMIC MODEL ESTIMATION.....	149
7.1	Introduction	149
7.2	Model specifications.....	149
7.2.1	Dynamic models	149
7.2.2	Estimation techniques	153
7.3	Estimation results of dynamic pseudo panel data models.....	154

7.3.1	Base model.....	154
7.3.2	Test of functional forms	157
7.3.3	Model diagnostics	163
7.3.4	Comparison of estimation techniques	164
7.4	Estimation of demand elasticities.....	167
7.5	Summary and discussion.....	174
CHAPTER 8 DEMAND FORECAST		176
8.1	Introduction	176
8.2	Model validation	176
8.3	Projection of predictors.....	179
8.4	Public transport demand forecast.....	182
8.5	Sensitivity analysis.....	184
8.6	Summary.....	189
CHAPTER 9 CONCLUSIONS		191
9.1	Summary of research findings	191
9.2	Research contributions	194
9.2.1	Contributions to the literature and research methodology	194
9.2.2	Contributions to practical urban and transport planning.....	195
9.3	Limitations of this study	196
9.4	Directions for future research	198
9.5	Concluding remarks.....	200
REFERENCES.....		202
APPENDICES.....		211
Appendix 1		211
Appendix 2.....		212
Appendix 3.....		214

List of Figures

Figure 1.1 A Summary of Research Approaches to Public Transport Demand Modelling.....	5
Figure 3.1 The Sydney Greater Metropolitan Area	52
Figure 3.2 Extended Railway Lines of CityRail after 1997	55
Figure 3.3 Pseudo Nodes in a Cul-de-sac Built Environment	62
Figure 3.4 Pseudo Nodes in a Built Environment with a Grid Road Network	62
Figure 3.5 Adaptive Kernels in Local Model Estimation.....	68
Figure 3.6 Bandwidth of a Kernel.....	68
Figure 3.7 Location of the Sydney Urban Area	74
Figure 3.8 Map of the Local Model Estimates of Price in the Sydney Urban Area	75
Figure 3.9 Map of the Local Model Estimates of Pseudo Nodes in the Sydney Urban Area.....	76
Figure 3.10 Map of the Local Model Estimates of Bus Frequency in the Sydney Urban Area.....	77
Figure 3.11 Map of the Local Model Estimates of Distance to CBD in the Sydney Urban Area.....	78
Figure 3.12 The Residuals of the Local Model Estimation in the Sydney Urban Area	79
Figure 4.1 Train and Bus Mode Share by Age Group in 2009/10.....	85
Figure 4.2 Average Public Transport Trips by Travel Zones.....	86
Figure 4.3 Moving-averaged Population Density versus Population Density in 2006	92
Figure 4.4 Number of Public Transport Trips by Age for Different Birth Year Groups	96
Figure 4.5 Box Plot of Number of Public Transport Trips by Birth Year Groups	97
Figure 4.6 Box Plot of Number of Public Transport Trips of Distance-to-CBD Groups	97
Figure 4.7 Time Trends of Number of Public Transport Trips from 1997-2009 by Group.....	99

Figure 4.8 Time Trends of Public Transport Trip Price from 1997-2009 by Groups	100
Figure 4.9 Time Trends of Personal Income from 1997-2009 by Groups	101
Figure 5.1 Density Plots of λ_1 Estimates from Pooled OLS and FE Estimators	117
Figure 5.2 Density Plots of β_1 Estimates from Pooled OLS and FE Estimators	117
Figure 5.3 Density Plots of λ_1 Estimates from Pooled OLS and FE Estimators	118
Figure 5.4 Density Plots of β_1 Estimates from Pooled OLS and FE Estimators	118
Figure 6.1 Scatter plot of Residuals and Fitted Values	139
Figure 6.2 Scatter Plot of Residuals and Price	140
Figure 6.3 Scatter Plot of Residuals and Income	141
Figure 6.4 Scatter plot of Residuals and Age	142
Figure 6.5 Scatter Plot of Residuals and Bus Frequency.....	142
Figure 6.6 Scatter Plot of Residuals and Population Density	143
Figure 6.7 Scatter plot of Residuals and Pseudo Nodes.....	144
Figure 7.1 Scatter Plot of Residuals and Fitted Values from the Dynamic Model	163
Figure 7.2 Scatter Plot of Residuals and the Lagged Variable from the Dynamic Model	164
Figure 7.3 The Speed of Public Transport Demand Adjustments	168
Figure 8.1 A Built Environment with Lower Population Density as a Baseline	186
Figure 8.2 A Built Environment with Higher Population Density for	187
Figure 8.3 A Built Environment with more Pseudo Nodes as a Baseline.....	188
Figure 8.4 A Built Environment with Fewer Pseudo Nodes for Sensitivity Analysis	188
Figure 8.5 A comparison of Public Transport Demand Growth of all Scenarios	189

List of Tables

Table 2.1 Summary of Selected Public Transport Demand Elasticity Studies.....	12
Table 2.2 Categories of Explanatory Variables in Demand Models (Number of studies using these variables)	14
Table 2.3 The Measurement of Land Use Factors from Selected Previous Studies	18
Table 2.4 Weighted Average Elasticity of Public Transport Demand with Land Use Factors	19
Table 2.5 Selected Studies on Public Transport Demand Elasticities with respect to Land Use Measures	20
Table 2.6 A Comparison of Genuine Panels, Repeated Cross-sectional Data, and Pseudo Panels	24
Table 2.7 A Comparison of Selected Previous Pseudo Panel Data Studies.....	25
Table 2.8 Estimation Techniques in Previous Applied Pseudo Panel Studies.....	42
Table 2.9 Panel Data Model Assumption Tests.....	45
Table 3.1 Demographics of the Sydney Greater Metropolitan Area.....	53
Table 3.2 Statistics of Trip Modes in the SGMA in 2010/2011	54
Table 3.3 CityRail Network Changes since 1997	55
Table 3.4 Ticket Journey Multipliers.....	59
Table 3.5 Summary and Descriptive Statistics of Variables	64
Table 3.6 Descriptive Statistics of Variables	69
Table 3.7 Correlation Matrix.....	70
Table 3.8 Global Model Estimation Results	71
Table 3.9 Average Elasticity to Public Transport Demand	73
Table 3.10 Results of the Monte Carlo Test for Spatial Variability	74
Table 4.1 Historical Statistics of Public Transport Trips from.....	83
Table 4.2 An Evaluation of Grouping Criteria	84
Table 4.3 A Comparison between Two Different Pseudo Panel Datasets	89
Table 4.4 Results of Pseudo Panel Construction	90
Table 4.5 A Summary of Variables in the Pseudo Panel Dataset	93
Table 4.6 Between-group and Within-group Variances of all Variables	95
Table 5.1 Scenario Design for Monte Carlo Experiments	109

Table 5.2 Summary of Estimator Properties	111
Table 5.3 Simulation Results for Static Models	113
Table 5.4 Simulation Results (Scenarios 1 – 4) for Dynamic Models	115
Table 5.5 Simulation Results (Scenarios 3/5/6) for Dynamic Models	120
Table 5.6 of Scenario 3 and Scenario 7 in the Static Model.....	121
Table 5.7 Comparisons of Scenario 3 and Scenario 7 in the Dynamic Model	122
Table 6.1 Descriptive Statistics of Variables	130
Table 6.2 Correlation Matrix of Variables in the Pseudo Panel Dataset	132
Table 6.3 Pooled OLS Estimation Results of Static Linear Model (Base Model)	134
Table 6.4 Comparison of Static Model Functional Forms.....	137
Table 6.5 A Comparison of Static Model Estimation Results with Various Estimators.....	146
Table 7.1 Pooled OLS Estimation Results of the Linear Dynamic Model (Base Model).....	155
Table 7.2 A Comparison of Pooled OLS Estimation Results between the Static and Dynamic Base Models	157
Table 7.3 Evaluation of Dynamic Model Functional forms.....	158
Table 7.4 A Comparison of the Best Static and Dynamic Functional Forms.....	159
Table 7.5 Test of Dynamic Model Specifications	161
Table 7.6 Dynamic Model Estimation Results using Various Estimators	165
Table 7.7 Demand Elasticities Derived from the	169
Table 7.8 Descriptive Statistics of Low Income and High Income Cohorts Divided by the Median Income	172
Table 7.9 Estimation Results classified by Personal Income.....	172
Table 8.1 Model Estimation Results Using Data from 1997-2009 and 1997-2007	178
Table 8.2 Predicted Public Transport Demand for 2008 and 2009.....	179
Table 8.3 Projections of Predictors for Demand Forecasting.....	180
Table 8.4 Results of Annual Public Transport Demand Forecast	183
Table 8.5 Sensitivity Analysis of Public Transport Demand Forecasting.....	185

List of Acronyms

ABS: Australian Bureau of Statistics
BTS: Bureau of Transport Statistics
CBD: Central Business District
CD: Census District
CPI: Consumer Price Index
ECM: Error Correction Model
FE: Fixed Effect
FGLS: Feasible Generalized Least Squares
GIS: Geographical Information System
GLS: Generalised Least Squares
GMM: Generalised Method of Moments
GWR: Geographically Weighted Regression
IV: Instrumental Variable
LGA: Local Government Area
LSDV: Least Squares Dummy Variable
NSW: New South Wales
OLS: Ordinary Least Squares
PAM: Partial Adjustment Model
PCSE: Panel-Corrected Standard Error
RE: Random Effect
RESET: Regression Specification Error Test
RMSE: Root Mean Square Error
SD: Statistical Division
SGMA: Sydney Greater Metropolitan Area
SHTS: Sydney Household Travel Survey
SP: Stated Preference
SSD: Statistical Subdivision
TZ: Travel Zone
VIF: Variance Inflation Factor
WLS: Weighted Least Squares

Notational Glossary

Variable/ Parameter	Description	Units
A	proportion of demand adjustment	
C	Total number of cohorts in the pseudo panel dataset	
$\bar{D}_{g,t}$	Average public transport trips per person per day for group g at time period t	Trips
$\bar{D}_{g,t-1}$	Average public transport trips per person per day for group g at time period $t-1$	Trips
d_i	Distance between the i th observation and the location (u_j, v_j) in GWR analysis	
E_i'	A vector of socio-economic variables	
e_k	Demand elasticity of variable k	
G	Number of groups in the pseudo panel dataset	
h	Bandwidth of the kernel in GWR analysis	
L_i'	A vector of land use variables	
M_i	Proportion of land use type i in a Travel Zone	percentage
n_c	Cohort size	individuals
P_i	Public transport trip price	AUD
Q	Total number of land use types	
S_i	Quality of service (bus frequency)	services
T	number of years for A percent of demand to adjust	
w_i	Geographical weight for an observation i in GWR analysis	
(u_i, v_i)	Geographical coordinates	
u_{it}	Composite error term in a general panel data model for a unit i at time period t	
\bar{x}_{gt}	Independent variable in a pseudo panel data model for a group g at time period t	
x_{it}	Independent variable in a general panel data model for a unit i at time period t	
x_{it-1}	Lagged independent variable in a general panel data model for a unit i at time period $t - 1$	
\bar{y}_{gt}	Dependent variable in a pseudo panel data model for a group g at time period t	
y_{it}	Dependent variable in a general panel data model for a unit i at time period t	
y_{it-1}	Lagged dependent variable in a general panel data model for a unit i at time period $t - 1$	

z_{it}	Instrument in the Instrumental Variable estimation	
$\bar{\alpha}_{gt}$	Unobserved group effect in a pseudo panel data model	
α_i	Unobserved group effect in a genuine panel data model	
β_0	Parameter of constant in a general panel data model	
β_1	Parameter of x_{it} in a general panel data model	
β_2	Parameter of the lagged independent variable	
λ	Parameter of the lagged dependent variable	
ε_{it}	Independent identically distributed error term in a panel data model	
$\bar{\omega}_{gt}$	Time-varying cohort effect	
$\sigma_{\bar{\alpha}}^2$	Variance of unobserved group effect	
$\sigma_{B,x}^2$	Between-group variance of the exogenous variable	
$\sigma_{W,x}^2$	Within-group variance of the exogenous variable	

CHAPTER 1 INTRODUCTION

This thesis studies public transport demand in the Sydney Greater Metropolitan Area (SGMA) incorporating various measures of land use characteristics using a pseudo panel approach to take the temporal effect of travel demand changes into account. This chapter first introduces the research background and identifies the research questions addressed in this study. The framework of research approaches is presented next in Section 1.2. Section 1.3 highlights the key contributions of this study in terms of contributions to the literature and research methodology as well as the contribution to practice in urban transport planning and policy implementation. Finally, this chapter describes the structure of the thesis in Section 1.4.

1.1 Background and research questions

The importance of public transport systems has been receiving substantial attention in the urban and transport planning sectors. Understanding public transport demand in terms of the associations between public transport use and its determinants provides important information for transport policy formulation and implementation. The associations between public transport demand and its determinants can be investigated through public transport demand models, which have been extensively researched in transport literature. However, although widely discussed, there are some components which have not been fully incorporated in previous public transport demand studies. One is the temporal effect of travel demand adjustment which has not been captured by conventional static demand models. The other is the integration of public transport demand with a comprehensive set of land use variables, despite the strong connections established in the literature of travel behaviour and built environment. This study focuses on these two elements of travel demand determinants which are under-researched in the literature, using the SGMA as a case study.

The temporal effect of demand adjustment, also referred as the dynamics of travel demand, relates to the way in which travellers do not tend to adjust their travel behaviour instantly in response to transport system changes. The reasons

for the lagged demand adjustments may come from travellers' habits, imperfect information, or changes of residential or workplace locations (Dargay and Hanly, 2002), which in turn leads to the differences in the short-run and long-run demand elasticities. International evidence has suggested that this temporal effect is significant in the determination of public transport demand where this has come from comparing multiple state-wide or worldwide transport systems at an aggregate level (Bresson et al., 2003, Graham et al., 2009, Souche, 2010). The investigation of the temporal effect would be of more benefit to urban transport planning if the magnitude of short-run and long-run demand elasticities could be identified for a single transport system or a specific study area, but this research is not common in the literature because it relies on longitudinal data, normally collected from continuous travel surveys, which are difficult to conduct at a large scale either in space or in time. Thus, although lagged demand adjustment has been suggested to be significant in travel demand analysis, there is a need to demonstrate this in public transport demand models to provide robust evidence for transport planners who typically only have information of short-run demand changes. This context provides the background for research questions in study which are defined as follows:

Question 1: *What are the determinants of public transport demand and the demand elasticities with respect to each of the determinants in the SGMA?*

Question 2: *Is the temporal effect of public transport demand significant in the SGMA? What are the short-run and long-run demand elasticities if the temporal effect is significant?*

The other element of public transport demand modelling addressed in this study is the connections between public transport demand and land use. The association between travel behaviour and land use characteristics of a built environment has been identified and widely recognised in transport research. However, conventional public transport demand models tend to only take account of land use density (Dargay et al., 2010, Souche, 2010). Other land use variables such as land use diversity, urban design, and accessibility to public transport

services have not been fully incorporated in public transport demand models, although their impact on travel behaviour has been demonstrated (Cervero and Kockelman, 1997). Ignoring these other factors of land use characteristics may not only under-estimate the influence of land use characteristics on public transport demand, but also restrict urban and transport planners in their use of strategic land use planning to increase public transport usage. To investigate the associations between public transport demand and land use factors in the context of the SGMA, the following research question is analysed in this study:

***Question 3:** What are the magnitudes of the impact of land use density, diversity, design, and accessibility on public transport demand in the SGMA?*

In summary, this study constructs a public transport demand model incorporating the temporal effect and land use characteristics. The demand model is estimated to identify the relationship between public transport demand and its determinants, and is used to forecast public transport demand for the SGMA. The research approach is summarised in the next section.

1.2 Research approach

The general research approaches to public transport demand modelling are summarised in

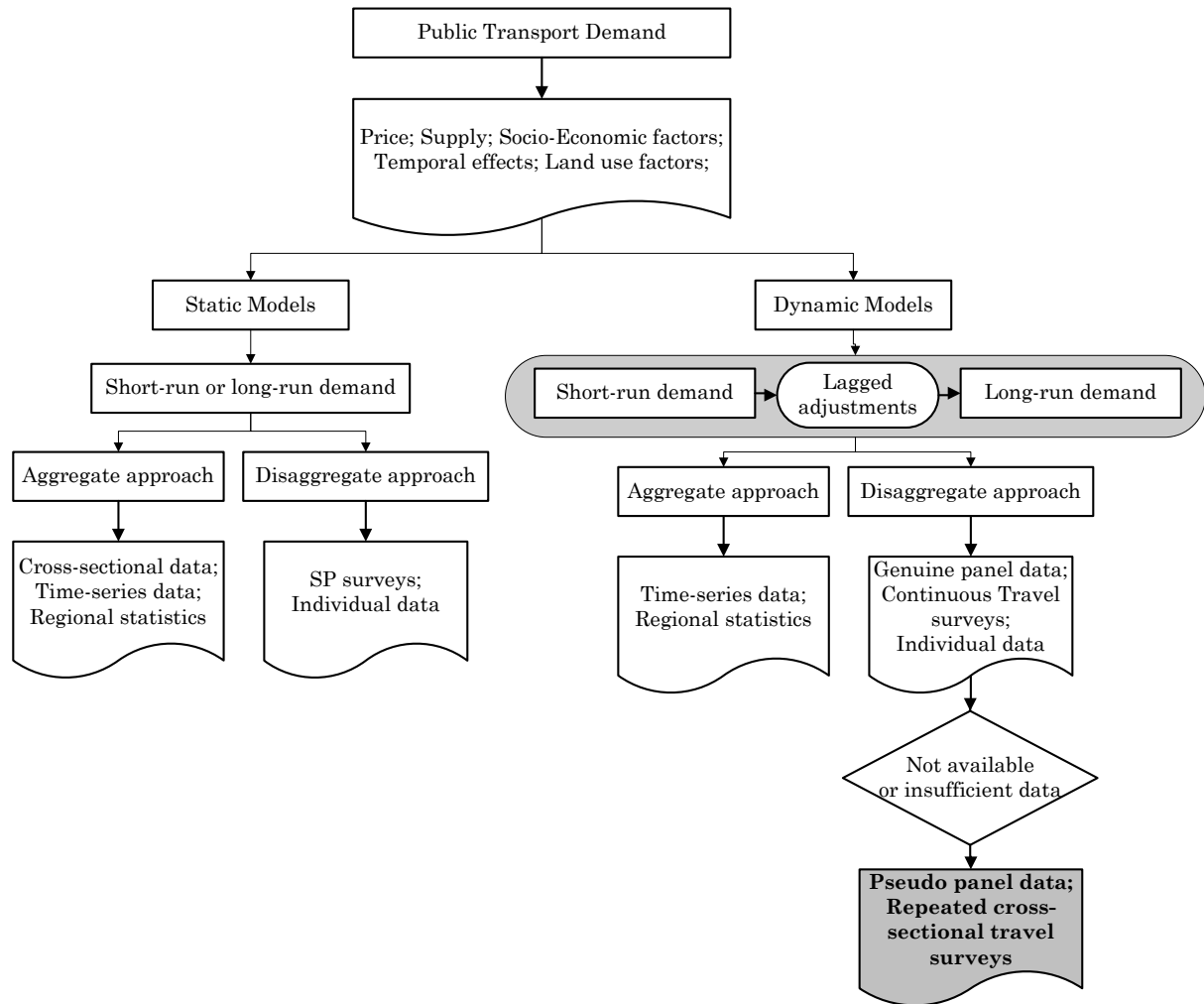


Figure 1.1. Public transport demand is a function of various factors including price, supply, socio-economic variables, and the two elements highlighted in the previous section: the temporal effect and land use factors. The existence of the temporal effect leads to a difference between short-run and long-run demand as a result of the lagged demand adjustment. The public transport demand model can be specified in a static or dynamic form, and the data used for model estimation are categorised into aggregate and disaggregate data. When the static model is employed, either short-run demand or long-run demand can be estimated depending on the type of data in use.

Short-run demand can be analysed using cross-sectional data based on regional statistics at an aggregate level, or using Stated-Preference Survey (SP) data or

combined SP and Reveal-Preference Survey data (RP) to retrieve individual information, whereas long-run demand requires time-series data with the status of long-run equilibrium being assumed in a static model estimation. However, Goodwin (1992) suggested that the assumption of long-run equilibrium for static model calibration is subject to model specification errors and thus a dynamic model is preferred for time-series data.

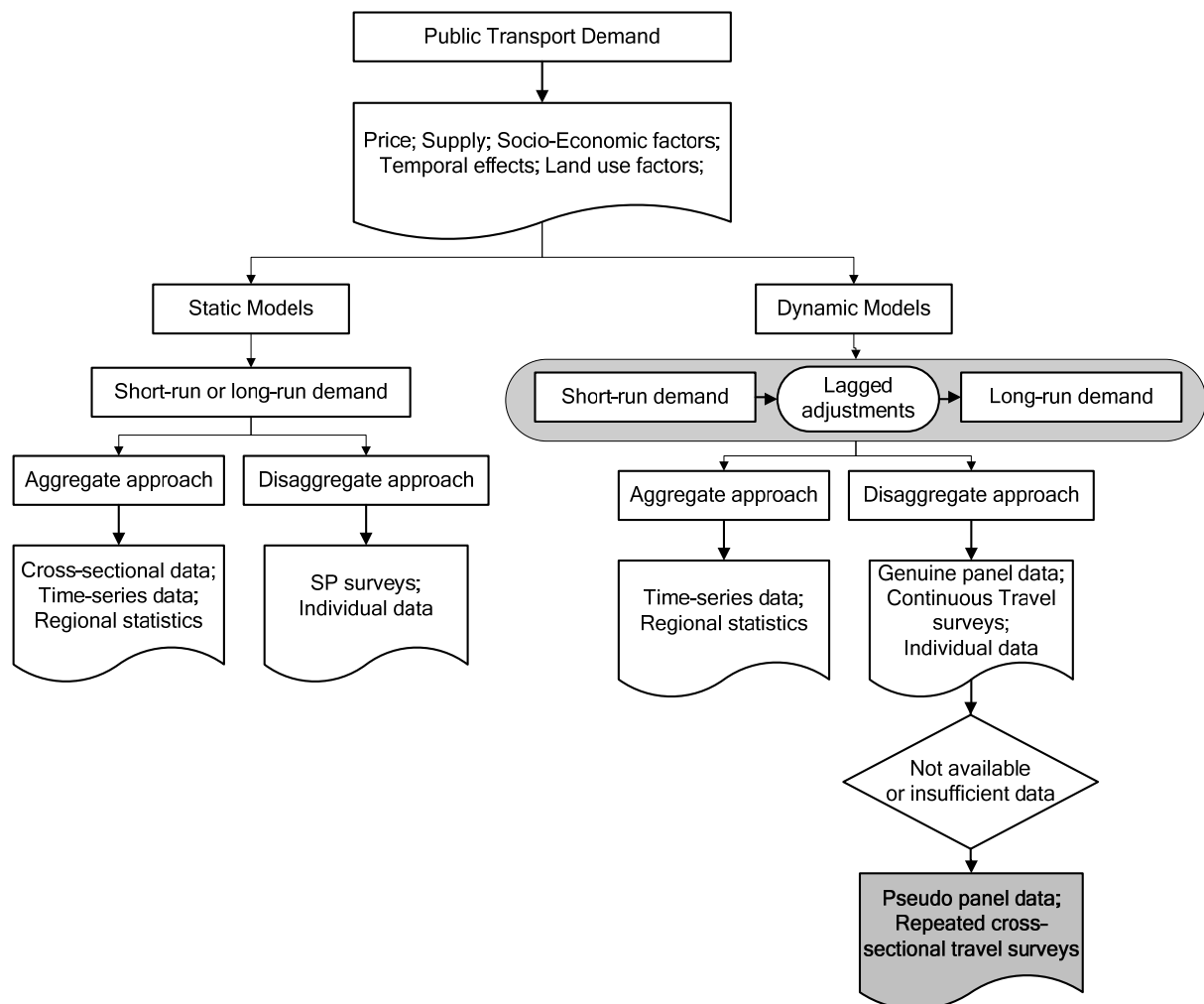


Figure 1.1 A Summary of Research Approaches to Public Transport Demand Modelling

A dynamic demand model includes lagged variables to capture the lagged demand adjustment as a result of the temporal effect of travel behaviour, which is unable to be controlled in a static model. In terms of the data used for dynamic model estimation, previous studies usually used aggregate time-series data from regional statistics of several study areas to give panel data analysis with sufficient sample observations and time scales. The drawback of this approach is

that aggregate data do not provide individual travel information and the research results are not representative of a specific study area. To incorporate individual travel information in travel demand models for a specific study area, the ideal approach is to use panel data based on individual records, usually collected from continuous household travel surveys over time, which is known as genuine panel data. However, genuine panel data are not commonly available in transport due to their high costs and sample attrition problems, and thus a disaggregate approach is not common in longitudinal public transport demand analysis. As compared to genuine panel data, repeated cross-sectional travel surveys are more commonly available in some countries but the shortcoming of the repeated cross-sectional data is that individuals are not traced over time. Instead, new samples are drawn in each wave of the survey although this approach provides more sufficient and consistent survey data it is not possible to trace travel behaviour change. However, a pseudo panel approach does enable panel data analysis using repeated cross-sectional data (Deaton, 1985), with results being able to distinguish the short-run and long-run demand by employing a dynamic pseudo panel data model.

The pseudo panel approach has been increasingly applied to travel demand studies because of its sound theoretical grounding and its ability to accommodate the dynamics of travel behaviour using individual travel survey data. However, some issues related to applied pseudo panel research still remain under-researched and require further investigation, such as the construction of pseudo panel data for limited sample observations and the estimation techniques for dynamic pseudo panel data models. These issues are discussed and addressed in this study.

1.3 Thesis contributions

In this section, the contributions of this study are classified into scientific contributions and practical contributions. Scientific contributions include new knowledge added to the relevant literature that has not been fully addressed in the past and refinements of research methodology in the research field which collectively benefits future research. Practical contributions are research findings

that are practice-ready which can be applied to urban transport planning with policy implications. The specific contributions from these two perspectives are highlighted as follows.

Scientific contribution to the literature and research methodology:

- This thesis incorporates a comprehensive set of land use variables including land use density, diversity, design, and accessibility in a public transport demand model, together with price variable and other determinants as identified in the literature of public transport demand modelling (Chapter 3).
- This thesis introduces the principles and process of pseudo panel data constructions which have not been comprehensively reviewed in the literature (Chapter 4).
- This thesis applies the pseudo panel data approach to public transport demand analysis and address the research issues that constrain its application. The application of a pseudo panel data approach to public transport has not yet been evident in the literature (Chapter 4).
- This thesis examines estimation techniques for pseudo panel data models through a Monte Carlo simulation experiment. The issue of pseudo panel data estimation has been discussed in the literature but the results are still inconclusive with different research findings from different studies. The Monte Carlo experiment suggests an appropriate estimator to estimate pseudo panel data and provides a guideline for future applied pseudo panel research (Chapter 5).
- This thesis employs a dynamic Partial Adjustment Model to investigate the temporal effect of travel demand public transport demand using the constructed pseudo panel dataset. This closes the research gap in public transport demand modelling which conventionally was unable to incorporate the temporal effect for a specific study area due to the unavailability of genuine panel data (Chapter 7).

Practical contributions to urban transport planning and policy

- This thesis shows the importance of taking account of spatial variation in the analysis of associations between public transport demand and its

determinants which provide policy implications for urban and transport planning (Chapter 3).

- This thesis uses the currently-available repeated cross-sectional household travel surveys to conduct a longitudinal study with individual information incorporated (Chapter 4).
- This thesis estimates the short-run and long-run public transport demand elasticities showing the significance of lagged demand adjustments which is critical to long-term transport planning (Chapter 7).
- This thesis estimates the short-term and long-term impacts of land use characteristics on public transport demand and shows the differences of land use elements to provide evidence for long-term urban planning (Chapter 7).
- This thesis forecasts public transport demand for the study area using the dynamic demand model with sensitivity analysis to forecast public transport demand under different policy scenarios (Chapter 8)

1.4 Thesis outline

This thesis is organised in nine chapters. The present chapter has introduced the background to the research and the research questions to be addressed in this study. It outlines the framework of research approach and the thesis contribution. The contents of the following chapters in this thesis are outlined in the rest of this section.

Chapter 2 reviews the literature for each of the components of this study. The literature of public transport demand modelling and studies on the relationship between travel behaviour and land use are first discussed. This discussion points to the way in which a pseudo panel approach could be applied to address the research questions. The pseudo panel approach is then comprehensively reviewed from its theoretical development, modelling and estimation issues, and the applications in travel demand analysis. This review identifies the research gaps in the pseudo panel literature which are accommodated in the following chapters.

Chapter 3 introduces the study area of this research and the data sources comprising travel-related and land use data for this study. The two domains of data are integrated in this chapter with an exploratory analysis on the relationship between public transport demand and land use characteristics. This analysis uses Geographically Weighted Regression based on the pooled Sydney Household Travel Survey data from 1997 to 2009. This exploratory analysis validates the selected explanatory variables of public transport demand and provides suggestions for pseudo panel data construction in Chapter 4.

Chapter 4 presents the pseudo panel data approach for this study. This chapter details the process and principles of constructing the pseudo panel dataset, and specifies the general form of the pseudo panel data model. The estimation techniques for pseudo panel data models are also discussed which highlights the need of examining the performance of various estimators for pseudo panel data estimation.

Chapter 5 conducts a Monte Carlo simulation experiment to examine the estimation techniques developed for genuine panel data estimation which are commonly applied to pseudo panel data models. The simulation results provide insights into the estimation bias and efficiency for each of the panel data estimators with their potential causes under various data properties. This experiment in turn suggests guidelines for estimating pseudo panel data models in Chapter 6 and 7.

Chapter 6 and Chapter 7 present the static and dynamic pseudo panel data models respectively. These two chapters specify the public transport demand model in different functional forms. The estimation results are evaluated to justify the best public transport demand model. The suggested demand model is then used to estimate the demand elasticities with respect to each of the determinants of public transport demand in the study area.

Chapter 8 conducts demand forecasting using the public transport demand model from Chapter 7. The demand model is validated and then used to forecast

demand for the SGMA. The forecast demand is compared to the actual public transport demand in 2009 and 2010 to evaluate the forecasting power of the dynamic pseudo panel data model.

Chapter 9 summarises the research findings and research contributions to the literature and policy implications, followed by a discussion on the limitations of this study and directions for future research before concluding.

CHAPTER 2 LITERATURE REVIEW

The literature review of this study comprises five sections in this chapter. Section 2.1 reviews the previous public transport demand elasticity studies, and highlights the importance of the temporal effect of travel demand by discussing the differences between short-run and long-run demand elasticities. Section 2.2 reviews the previous work on identifying the association between travel behaviour and land use characteristics. Section 2.3 introduces the background of the pseudo panel data approach and its previous applications in transport research. Section 2.4 and 2.5 discuss the issues of estimation techniques and model assumption testing which have not been fully discussed in previous pseudo panel data studies. The literature review identifies the research gaps to be addressed in this study which are summarised in Section 2.5.

2.1 Public transport demand elasticity

There has been an extensive body of demand elasticity studies in the field of public transport research. Some excellent review papers and meta-analysis studies have been reported to summarise the research outcomes and identify the research contribution (Goodwin, 1992, Oum et al., 1992, Nijkamp and Pepping, 1998, Kremers et al., 2002, Balcombe et al., 2004, Paulley et al., 2006, Holmgren, 2007, Hensher, 2008). In general, demand elasticity studies show a great variation in results due to six characteristics: (1) type of data used (aggregate or disaggregate), (2) time frame analysed (month, quarter, or year), (3) model structure (static or dynamic), (4) econometric technique used (pooled OLS or Fixed Effect estimator), (5) specification of the dependent variable (travel demand or mode choice), and (6) demand specification (Graham et al., 2009). Table 2.1 summarises previous studies on public transport demand elasticities in recent decades. This table shows the diversity of research scope, approaches, and results. The comparison of determinants, the distinction between short-run and long-run demand elasticities, and the evaluation of methodology is reviewed in the following sub-sections.

Table 2.1 Summary of Selected Public Transport Demand Elasticity Studies

Author	Mode	Area	Data Source	Methodology	Explanatory Variables	Price Elasticity	
						Short-Run	Long-Run
Voith (1991)	Rail	US	Transport Authority (aggregate)	Dynamic econometric model	Fare; Alternative costs; Peak/off-peak; Speed; Vehicle-km	-0.62	-1.59
Hensher (1998)	Rail Bus Car	Australia	SP and RP survey (disaggregate)	Discrete choice model	Fare; Travel time	-0.22 -0.36 -0.2	n/a
Hensher and King (1998)	Car Bus	Australia	SP and RP survey (disaggregate)	Discrete choice model	Fare; Alternative costs; Travel time	note ¹	n/a
Dargay and Hanly (2002)	Bus	UK	Department of Transport (aggregate)	Dynamic econometric model	Fare; Alternative costs; Income; Vehicle-km; Population; Urban Density	-0.33	-0.62
Douglas et al. (2003)	Bus Rail Ferry	Australia	SP survey; Second best survey (disaggregate)	Scenario model	Fare; Travel time; Service interval	-0.36 -0.38 -0.56	n/a
Bresson et al. (2003)	Public transport	France	Bus Operators National Statistics (aggregate)	Dynamic econometric model	Fare; Income; Vehicle-km	-0.32	-0.61
	Bus	UK	National Statistics (aggregate)	Dynamic econometric model	Fare; Income; Vehicle-km	-0.51	-0.69
García-Ferrer et al. (2006)	Metro Bus	Spain	Dept. of Transport (aggregate)	Dynamic econometric model	Fare; Vehicle-km	-1.03 -1.07	n/a

Author	Mode	Area	Data Source	Methodology	Explanatory Variables	Price Elasticity	
						Short-Run	Long-Run
Graham et al. (2009)	Metro	Worldwide	Railway and Transport Strategy Centre (aggregate)	Dynamic econometric model	Fare; Income; Vehicle-km	-0.05	-0.33
Wang (2009)	Bus Rail	New Zealand	NZ Transport Agency (aggregate)	Dynamic econometric model	Fare; Fuel Price; Car ownership	-0.24 -0.83	-0.4 -1.31
Dargay et al. (2010)	Car Rail Coach Air	UK	British National Travel Survey (aggregate)	Dynamic econometric model	Fare; Alternative costs; Income; Population; Car ownership; Travel time; Age; Gender; Employment; Household characteristics	-0.30 -0.30 -0.20 -0.10	-1.00 -1.00 -0.80 -0.30
Souche (2010)	Car Public transport	Worldwide	IUTP (aggregate)	Dynamic econometric model	Fare; Alternative cost; Income; Land use density	-0.74 -0.22	n/a

¹Cross elasticities are examined with various ticket types in two scenarios.

2.1.1 Determinants of demand elasticity

Public transport demand is determined by various factors. The components of the explanatory variables selected for a demand model have a significant impact on the estimated demand elasticities. In regard to the choice of variables, previous researchers have highlighted some important factors to be considered in a demand model. Balcombe et al. (2004) had a full discussion on the relationship between public transport demand and these factors. This report explicitly analysed the effects of fare, quality of service, competing modes, income, car ownership, and land use factors on travel demand. This study concluded that fare, quality of service, and car ownership are the most significant factors for public transport demand. A meta-analysis reported by Holmgren (2007) suggested that an ideal demand model should include fare, car ownership, fuel price, quality of service, and income.

The literature reviewed in Table 2.1 demonstrates that existing studies show a great diversity of explanatory variables in the public transport demand models. In general, the composition of independent variables varies with the purposes of study, research methodology, and data availability. Table 2.2 lists all the variables examined in the previous studies reviewed, classified into four groups: (1) travel costs; (2) quality of public transport service; (3) socio-economic factors; and (4) land use factors, with the number of studies that include this variable in their demand models in brackets after each variable.

Table 2.2 Categories of Explanatory Variables in Demand Models (Number of studies using these variables)

Category	Explanatory Variables
Travel Costs	Fare (11); Alternative costs (5); Travel time (4); Fuel price (1)
Quality of Service	Vehicle-km (5); Peak/off-peak vehicles (1); Speed (1); Service Interval (1)
Socio-economic Factors	Income (6); Car ownership (2); Age (1); Gender (1); Employment status (1); Household characteristics (1)
Land Use Factors	Population (3); Land use density (2)

Source: summarised from Table 2.1

Table 2.2 shows that fare, alternative costs, vehicle-km, and income are the most commonly examined variables, with fuel price and travel time as variables reflecting related and alternative costs to public transport fares. On the other hand, land use factors, where used, appear to have less diversity of measures. Balcombe (2004) has documented that the land use could influence public transport demand through dispersion of activities, shape of urban area, density, clustering of trip ends, and settlement size. Cervero and Kockelman (1997) also suggested that the density, diversity, and design (3Ds) of land use would have impacts on the non-auto trips. However, only land use density has been used as a measure to represent land use characteristics in the reviewed studies, partly because of data availability and partly because of the complexity of land use characteristics. Therefore, although most studies show a common pattern of variable composition, some factors such as land use variables deserve more attention.

2.1.2 Short-run and long-run elasticity

Whilst modelling the travel demand for public transport systems, many researchers have pointed out the importance of distinguishing between short-run and long-run demand elasticities. Voith (1991) suggested that the ridership of public transport systems might not change immediately in response to system changes, so the public transport operators need to be aware of potential long-term effects when proposing their financial plans. Oum et al. (1992) indicated that long-run demand is likely to be more elastic than short-run demand because travellers have more options to change their travel behaviour in the long run as compared to the short run. Goodwin (1992) suggested that an individual's adjustment of travel behaviour, as affected by travel costs, is not immediate and because this reaction may take a longer time period, so this temporal effect should be taken into account in a travel demand model. In transport planning, the long-run demand is particularly of interest for strategic planning and, in this context, Litman (2004) pointed out that conventional travel demand models based on short-run elasticity may underestimate the long-term impacts of service changes on public transport ridership. Collectively these studies show the

importance of identifying long-run demand elasticities when studying public transport demand.

Despite the importance of distinguishing between short-run and long-run public transport demand, the existing literature does not identify uniform timeframes which distinguish between short-run and long-run travel demand. Studies have varied in their choice of absolute time period selected and this appears to depend on the context of the study. The long-run demand is the demand derived after individuals have fully adjusted their travel behaviour and reached the long-run equilibrium as a result of system changes (Oum et al., 1992), and the time period for individuals to fully adjust their travel behaviour in response to the changes in exogenous variables may take years (Batley et al., 2011). This time period tends to be affected by individuals' lagged adjustment of residential and work location choice, habits of travel, travel costs, imperfect information, and uncertainty (Dargay and Hanly, 2002). The lagged behaviour change has been identified from demand modelling through a distinction between short-run and long-run demand elasticities, with results showing long-run demand elasticities being greater than short-run demand elasticities if the lagged adjustment exists. As shown in Table 2.1, the estimated long-run demand elasticities are generally two times to three times greater than their associated short-run elasticities.

Previous studies that have attempted to identify long-run demand elasticities have mostly applied a dynamic Partial Adjustment Model (PAM) which allows a lagged dependent variable to take the time effects into consideration. The lagged dependent variable is used to control for the fact that the current demand is affected by demand in previous time periods. The coefficient of the lagged dependent variable in a demand model therefore represents the speed of adjustment between short-run demand and long-run demand. In a dynamic model, the time periods for short-run demand and long-run demand are thus not fixed but vary with the speed of behaviour adjustment. By definition, the short-run time period refers to the shortest time period that can be measured using the data of the study (Batley et al., 2011). For example, a model that uses yearly data defines the short-run period as one year. From the estimated model, the short-

run elasticities are directly derived from the coefficients of the explanatory variables, whereas the long-run elasticities are adjusted by the coefficient of the lagged dependent variable with the “long-run” being defined as the time period for 95 percent of the demand response to be adjusted (Jevons et al., 2005). A greater coefficient of the lagged dependent variable means that the demand is more influenced by previous demand which will result in greater long-run elasticities. Therefore, if individuals’ behaviour is highly affected by their previous decision, they will need to take a longer time to adjust their behaviour.

However, although well-recognised in the literature, the distinction between short-run and long-run demand has mostly studied at an aggregate level by comparing multiple systems across different regions (Bresson et al., 2003; Graham et al. 2009, Souche, 2010). There is no much work focusing on a specific study area using the individual level of data. As a result, the local transport planners lack the information about individuals’ long-run demand changes and thus impede the long-term transport planning.

2.2 The relationship between land use and travel behaviour

2.2.1 Evidence from previous studies

There has been extensive work on identifying the interactions between travel behaviour and land use characteristics outside studies concerned with explaining public transport demand. Travellers’ choice of trip mode can be influenced by land use characteristics such as land use density, diversity, design, and accessibility (Cervero and Kockelman, 1997). The measurement of these land use factors in previous studies is shown in Table 2.3. In principle, land use density generally refers to housing, population, or employment density. Diversity is used to illustrate the mixed nature of land use and is normally measured as the entropy of the land mix. Land use design refers to the connectivity or walkability of neighbourhood environment, such as the number of intersections, whereas accessibility usually refers to the access to the local public transport station or trip destinations.

Table 2.3 The Measurement of Land Use Factors from Selected Previous Studies

Author	Dependent Variables	Measurement of land use factors			
		Density	Diversity	Design	Accessibility
Cervero and Kockelman (1997)	Vehicle mile; Probability of non-car travel	Intensity factors;	Land use mix; Vertical mixing; Population within ¼ mile of a store	Four-way intersections; Quadrilaterals; Sidewalk width; Parking	Accessibility index
Kitamura et al. (1997)	No. of vehicles; Trip distance; Mode share	-	-	Sidewalk width; Proportion of parking, four-way intersections, and quadrilateral blocks	Distance to rail stations; Distance to the nearest park
Cervero (2002)	Mode Choice	Population and employment density	Employment and population relative to county ratio	Ratio of sidewalk miles to road miles	Proportion of households within 0.5 mile of metro stations
Rajamani (2003)	Mode Choice	Population density	Land use mix	Park area per housing unit; Cul-de-sacs	Distance to bus stops
Rodriguez and Joo (2004)	Mode Choice	Population density	-	Percentage of shortest route to closest bus stop with sidewalk	-
Zhang (2004)-Boston	Mode Choice	-	Land use mix; retail floor area ratio	Intersection density	-
Zhang (2004)-Hong Kong	Mode Choice	Population density; Job density	Land use mix	-	Distance to the nearest public transport station
Bentol et al. (2005)	Mode Choice	Population density	Job-housing balance	Population Centrality	-
Cervero (2006)	Station boarding	Housing density	Land use mix	Parking supply	-
Pinjari et al. (2007)	Mode choice	Household density; Employment density	L and use mix	Street block density; Bicycle facility density	Access time to bus stop
Frank et al. (2008)	Mode choice	Population density; Job density	Land use mix measures	Intersection density	-
Estupiñán and Rodríguez (2008)	Station boarding	Population density	Land use mix	Bike path; Sidewalk design; No. of intersections	-
Sohn and Shim (2010)	Station boarding	Population density; Employment density	Land use mix	Road length; No. of dead ends; No. of intersections	Accessibility indices
Buehler (2011)	Mode choice	Population density	Land use mix	-	Distance to public transport
Sung and Oh (2011)	Station boarding	Residential density; Commercial density;	Land use mix; Commercial/business mix;	Road length; Road width; No. of intersections; Dead end roads	Subway accessibility

From the studies reviewed in Table 2.3, the land use factors show different levels of influence on travel behaviour which are reflected in their elasticity. Ewing and Cervero (2010) estimated the weighted average elasticities of public transport to these land use factors from existing studies as shown in Table 2.4. The results indicate that public transport demand appears to be inelastic to the individual land use factors, although the relationship is significant. However, Ewing and Cervero (2010) suggested that even if individually the effect is small, the combined contribution of all these land use factors could be significant.

Table 2.4 Weighted Average Elasticity of Public Transport Demand with Land Use Factors

Criterion	Measures	Elasticity
Density	Household/ Population density	0.07
	Job density	0.01
Diversity	Land use mix (entropy)	0.12
Design	Intersection/street density	0.23
	Percentage of four-way intersections	0.29
Accessibility	Distance to nearest public transport stop	-0.29

Source: Ewing and Cervero (2010)

Table 2.5 summarises the research methodology and results from these studies reported in literature. In terms of methodology, the choice modelling approach is the most widely adopted method because the choice model is capable of examining the probability of an individual's choice of travel modes. The estimated elasticities with respect to land use factors show a variety of results. In general, most elasticities estimated in previous studies show a positive sign indicating that public transport use increases with higher density, diversity, accessibility, and more walking-supportive urban design. However, some inverse signs and insignificant estimates are shown implying that the interaction between travel behaviour and land use factors might be uncertain in some cases. This uncertainty is possibly from the limitations of the modelling approach or from data availability, with results varying by locations.

Table 2.5 Selected Studies on Public Transport Demand Elasticities with respect to Land Use Measures

Author	Area	Methodology	Data	Elasticities of public transport trips to land use factors			
				Density	Diversity	Design	Accessibility
Cervero and Kockelman (1997)	San Francisco	Multiple regression model; Logit choice model	Travel Diary; Census	0.084~0.113	0.365	0.087~0.183	-
Kitamura et al. (1997)	San Francisco	Linear regression model	Field Survey	-	-	-	-
Cervero (2002)	North America	Logit choice model	Household Travel Survey	0.268~0.511	0.452~0.615	0.327	0.195
Rajamani (2003)	North America	Logit choice model	Activity survey	0.0775	-0.037	-0.012~0.0004	0.418
Rodriguez and Joo (2004)	North America	Multinomial choice model	Field survey; Census	-0.204~ -0.537	-	0.251~2.762	-
Zhang (2004)	Boston	Discrete choice model	Household Travel Survey	0.004~0.118	0.121	-	0.044~0.083
	Hong Kong			0.005~0.014	-	-	-
Bento et al. (2005)	North America	Discrete choice model	Household Travel Survey	-2.70~ -2.41	insignificant	-5.35~4.26	-
Cervero (2006)	North America	Post-processing and direct models	National Database	0.145	0.043	0.045	
Frank et al. (2008)	Seattle	Choice model; Tour based trips	HTS; GIS database		0.01~0.34	0.14~0.26	-
Sung and Oh (2011)	Seoul	Multiple regression models	Smart card; GIS tools	0.106~0.176	0.116~0.223	-0.233~0.333	0.145

2.2.2 Land use factors in public transport demand modelling

Although the interaction between travel behaviour and land use characteristics has been demonstrated in the literature as discussed above, this has not been commonly integrated into the reported literature of public transport demand studies as reviewed in Section 2.1. The literature shows a distinct separation of two areas with one modelling public transport demand with respect to its determinants; and the other pursuing the relationship between travel mode choices and land use characteristics. These two areas of studies are clearly interrelated but it seems there has been little effort to integrate these two fields of knowledge.

As discussed in Section 2.1, a conventional public transport demand model is used for demand forecasting and elasticities investigation. The elasticities derived from demand models are clearly affected by the set of explanatory variables examined. However, the conventional public transport demand studies have not yet comprehensively incorporated land use variables. Although land use characteristics can be evaluated from various measures as discussed above, these measures have only entered the demand model studies in terms of population density. Other factors including land use diversity, urban design, and accessibility, which have demonstrated their importance in travel behaviour, have not been fully considered in demand modelling studies. Excluding factors from demand modelling that have a significant impact on travel demand potentially causes biased estimates. Thus, the integration between the two fields of knowledge is needed to strengthen the linkage between public transport demand and land use characteristics.

2.2.3 Self-selection of location choice

The association between travel behaviour and land use has been evident, but the causality that generates this association is not fully understood. A possible causality that links travel behaviour and land use characteristics is an individual's location choice, because the choice of residential location can be influenced by people's preference of mode choice. For example, people who do not have a car are more likely to choose a residential location close to public transport stations and thus walk and use public transport more. These attitudinal attributes of mode choice preferences involved in the decision of location choice is called 'self-selection' in the literature.

When investigating the relationship between travel behaviour and land use characteristics, it is important to understand the extent to which an individual's mode choices are attributed to the land use characteristics and self-selection. If self-selection exists but ignored, the impact of land use on travel behaviour will be overestimated. The self-selection problem has received significant attention and there is a rich body of literature which aims to identify the causality of

people's travel behaviour and location choice (Golob, 2003, Simma and Axhausen, 2004, Rivera and Tiglao, 2005, Bhat and Guo, 2007, Cao et al., 2007, Cervero, 2007, Pinjari et al., 2007). A comprehensive review paper (Mokhtarian and Cao, (2008) summarises seven approaches to investigate the self-selection problem: direct questioning using questionnaires to ask respondents' attitude toward their choices of household locations, statistical control, instrumental variables models, sample selection models, joint discrete choice models, structural equations models, and longitudinal designs. They concluded that a longitudinal structural equations modelling using genuine panel data or designed experiment is the most ideal approach to identify the causality of travel behaviour and land use characteristics. However, if attitudinal data are not available and the self-selection is not the main focus of the demand study, it can be controlled by controlled by incorporating the socio-economic factors of travellers in the demand model. This approach has been suggested by Zhang (2011) as a way of mitigating the impact of self-selection on travel demand.

2.3 Pseudo panel data

2.3.1 Background of pseudo panel data

As shown in Section 2.1.2, differences between short-run and long-run elasticities are to be expected. Given the importance of time effects on people's decisions on travel, the data used for estimating travel demand models need to identify the individuals' dynamics of travel behaviour. Dargay and Vythoulkas (1999) suggested two types of data that can be used for dynamic modelling. One is genuine panel data collected by conducting surveys with the same participants over a significantly long time period at a disaggregate level. The other is aggregate time-series data collected from observations on a larger group of people, aggregated at some spatial – usually regional - level.

Genuine panel data at a disaggregate level are ideal to investigate a traveller's long-term travel behaviour, but in practice this sort of data is rarely available. In comparison, aggregate time-series data are easier to obtain but they lack information at the individual level. Dargay and Vythoulkas (1999) suggested that using pseudo panel data can reduce the constraints of existing data availability.

Pseudo panel data were first introduced by Deaton (1985) in consumer economics. In the field of transport planning, it has been adopted by studies on forecasting car ownership and car use (Dargay and Vythoulkas, 1999, Dargay, 2001, Dargay, 2002, Dargay, 2007, Huang, 2007, Weis and Axhausen, 2009). The concept of a pseudo panel dataset is to use existing repeated cross-sectional data and to group individuals or households into cohorts by time-invariant variables such as birth year and household characteristics. This allows the identification of the patterns of travel behaviour in each defined cohort to be examined. In turn this approach can better explain individuals' travel behaviour since it has a micro-economic level basis for the examination of behaviour over time and thus provides the potential to adapt aggregate modelling approaches to the disaggregate context.

A comparison of the benefits and disadvantages of genuine panels, repeated cross-sectional data, and pseudo panels are summarised in Table 2.6. Genuine panels are the most ideal data sources to capture individuals' behavioural changes over time and to identify a causality effect. However, genuine panels also possess some deficiencies from sampling and survey process that may limit data availability. Compared to genuine panels, repeated cross-sectional data offer greater sample observations, because representative samples are drawn for each wave of the survey independently. The main shortcoming of repeated cross-sectional data is that the respondents are not traced over time, so the travel behaviour change over time cannot be captured from different individuals. A pseudo panel, grounded in sound theory, is an alternative approach to conduct a longitudinal study on behavioural changes. Although the cohort level of aggregation in a pseudo panel dataset loses some individual information, it has been suggested that this loss can be minimised if the variation within cohorts can be controlled to be much smaller than the variation between cohorts (Verbeek and Nijman, 1992).

Table 2.6 A Comparison of Genuine Panels, Repeated Cross-sectional Data, and Pseudo Panels

	Pros	Cons
Genuine Panels	<ul style="list-style-type: none"> • Capture behavioural change over time • Identify causality effect • Ideal for forecasting purpose 	<ul style="list-style-type: none"> • Coverage limitation • Sample attrition • Panel conditioning and fatigue • Data are rarely available
Repeated Cross-Sectional Data	<ul style="list-style-type: none"> • Sufficient data sources • Steady sampling condition 	<ul style="list-style-type: none"> • Larger sampling errors • Not tracking the same individuals
Pseudo Panels	<ul style="list-style-type: none"> • Identify the dynamics of travel behaviour • Appropriate if the samples within cohorts are grouped homogenously • Tracking the same observations based on a cohort level 	<ul style="list-style-type: none"> • Loss of individual information after cohort aggregation • Loss samples after matching cohorts

Summarised from Verbeek (1992), Yee and Niemeier (1996), and Raimond and Hensher (1997)

2.3.2 Principles of pseudo panel data construction

A pseudo panel dataset is created by using existing repeated cross-sectional data to create cohorts by forming individuals or households into groups by time-invariant variables such as birth year. Each created group is constituted of cohorts with the same grouping criteria identified over the observed time period. After forming the pseudo panel dataset, a cohort is treated as a single observation in the dataset and the mean values of the variables are computed to represent the observations. Provided that the pseudo panel dataset can be created in a way to generate sufficient inter-group heterogeneity, the pseudo-panel approach can allow the identification of the patterns of travel behaviour from the defined groups.

The aim of cohort creation is to reduce the variation within the cohorts and increase the variation between cohorts to ensure each cohort can be treated as an independent individual and be traced over time (Verbeek, 1992). From Table 2.7 which summarises previous pseudo panel data studies, birth year is the most commonly used variable to create groups. It is also clear that other variables can be adopted to complement birth year if further subdivision of cohorts becomes necessary. The selection of complementary criteria depends on the context of study. For example, Bernard et al. (2010) chose household location and house size because people living in the same region and the same size of house share common features of electricity consumption. Therefore, although birth year is the

most obvious variable to create cohorts and its representativeness has been demonstrated from previous studies, other variables complementary to birth year can be used to make the observations within cohorts more homogenous or create additional cohorts.

Table 2.7 A Comparison of Selected Previous Pseudo Panel Data Studies

Author	Context of Study	Study Area	Grouping Criteria
Gassner (1998)	Telephone access	UK	Birth year
Dargay and Vythoulikas (1999)	Car ownership	UK	Birth year
Dargay (2002)	Car ownership	UK	Birth year
Gardes et al. (2005)	Food consumption	US	Birth year; Education
Dargay (2007)	Car travel demand	UK	Birth year
Huang (2007)	Car ownership	UK	Birth year
Weis and Axhausen (2009)	Travel demand	Switzerland	Birth year; Gender; Region
Warunsiri and McNown (2010)	Return to education	Thailand	Birth year
Bernard et al. (2011)	Electricity	Canada	Region; House size

2.3.3 Pseudo panel data in transport literature

Although the pseudo panel data approach has been developing since 1985, there is not much empirical work applied in transport field, and previous pseudo panel data studies in transport have focused on car travel studies rather than public transport.

The first application of the pseudo panel approach in transport was conducted by Dargay and Vythoulikas (1999). They constructed a pseudo panel dataset from the

UK Family Expenditure Surveys from 1970 to 1994, and studied the relationship between car ownership and its determinants including income, car costs, public transport fare, and the socio-demographic factors of the households. The short-run and long-run elasticities were investigated by employing a dynamic Partial Adjustment Model. The results show the long-run elasticities to be generally three times greater than short-run elasticities. This was the first attempt applying a pseudo panel dataset to a transport study, and empirically it demonstrated that the pseudo panel data can be estimated as conventional panel data if the cohorts are created with sufficient inter-group variation.

Based on the UK data, Dargay (2002) extended this car ownership study to investigate the difference in car ownership between rural and urban areas. The results indicated that households in urban and rural areas do have different sensitivity to motoring costs for their household car ownership, and long-run elasticities are around 40 percent greater than short-run elasticities. This study showed the ability of a pseudo panel data approach to distinguish the difference in travel behaviour across cohorts possessing different characteristics.

Dargay (2007) further investigated the asymmetry of the relationship between car travel, car ownership, and household income. This study found that the relationship between car ownership and the changes in household income is not symmetric, because rising income encourages households to purchase more cars, whereas the falling income does not make households abandon their purchased cars. This finding can be regarded as showing the benefit gained from a pseudo panel approach over an aggregate approach because it provides a certain level of micro-economic underpinning travel behaviour.

Huang (2007) also created a pseudo panel dataset from the UK Family Expenditure Surveys. This study employed a linear dynamic econometric model as well as a non-linear discrete choice model to study the household car ownership in the UK. He suggested that although the cohort dataset is an intermediate level between disaggregate and aggregate level, it does not violate

the random utility theory of disaggregate choice modelling based on an individual level.

Weis and Axhausen (2009) studied induced travel based on a pseudo panel dataset constructed from the Swiss National Travel Survey. A Structural Equations Model (SEM) was applied to take account of the interactions among endogenous variables such as share of mobiles, number of trips, and travel distance. Their findings confirm that generalised cost has a significant impact on travellers' mobility. The results show that pseudo panel data can be empirically estimated in SEM as if they were conventional panel data.

The empirical evidence shows that the pseudo panel data can be treated as genuine panel data at a cohort level in the estimation process. Despite the loss of the real individuals' information, pseudo panel data still provides a deeper insight into the variation of respondents' socio-economic characteristics over time as compared to aggregate data. In terms of modelling flexibility, existing studies in both transport and other contexts have also demonstrated that the pseudo panel data can be applied in a wide range of models, including dynamic models, random utility models, and SEM.

However, the data requirement for pseudo panel data is still demanding. First, the cohort size has to be sufficiently large, suggested as more than one hundred individuals within each cohort (Verbeek and Nijman, 1992). Second, the number of observations also needs to be large enough to have valid statistical efficiency, thus a trade-off between cohort size and number of observation occurs. Third, the pseudo panel dataset must allow sufficient group-specific variation and thus choosing variables to be examined becomes a difficult task. This is because after averaging the cohort variables, the variation of variables across observations is reduced, and the explanatory power of the chosen variables becomes lower. These issues are obstacles for empirical applications if the repeated cross-sectional data do not provide sufficient and quality samples.

2.3.4 Issues of application to public transport

The issues of a pseudo panel data approach mentioned above appear to be more severe for public transport demand studies, and thus explain the lack of pseudo panel studies focusing on public transport in the literature. In practice, the challenges to be overcome in the context of public transport include:

- (1) The limited number of public transport trips as a result of low public transport usage.
- (2) The interpretation of public transport demand at the cohort level needs to be defined, whether demand is the number of trips or mode share.
- (3) The difficulty in generating a group-specific price variable for public transport, given the complexity of fare types and ticket types.

Although repeated cross-sectional travel surveys offer a consistent and large scale data, the number of observations will be considerably reduced after forming the cohorts. Besides, the cohort size also needs to reach a certain level of threshold to reduce the measurement errors. If the research context is car travel or car ownership, this issue is less problematic since the number of car users from travel surveys is mostly sufficient for statistical analysis. In contrast, where the public transport usage is only a minor proportion of total trips, the low number of public transport observations will limit the flexibility of the cohort construction and lead to the loss of the statistical power as a result of small cohort size or small number of observations.

The second issue in public transport studies is how the public transport demand should be interpreted. Compared to car travel studies, in which the car travel distance or the number of household car ownership are used to explain the car travel demand, the interpretation of public transport demand at a cohort level appears to be more complicated. The travel distance used in car travel studies should not be applied to public transport trips, because the travel distance of public transport trips is expected to be highly correlated with the individual location information which determines the distance between travellers' origin and destination, thus confounding the relationship between the level of public transport use and its explanatory factors. Apart from pseudo panel studies,

conventional public transport demand research tends to use passenger-km or ridership for an aggregate study, and discrete mode choice for disaggregate study as shown in Table 2.1. Both approaches can not be directly applied to a cohort analysis because of the different aggregation level. Besides, the variable used in a cohort analysis is expected to have some group-specific characteristics, which means that respondents assigned to different groups are expected to carry sufficient variation. Therefore, careful attention must be given to the choice of a reasonable dependent variable used to explain the public transport demand at a cohort level.

In the context of public transport demand elasticity, the travel cost as the price variable is certainly one of the main research interests. At a cohort level, the price variable ought to possess some group-specific variation. This issue has been identified in Dargay and Vythoulkas (1999) where a weighted train and bus fare was used which was not group-specific in this study. They pointed out that more transport related questions could be analysed if more variation in the price variable can be established.

Weis and Axhausen (1999) used a travel price index to represent a measure of the travel price relative to the general consumer price. This measure had variation over time but was constant across the defined groups. In public transport demand studies, it is important to incorporate the variation of public transport fares, as applied to different groups of users, such as concession price for students and pensioners as price variation is expected to have an impact on the public transport demand. This will be especially the case for pseudo panel studies that use birth year to create groups and can identify the travel behaviour across different generations.

2.4 Panel data model estimation

Pseudo panel data models have been empirically estimated in previous applied research, and there have been some technical reports examining the performance of various estimators on pseudo panel data. This section reviews the conventional estimation techniques developed in the literature of panel data analysis, and

discusses the estimation techniques applied in previous pseudo panel data studies.

2.4.1 Static genuine panel data model and estimation

The theoretical background of the estimation techniques and model assumptions of genuine panel data are the fundamental knowledge of pseudo panel data models. A simple panel data model can be described as Equation (2.1):

$$y_{it} = \beta_0 + \beta_1 x_{it} + u_{it}, \quad i=1, \dots, N; t=1, \dots, T \quad \text{Equation (2.1)}$$

where i denotes the panel units (eg. firm, country, or household), t denotes the time period, and u_{it} represents the error term.

The main concern of genuine panel data estimation is that the panel units are likely to possess unobserved heterogeneity that are correlated with u_{it} , and this leads to biased and inconsistent estimates using the pooled Ordinary Least Squares estimator (OLS) through the violation of the error term independency. To control for unobserved heterogeneity, various estimators have been developed for genuine panel data models such as the Fixed Effect (FE), Random Effect estimator (RE), and Instrumental Variable estimator (IV). These estimators are based on different model assumptions and are used to accommodate different panel data properties and model forms. A comprehensive introduction to these developed estimators is summarised in Hsiao (1986).

Given that unobserved heterogeneity is hidden in the error term as a part of u_{it} in Equation (2.1), the FE model modifies Equation (2.1) by adding a time-invariant unobserved factor α_i assumed to be correlated to x_{it} and replacing u_{it} by ε_{it} which has a mean value of zero and is independent of x_{it} as described in Equation (2.2). In the FE estimation, α_i is eliminated through a demeaned transformation to ensure the error term is not correlated with the regressors.

$$y_{it} = \beta_0 + \beta_1 x_{it} + \alpha_i + \varepsilon_{it}, \quad E(\varepsilon_{it}) = 0 \quad \text{Equation (2.2)}$$

To eliminate α_i , the FE estimator averages Equation (2.2) over time period t :

$$\bar{y}_i = \beta_0 + \beta_1 \bar{x}_i + \alpha_i + \bar{\varepsilon}_i \quad \text{Equation (2.3)}$$

Subtracting Equation (2.3) from Equation (2.2) :

$$\begin{aligned} (y_{it} - \bar{y}_i) &= (\beta_0 - \beta_0) + \beta_1(x_{it} - \bar{x}_i) + (\alpha_i - \alpha_i) + (\varepsilon_{it} - \bar{\varepsilon}_i) \\ &= \beta_1(x_{it} - \bar{x}_i) + (\varepsilon_{it} - \bar{\varepsilon}_i) \end{aligned} \quad \text{Equation (2.4)}$$

In Equation (2.4), the unobserved individual effect α_i is eliminated, so the OLS estimator will be unbiased after transformation. This OLS estimator, based on the demeaned deviation, is called the FE estimator or within estimator which only take accounts of the within variation in the variables by treating α_i as fixed parameters.

Another way to estimate a FE model is called Least Squares Dummy Variable (LSDV) estimator. This is undertaken by including a dummy variable to each of $N - 1$ panel units as an explanatory variable. This approach can be written as Equation (2.5). In Equation (2.5), each dummy variable d in unit i is used to incorporate the individual effect. The estimation results from the FE estimator and the LSDV estimator are identical in terms of slope coefficients and standard errors, but LSDV takes account of the dummy variables in the estimation process and thus loses degrees of freedom.

$$y_{it} = \alpha_1^* d_{1i} + \alpha_2^* d_{2i} + \dots + \alpha_N^* d_{Ni} + \beta_1 x_{it} + \varepsilon_{it} \quad \text{Equation (2.5)}$$

The FE estimator is used to accommodate the correlation between unobserved heterogeneity α_i and explanatory variables x_{it} . However, if no correlation exists, then there is no efficiency gained from using a FE model. Instead, a Random Effect (RE) estimator is preferred over the pooled OLS estimator which is still biased because of the unobserved heterogeneity in α_i leads to serial correlation of error terms. This can be seen in Equation (2.6) where the composite error term u_{it} is serial correlated due to the presence of α_i .

$$y_{it} = \beta_0 + \beta_1 x_{it} + u_{it}, \quad u_{it} = \alpha_i + \varepsilon_{it} \quad \text{Equation (2.6)}$$

The RE estimator employs a Generalised Least Squares (GLS) transformation by introducing a scalar, θ , defined as Equation (2.7).

$$\theta = 1 - \sqrt{\frac{\sigma_\varepsilon^2}{\sigma_\varepsilon^2 + T\sigma_\alpha^2}} \quad \text{Equation (2.7)}$$

The variables in Equation (2.6) are next averaged and multiplied by θ , and subtracted from Equation (2.6) to give Equation (2.8) where the transformed error terms are serially uncorrelated. Given that θ is only constituted of two parameters σ_ε^2 and σ_α^2 , these can be estimated through a LSDV estimator and a OLS estimator respectively. This process yields the RE estimator using the GLS transformation based on the assumption that the unobserved heterogeneity is uncorrelated with explanatory variables.

$$(y_{it} - \theta \bar{y}_i) = (\beta_0 - \theta \beta_0) + \beta_1 (x_{it} - \theta \bar{x}_i) + (\varepsilon_{it} - \theta \bar{\varepsilon}_i) \quad \text{Equation (2.8)}$$

The choice between the FE and RE estimator depends on the nature of model forms and data properties. The key distinction is that the FE estimator allows correlation between unobserved heterogeneity and explanatory variables, whereas the RE estimator assumes that they are uncorrelated. If the assumption is that there is no correlation between unobserved heterogeneity and explanatory variables, then the RE estimator will provide more efficient estimation over the FE estimator. Conversely, if this assumption is violated, then the RE estimator is inconsistent so the FE estimator is preferred. This assumption can be examined by employing a Hausman's test in which the null hypothesis is that the correlation between unobserved heterogeneity and explanatory variables is zero.

Another characteristic of the FE estimator is that it only takes accounts of within-group variation and thus any time-invariant variables such as gender cannot be estimated through a FE estimator, whereas the RE estimator which is a weighted estimation of the between-group variation and within-group variation

is able to incorporate time-invariant variables. Plümper and Troeger (2007; 2011) have examined the performance of the FE estimator when estimating time-invariant or rarely changing variables in the model and found that the FE estimator will be inefficient under this circumstance. Therefore, the FE estimator may not be preferred if the model includes some variables that are of interest but rarely changing over time.

2.4.2 Dynamic panel data model and estimation

Genuine panel data models capture the dynamic economic behaviour through a dynamic panel data model. In transport, individuals' travel behaviour in response to transport system changes is suggested to be affected by their lagged adjustments of residential and work location choice, car ownership, habits of travel, travel costs, life cycle changes, imperfect information, and uncertainty (Voith, 1991; Goodwin, 1992; Oum et al. 1992, Dargay and Hanly, 2002, Batley et al., 2011). The dynamic economic behaviour in genuine panel data analysis can be modelled using a dynamic functional form. A general dynamic panel data model form can be specified as Equation (2.9).

$$y_{it} = \beta_0 + \sum_{m=1}^p \lambda_m y_{i,t-m} + \sum_{j=1}^n \sum_{m=0}^q \beta_{jq} x_{i,jt-m} + \varepsilon_{it} \quad i=1,\dots,N; t=t-q,\dots,T \quad \text{Equation (2.9)}$$

Equation (2.9) is known as an autoregressive distributed lag model where p denotes the number of lags of y_{it} , q denotes number of lags of x_{it} , and n is the number of exogenous regressors. ε_{it} represents the error term. The model can be simplified when $p = q = n = 1$ as Equation (2.10).

$$y_{it} = \beta_0 + \lambda_1 y_{it-1} + \beta_1 x_{it} + \beta_2 x_{it-1} + \varepsilon_{it} \quad \text{Equation (2.10)}$$

This simplified general dynamic model identifies that the current economic behaviour is expected to be affected by the exogenous variable x_{it} , and its lagged value at x_{it-1} as well as the lagged value of the dependent variable y_{it-1} . The estimated coefficients β_1 and β_2 are short-run multipliers which represent the

effect on y_{it} of a unit change in x_{it} and x_{it-1} , whereas λ_1 is used to derive the long-run multiplier $(\frac{\beta_1 + \beta_2}{1 - \lambda_1})$.

Various dynamic model forms have been developed including models incorporating the lagged adjustments higher than the first order and they have been empirically employed according to the nature of economic behaviour assumed. The Partial Adjustment Model (PAM) as specified in Equation (2.11) is commonly applied to take account of the effect of previous behaviour on current behaviour where the lagged dependent variable, y_{it-1} , is used to represent the economic behaviour in previous time period $t - 1$, and β_2 is assumed to be zero which assumes that the lagged value of X_{it} has no significant impact on Y_{it} .

$$y_{it} = \lambda y_{it-1} + \beta_1 x_{it} + u_{it}, \quad u_{it} = \alpha_i + \varepsilon_{it}, \quad i = 1, \dots, N, \quad t = 1, \dots, T \quad \text{Equation (2.11)}$$

The major problem in estimating the dynamic model (Equation (2.9)) is that without controlling for the individual effect α_i , y_{it-1} is correlated with the composite error term μ_{it} because α_i , which does not change over time, will influence y_{it-1} in the estimation process. Thus, using pooled OLS estimation will result in biased estimates of λ . In a static model, where $\lambda = 0$ in Equation (2.9), the correlation between α_i and μ_{it} can be controlled by using a FE estimator as discussed above. However, in the presence of y_{it-1} , pooled OLS estimation is biased upward and FE estimation is biased downwards (also known as Nickell bias, Nickell (1981)) as a result of the endogeneity between y_{it-1} and the error term.

The FE estimator transforms model (Equation (2.9)) to eliminate α_i and then gives Equation (2.12).

$$(y_{it} - \bar{y}_i) = \lambda(y_{it-1} - \bar{y}_{it-1}) + \beta_1(x_{it} - \bar{x}_i) + (\alpha_i - \alpha_i) + (\varepsilon_{it} - \bar{\varepsilon}_i) \quad \text{Equation (2.12)}$$

where

$$y_{it-1} - \bar{y}_{it-1} = y_{it-1} - \frac{1}{T-1}(y_{i1} + \dots + y_{it} + \dots + y_{iT-1})$$

$$\varepsilon_{it} - \bar{\varepsilon}_i = \varepsilon_{it} - \frac{1}{T-1}(\varepsilon_{i2} + \dots + \varepsilon_{it-1} + \dots + \varepsilon_{iT})$$

The issue in Equation (2.12) is that $(-\frac{y_{it}}{T-1})$ is correlated with ε_{it} , and $(-\frac{\varepsilon_{it-1}}{T-1})$ is correlated with y_{it-1} . Therefore, although α_i is eliminated through the transformation, the estimates are still biased in finite T because the transformed lagged dependent variable and the transformed error term are correlated.

The RE estimator is also biased in the dynamic model, because the GLS transformation does not eliminate α_i in the estimation process as shown in Equation (2.13). Therefore, the presence of α_i leads to biased estimates.

$$(y_{it} - \theta \bar{y}_i) = \lambda(y_{it-1} - \theta \bar{y}_{it-1}) + \beta_1(x_{it} - \theta \bar{x}_i) + (\alpha_i - \theta \bar{\alpha}_i) + (\varepsilon_{it} - \theta \bar{\varepsilon}_i) \quad \text{Equation (2.13)}$$

Anderson and Hsiao (1981) proposed an Instrumental Variable (IV) estimator to address this endogeneity problem. This method introduces an instrument, z_{it} , which is correlated with Δy_{it-1} but uncorrelated with $\Delta \varepsilon_{it}$, so the parameters are estimated in two stages. In the first stage Δy_{it-1} is regressed by z_{it} using an OLS estimator in Equation (2.14). In the second stage, Δy_{it} is regressed by the fitted values generated from the first stage using an OLS estimator in Equation (2.15). Thus, this IV estimator is also called the two-stage least square estimator (2SLS).

$$\Delta y_{it-1} = \delta z_{it} + \omega_{it} \quad \text{Equation (2.14)}$$

$$\Delta y_{it} = \lambda \hat{y}_{it-1} + \Delta \varepsilon_{it} = \lambda(\hat{\delta} z_{it}) + \Delta \varepsilon_{it} \quad \text{Equation (2.15)}$$

As it is important that the instrument z_{it} has to be correlated with Δy_{it-1} but uncorrelated with $\Delta \varepsilon_{it}$, Anderson and Hsiao (1981) proposed to use y_{it-2} as an instrument because it lies in the assumption, as seen in Equation (2.16).

$$\begin{aligned} \text{cov}(y_{it-2}, \Delta y_{it-1}) &= \text{cov}[y_{it-2}, (y_{it-1} - y_{it-2})] = E(y_{it-2}(y_{it-1} - y_{it-2})) \neq 0 \\ \text{cov}(y_{it-2}, \Delta \varepsilon_{it}) &= \text{cov}[y_{it-2}, (\varepsilon_{it} - \varepsilon_{it-1})] = E(y_{it-2}(\varepsilon_{it} - \varepsilon_{it-1})) = 0 \end{aligned} \quad \text{Equation (2.16)}$$

However, with regards to the IV estimator, Arellano and Bond (1991) suggested that the 2SLS estimator is inefficient because the first-differenced transformation is likely to produce serial correlation. Thus, they proposed to use the Generalised Method of Moments (GMM) estimator, as a form of IV estimation, to estimate the parameters more efficiently than the 2SLS estimator by imposing the moment conditions:

$$E(Z_i' \Delta u_i) = 0 \text{ for } i=1,2,\dots,N \quad \text{Equation (2.17)}$$

In general, if there is no serial correlation in the error terms, both IV and GMM estimators are consistent. If there is still serial correlation present after using instrument y_{it-2} , then further lags of dependent variables y_{it-3} or y_{it-4} should be employed. In the GMM framework the serial correlation can be tested and adjusted by using robust standard errors (Arellano and Bond, 1991).

The dynamic panel data differs from static panel data model by including a lagged dependent variable. Thus, on top of the static panel data model assumptions, a dynamic model produces another issue of concern- the correlation between lagged dependent variable and error terms which violates the assumption of strict exogeneity in a panel data model. Advanced IV estimators have been introduced to accommodate this problem, but the assumptions of the instruments also need to be tested. In choosing between the Anderson-Hsiao IV estimator and Arellano-Bond GMM estimator, Halaby (2004) identified that neither of the estimators has shown uniform superiority in every circumstance, and thus analysts should experiment both estimators in applied work.

Although the IV estimators have been acknowledged as an effective tool to deal with the endogeneity problem in dynamic panel data modelling, this method also suffers from some restrictions. Kiviet (1995) demonstrated that the IV estimation methods may lead to small sample bias and large standard errors which together

result in poor efficiency. Bruno (2005a) pointed out that the IV estimators are appropriate when the number of cross-section units (N) is large, but when N is small, IV estimators will lead to problematic estimates. On the other hand, although the FE estimator with standard OLS is biased in the dynamic model, in principle it generates smaller standard errors than IV estimators and thus makes the statistical inference more reliable (Beck and Katz, 2011). Therefore, if the bias in the standard OLS can be approximated, the corrected estimates with smaller standard errors from the standard OLS estimation will be favoured over the IV estimator.

Based on this concept, Kiviet (1995) developed a bias-corrected least squares dummy variables (LSDV) estimator that approximates and removes the bias from the standard LSDV estimator for dynamic panel data in the following way. Consider the PAM dynamic model as Equation (2.11) with observations collected over time and across panel units give Equation (2.18). Kiviet's work was extended and simplified as a more general form in Bun and Kiviet (2003) as follows:

$$y = W\delta + (I_N \otimes I_T)\alpha + \varepsilon \quad \text{Equation (2.18)}$$

where $\delta = (\lambda, \beta_1)$; y and $W = (y - 1, X)$ are $(NT \times 1)$ and $(NT \times k)$ matrices of stacked observations; $I_N \otimes I_T$ is the matrix of individual dummies where I_T refers to the $(T \times 1)$ vector of all unity elements. ε is $(NT \times 1)$ vector of error terms.

Using OLS yields the LSDV estimator as Equation (2.19).

$$\hat{\delta}_{LSDV} = (W'AW)^{-1}W'Ay \text{ where } A = I_N \otimes (I_T - \frac{1}{T}I_T I_T')$$
 Equation (2.19)

The bias of the LSDV estimator when N is infinite has been examined by Nickell (1981) and is of order $O(T^{-1})$. Bun and Kiviet (2003) suggested that in small samples there is additional bias given by Equation (2.20).

$$E(\hat{\delta}_{LSDV} - \delta) = O(T^{-1}) + O(N^{-1}T^{-1}) + O(N^{-1}T^{-2}) + O(N^{-2}T^{-2}) \quad \text{Equation (2.20)}$$

The total bias can be approximated by applying a Monte Carlo simulation. Kiviet (1995) concluded that the standard LSDV estimator generates smaller standard errors than the IV estimators for small samples suggesting the LSDV estimator is potentially more efficient at reducing the bias in dynamic modelling. The Monte Carlo experiment conducted in Judson and Owen (1999) also supported the use of the bias-corrected LSDV estimator in balanced panel data when N is small.

Some panel data observations can be missed or dropped during the observed time period, and this is known as unbalanced panel data. Focusing on unbalanced panel data, Bruno (2005b) extended the corrected LSDV estimator to dynamic unbalanced panel data by imposing a selection rule which selects the observations identified in both current time t and $t - 1$. The select rule is given by Equation (2.21).

$$s_{it} = \begin{cases} 1 & \text{if } (r_{it}, r_{it-1}) = (1,1) \\ 0 & \text{otherwise} \end{cases} \quad i=1,\dots,N \text{ and } t=1,\dots,T \quad \text{Equation (2.21)}$$

where r_{it} is the selection indicator.

Thus, the dynamic panel data model can be modified as Equation (2.22)

$$s_{it} y_{it} = s_{it} (\lambda y_{it-1} + \beta_1 x_{it} + u_{it}), \quad u_{it} = \alpha_i + \varepsilon_{it}, \quad i = 1, \dots, N, \quad t = 1, \dots, T \quad \text{Equation (2.22)}$$

Therefore, the LSDV estimator differs from Equation (2.18) by imposing the selection rule as shown in Equation (2.23).

$$\begin{aligned} \hat{\delta}_{LSDV} &= (W' A_S W)^{-1} W' A_S y \\ A_S &= S(I - D(D' S D)^{-1} D') S \end{aligned} \quad \text{Equation (2.23)}$$

where S refers to the $(NT \times NT)$ of usable observations excluding missing values and $D = I_N \otimes I_T$

Bruno (2005) suggested that in finite samples, if all the regressors other than the lagged dependent variable are exogenous, then the bias-corrected LSDV estimator is preferred to IV estimators. As the LSDV estimator is identical to the FE estimator with standard OLS in terms of the parameter estimates and standard errors, the corrected-bias method can be applied to both the LSDV estimator and the FE estimator with standard OLS in practice. However, this approach is only valid when the all the explanatory variables are exogenous, which is a rather strong assumption and is difficult to justify in practice. In addition, the corrected-bias LSDV method can only take account of time-series variation in the same way as the conventional FE and LSDV estimators do. Hence, this method has been rarely employed in applied panel data analysis, and thus is not considered for the empirical analysis of this study in the following chapters.

2.4.3 Pseudo panel data model and estimation

The pseudo panel data model introduced by Deaton (1985) is specified as follows:

$$\bar{y}_{gt} = \beta_0 + \beta_1 \bar{x}_{gt} + \bar{\alpha}_{gt} + \bar{\varepsilon}_{gt} \quad \text{Equation (2.24)}$$

Compared to the genuine panel data Equation (2.1), Equation (2.24) uses the subscript g instead of i to denote the created groups in the pseudo panel data. The variables \bar{y}_{gt} and \bar{x}_{gt} represent the way in which the variables are the mean values of each cohort. The critical element of Equation (2.24) that distinguishes the pseudo panel data model from the genuine panel data model is that the average unobserved group effect $\bar{\alpha}_{gt}$ is time-varying because the cohorts are constituted of different members although they are defined in the same group, whereas in genuine panel data the individual effect is time-invariant and denoted as α_i . The result is that the time-varying group effects will not be eliminated through the demeaned transformation in the FE estimation (as shown in Equation (2.12)), so the conventional FE estimator will be biased whether in the static or dynamic model when pseudo panel data are in use.

Deaton (1985) highlighted this by emphasising sample cohort means in pseudo panel data sets are consistent but “error-ridden” estimates of the true population means which are unobservable. Deaton proposed using an errors-in-variable estimator to estimate this population relationship. Verbeek and Nijman (1992) conducted an empirical analysis to compare the estimates obtained using the FE estimator with standard OLS for a genuine panel dataset and a pseudo panel dataset created from the genuine panel data. They found the difference between the estimates from genuine panel data and from pseudo panel data can be reasonably ignored if the cohort size is sufficiently large. Large in this context was a cohort size greater than one hundred individuals with smaller cohorts remaining reliable if they contained sufficient inter-cohort variation. This result implies that with sufficiently large cohorts the time-varying $\bar{\alpha}_{gt}$ can be treated as a constant over time as $\bar{\alpha}_g$, so that the pseudo panel data can be used in estimation as if they were genuine panel data, using conventional estimation techniques.

Given that one of the advantages of using pseudo panel data is that it permits longitudinal analysis where no genuine panel data exist, capturing the dynamics of the behaviour is particularly valuable from the use of pseudo panel data. Moffitt (1993) extended Deaton’s (1985) work to a dynamic context by using the partial adjustment dynamic model, concluding that a dynamic model can be consistently estimated with an IV-2SLS estimator, as discussed in the previous section.

McKenzie (2004) has more recently studied a dynamic model in a pseudo panel data context where the cohorts displayed inter-group heterogeneity. Using Monte Carlo simulation, he investigated dynamic model estimation using an OLS estimator and a GMM estimator separately. The simulation results demonstrated that a downward bias occurs for an OLS estimator if the cohort size is small, but this bias reduces if both cohort size and time periods become large. On the other hand, the bias from the GMM estimator is less severe but the estimates from the GMM estimator had more variability than the OLS estimator.

Verbeek and Vella (2005) further considered the estimation techniques for dynamic pseudo panel models. Monte-Carlo simulations identified that imposing time-varying instruments as in Moffitt (1993) resulted in severe estimation bias but this bias declined if the cohort size is larger than one hundred individuals and inter-group variation is present in the explanatory variables. Their conclusion was that a necessary condition of having a consistent estimation result in pseudo panel dynamic estimation is for the explanatory variables to have time-varying and inter-cohort variation as with genuine panel data.

Inoue (2008) proposed using GMM estimation to estimate dynamic pseudo panel data. He found that the GMM estimator was more precise than the FE estimator if the cohort size is large relative to the number of cohorts and the time periods. However, Inoue confirmed the finding from Verbeek and Vella (2005) that the identification condition may fail if the cohort-specific variation is not present.

In general, if the pseudo panel data can be created in a way that allows sufficiently large cohort size and inter-group variation, the literature suggests that pseudo panel data can be estimated as genuine panel data using conventional estimation methods. Table 2.8 summarises recent pseudo panel studies and the estimation techniques used in the analysis.

For studies using static models, most studies adopted the FE estimator to estimate the pseudo panel data (Gassener, 1998; Gardes et al., 2005; Huang, 2007; Weis and Axhausen, 2009; Warunsiri and McNown, 2010). These studies refer to the results of Deaton (1985) and Verbeek and Nijman (1992) which allows time-varying unobserved heterogeneity to be ignored if the cohort size is sufficiently large giving theoretically unbiased and consistent results using the FE estimator. Some studies also used other estimators and compared the results with the FE estimator. Gassner (1998) estimated the pseudo panel data with both the FE and RE estimator. The Hausman's test suggested that the FE estimator was preferred.

Table 2.8 Estimation Techniques in Previous Applied Pseudo Panel Studies

Author	Context of Study	Area	Observations ¹	Model	Estimation technique
Gassner (1998)	Telephone Access	UK	324 (G=27, T=12)	Static	FE; RE ²
Dargay and Vythoulkas (1999)	Car ownership	UK	165 (G=16, T=12)	Dynamic	FE; Pooled OLS; RE; RE-IV
Dargay (2002)	Car ownership	UK	134, 152, 159 ³ (G=15, T=14)	Dynamic	FE
Gardes et al. (2005)	Food consumption	US	90 (G=18, T=5)	Static	Between; FE; First-differences
Dargay (2007)	Car travel	UK	256 (G=16, T=20)	Dynamic	FE
Huang (2007)	Car ownership	UK	254 (G=16, T=19)	Static; Dynamic	Pooled OLS; FE
Weis and Axhausen (2009)	Travel demand	Switzerland	838 (G =140, T=7)	Static; SEM ⁴	FE
Warunsiri and McNown (2010)	Return to education	Thailand	220; (G =11, T=20) 440 ⁵ (G =22, T=20)	Static	Pooled OLS; FE; IV
Bernard et al. (2011)	Electricity	Canada	100 (G =25, T=4)	Dynamic	IV-dummy

¹Observations included in the estimation. The number of observations in some studies may be smaller than the product of T and G as a result of dropping cohorts less than one hundred individuals.

²The RE estimator is employed but rejected after the Hausman's test.

³Households are grouped based on the geographical locations. This study created 134 cohorts in rural areas, 152 cohorts in urban areas, and 159 cohorts in other areas.

⁴Structural Equation Model.

⁵220 cohorts from 2-year band age grouping and 440 cohorts from 1-year band age grouping.

Huang (2007) conducted a Likelihood Ratio Test and a Ramsey Regression Equation Specification Error Test (RESET) and concluded the FE estimator was more favourable than the pooled OLS estimator. Warunsiri and McNown (2010) compared the estimation results from the pooled OLS, FE, and IV estimators. They found that the pooled OLS estimator without cohort dummies had a downward estimation bias as a result of not controlling for unobserved heterogeneity. The FE estimator and IV estimator generated similar slope

coefficients but the standard errors from the IV estimator were larger than those from the FE estimation thus confirming the results of McKenzie (2004). This reported applied research for a static model suggests that the FE estimator appears to be a more appropriate estimator to estimate pseudo panel data as compared to the pooled OLS estimator.

Estimation of dynamic models using pseudo panel data has also been reported. As with genuine panel data, the lagged dependent variable is likely to be correlated with the error term and cause estimation bias. Dargay and Vythoukcas (1999) and Bernard et al. (2011) used the IV estimator to address this endogeneity problem and the results from Dargay and Vythoukcas (1999) show that the IV estimator should be chosen over the FE estimator.

2.5 Diagnostics and correction of panel data model assumptions

The estimation techniques for genuine panel data and pseudo panel data have been widely discussed in the literature. However, the diagnostics of the basic assumptions underpinning these estimation techniques which are routinely tested in genuine panel data analysis do not seem to have transferred to the pseudo panel data literature. This relates in particular to the structure and assumptions of the error term which are similar to multiple regression models. These are:

- (1) Error term ε_{it} is homoscedastic.
- (2) Error term ε_{it} is not serial correlated.
- (3) Error term ε_{it} is not cross-sectionally dependent.
- (4) Strict exogeneity so the error term is not correlated with explanatory variables.

Assumptions (1) to assumption (3) are captured by Equation (2.25) with assumption (4) by Equation (2.26).

$$E(\varepsilon_{it}\varepsilon_{js}) = \begin{cases} \sigma_\varepsilon^2 & \text{for } j=i \text{ and } t=s; \\ 0 & \text{otherwise} \end{cases} \quad \text{Equation (2.25)}$$

$$E[\varepsilon_{it} | x_{i1}, \dots, x_{iT}] = 0 \quad \text{Equation (2.26)}$$

These four assumptions apply to both static and dynamic models, and for both genuine panel data and pseudo panel data with assumption (4) being particularly important for dynamic models where the lagged dependent variable is included as an explanatory variable. In reported pseudo panel studies, discussion has focussed on time-varying unobserved heterogeneity with little consideration of the implications of the failure to meet the assumptions of the error term.

Reed and Ye (2011) showed that most genuine panel data suffer from serial correlation or cross-sectional dependence and conventional estimators cannot control both these effects (more generally referred as non-spherical errors). Ignoring the presence of non-spherical errors may cause biased or inefficient estimation, showing the importance of testing model assumptions to validate the estimation results.

In the literature concerned with genuine panel data analysis, some statistical methods have been developed to test model assumptions as summarised in Table 2.9. To test the homoscedasticity of the error term, Greene (2000) compared the Lagrange Multiplier (LM) test, likelihood ratio test, and standard Wald test, and proposed a modified Wald test for groupwise heteroscedasticity. For serial correlation, the Wooldridge's test is regarded as more robust than the Durbin-Watson test and the LM test (Wooldridge, 2002). The first test of cross-sectional dependence was introduced by Breusch and Pagan (1980) but this is only applied to panel data when the number of time periods exceed the number of panel units ($T > N$). It was extended by Pesaran (2004) to address the circumstance when N and T are both infinitely large and is known as the Pesaran Cross-section Dependence (CD) test. For strict exogeneity, Hausman's test can be used, as

suggested by Hayashi (2000) through a comparison of the results of a FE estimation and a FE-IV estimation.

Table 2.9 Panel Data Model Assumption Tests

Assumption test	Test Method	Developer
Heteroscedasticity	Modified Wald test	Greene (2000)
Serial correlation	Durbin-Watson test;	Durbin and Watson (1971);
	LM test;	Baltagi and Li (1991)
	Wooldridge test	Wooldridge (2002)
Cross-sectional dependence	Breusch-Pagan LM test ($T \gg N$)	Breusch and Pagan (1980);
	Pesaran CD test	Pesaran, M.H. (2004)
Strict exogeneity	Endogeneity test	Hayashi (2000)

To address heteroscedasticity, serial correlation, and cross-sectional dependence problems in panel data models, Reed and Ye (2011) summarised three estimators that can be used: OLS with robust standard errors, Feasible Generalized Least Squares (FGLS) and Panel-Corrected Standard Error (PCSE) estimator. If Equation (2.24) is the pseudo panel data model to be estimated, the first adjusts OLS estimators by imposing robust standard errors to control for heteroscedasticity, serial correlation, and cross-section dependence as described by Equation (2.27).

$$\hat{\beta} = (X'W^{-1}X)^{-1}X'W^{-1}Y$$

$$\text{Var}(\hat{\beta}) = (X'W^{-1}X)^{-1}(X'W^{-1}\hat{\Omega}W^{-1}X)(X'W^{-1}X)^{-1}$$

Equation (2.27)

where W is the weighting matrix and $\hat{\Omega}$ incorporates the estimated error variance-covariance matrix, X is the $T \times 1$ vector of observations of the exogenous explanatory variables, and Y is $T \times 1$ vector of observations of the dependent variable.

An extension of this is the FGLS introduced first by Parks (1967). The FGLS estimator uses a similar estimation formula as the standard OLS as shown in Equation (2.28).

$$\begin{aligned}\hat{\beta} &= (X' \hat{\Omega}^{-1} X)^{-1} X' \hat{\Omega}^{-1} Y \\ \text{Var}(\hat{\beta}) &= (X' \hat{\Omega}^{-1} X)^{-1}\end{aligned}\tag{2.28}$$

The difference between the standard OLS estimator and FGLS estimator is that the FGLS estimator allows for group-wise heteroscedasticity, first order serial correlation and time-invariant cross-sectional dependence. The FGLS estimator involves two transformations: first to eliminate the serial correlation by the OLS estimator, and second to correct the cross-sectional dependence by using the residuals from the first estimation.

Beck and Katz (1995) have demonstrated the FGLS estimator can only be applied when $T > N$ using a Monte Carlo simulation. However, even when $T > N$, FGLS tends to underestimate standard errors and thus inflate the confidence in the estimated parameters. Beck and Katz's simulations showed that the underestimation of standard errors with FGLS was most severe when $T = N$, with overconfidence of parameter significance reaching 408 percent when $T = N = 10$, as compared to the OLS estimator. A Panel-Corrected Standard Error (PCSE) estimator was developed by Beck and Katz (1995) to address this problem and was shown to perform better than FGLS. The PCSE estimator can be described as Equation (2.29).

$$\begin{aligned}\hat{\beta} &= (\tilde{X}' \tilde{X})^{-1} \tilde{X}' \tilde{Y} \\ \text{Var}(\hat{\beta}) &= (\tilde{X}' \tilde{X})^{-1} (\tilde{X}' \hat{\Sigma} \tilde{X}) (\tilde{X}' \tilde{X})^{-1}\end{aligned}\tag{2.29}$$

$$\begin{aligned}\tilde{Y} &= Y_t - \hat{\rho} Y_{t-1} \\ \tilde{X} &= X_t - \hat{\rho} X_{t-1}\end{aligned}\tag{2.30}$$

where \tilde{X} and \tilde{Y} are the transformed vectors of independent variables and dependent variable using Prais-Winsten transformation (Equation (2.30)) to control for serial correlation ($\hat{\rho}$) of the error term (Prais and Winsten, 1954). $\hat{\Sigma}$ is the variance matrix which is estimated by the residuals from the OLS estimation.

The literature shows that the presence of any heteroscedasticity, serial correlation, or cross-sectional dependence does not lead to biased coefficients but instead gives estimation inefficiency that makes the statistical inference unreliable. In contrast, the presence of endogeneity causes biased estimates. Whilst the literature for genuine panel data analysis also has developed estimation techniques to correct or allow estimation in the presence of relaxed model assumptions, these have not been highlighted or used in applied pseudo panel studies.

Dargay and Vythoukas (1999) noted the error terms in pseudo panel data estimation are likely to be heteroscedastic because the number of observations in each cohort is different. This was corrected by weighting all variables by the square root of the number of observations in each cohort (i.e., Weighted Least Squares method) in this study. Dargay (2007) also identified that the potential presence of serial correlation would lead to inconsistent estimates and tested for the presence of serial correlation but found none. Huang (2007) corrected for homoscedastic errors weighting the variables by the square root of the cohort size (as undertaken by Dargay and Vythoukas, 1999) but further investigated potential outliers by carrying out an analysis of residuals. Bernard et al. (2011) instrumented the error term to control for serial correlation and heteroscedasticity, but the analysis assumed no cross-sectional dependence in the error term between the defined groups which would require further validation. All studies using the Weighted Least Squares (WLS) method have not demonstrated how this eliminates the non-spherical errors, as they did not undertake or present further model assumption tests after the WLS estimation.

The discussion above suggests that violations in the error term assumptions in genuine panel data analysis can lead to biased or inefficient estimation. However, the error term assumptions have not been as rigorously examined in the applied pseudo panel data studies as in genuine panel data analysis. This review highlights the importance of model assumption diagnostics and correction techniques that has not been comprehensively applied in previous pseudo panel studies.

2.6 Research gaps and summary

The literature review on public transport demand elasticities and land use studies suggests that an ideal travel demand model should take three elements into account: the temporal effect, land use characteristics, and individuals' behaviour. Previous studies using either an aggregate modelling approach or a disaggregate modelling approach have not yet integrated these three elements in one study.

Aggregate modelling approaches are able to investigate time effects based on aggregate time-series data. However, few aggregate studies include multiple land use variables in demand models, and individuals' travel behaviour cannot be identified through aggregate data. These shortcomings can be overcome by a disaggregate modelling approach, because disaggregate models can incorporate more information about individuals' location characteristics as well as individuals' choices of travel. However, disaggregate modelling, as reported in the literature, has not properly taken account of the temporal effect of demand changes.

The limitation in disaggregate data seems to arise from a lack of data at the appropriate level of disaggregation to conduct a time-series study to determine the long-run demand elasticities. Given that land use factors can generate long-term effects on travel behaviour, the approach applied to investigate public transport demand elasticities which takes account of land use characteristics needs to have the ability to identify the temporal effect, and distinguish between short-run and long-run elasticities.

As genuine panel data are rarely available, pseudo panel data analysis provides a potential solution to the research gap identified above. Pseudo panel data, derived from repeated cross-sectional data can be established from travel surveys of individuals, creating cohorts which allow the temporal effect to be identified. A pseudo panel approach creates a dataset which allows both static and dynamic econometric models to be applied at a cohort level whilst incorporating the potential for behavioural grounding. Therefore, combining a disaggregate

modelling approach and pseudo panel data analysis can be a potential method to investigate the effect of time and land use on individuals' travel behaviour.

This study applies the pseudo panel data approach to identify the short-run and long-run public transport demand elasticities in Sydney by constructing a dynamic public transport demand model incorporating a comprehensive land use dataset together with price, socio-economic variables, as well as the supply variable as identified being important in the public transport demand literature. The dataset including all the variables are introduced in Chapter 3.

The review on previous pseudo panel data studies highlights the current practices of pseudo panel data approach in transport literature. Although the pseudo panel approach has been increasingly applied in travel demand analysis, applications to public transport demand have not yet been evident in the literature. As discussed in Section 2.3.4, when applying the pseudo panel approach to public transport demand studies, the dependent variable used to represent the public transport demand need to possess sufficient variation across defined groups. This is investigated in Chapter 4 in which the pseudo panel dataset is presented and the inter-group variation of the variables is examined. The other issue highlighted in Section 2.3.4 is the price variable for public transport. The public transport price variables used for previous pseudo panel studies on car travel demand were generic across groups, which limited the explanatory power of effects of price variation on the demand change. This study derives the public transport price variable from the household travel survey data, which is specific to each single trip and individual traveller. The inter-group variations of the price variable as well as other explanatory variables are also examined in Chapter 4.

Another constraint in applying pseudo panel analysis to public transport demand is the lack of sufficient public transport observations in most study areas, because conventionally the pseudo panel data construction requires at least more than one hundred members in each cohort as a rule of thumb. This constraint in turn limits the statistical power of conventional estimation techniques such as

the IV estimator that requires a large sample size to be efficient. As a result, the FE estimator becomes the most commonly applied technique to estimate pseudo panel data models although it is not able to take account of the between-group variation, which is usually substantial in a pseudo panel dataset. This indicates that the estimation techniques for pseudo panel data models need to be further examined, especially for applied pseudo panel datasets that possess some unique properties different from genuine panel data. In Chapter 5, a Monte Carlo simulation experiment is presented to evaluate the performance of various estimators under the scenarios of applied pseudo panel data. The simulation results in turn provide suggestions for empirical model estimations which are presented in Chapter 6 and Chapter 7.

CHAPTER 3 DESCRIPTION OF CASE STUDY

3.1 Introduction

Chapter 1 and Chapter 2 have discussed the research questions of this study and the research gaps identified in the literature. This chapter introduces the study area of this study and presents an exploratory analysis on the association between public transport demand and land use characteristics. The Sydney Greater Metropolitan Area (SGMA) is chosen as the study area and its general demographic and geographic characteristics as well as the public transport network in the SGMA are presented in Section 3.2. Section 3.3 summarises the data sources collected for this study and defines the variables of the dataset. Section 3.4 presents a preliminary analysis on the relationship between public transport demand and its explanatory variables using a Geographically Weighted Regression (GWR) approach¹. This exploratory analysis investigates the variation of public transport demand in the SGMA with respect to the explanatory variables hypothesised to be important whilst taking account of geographical variations in order to define the geographical boundaries for the pseudo panel data analysis in Chapter 4.

3.2 The Sydney Greater Metropolitan Area

3.2.1 *Demographics and geography*

Sydney is the most populous city in Australia and the state capital city of New South Wales. The metropolitan area of Sydney is defined by the Sydney Greater Metropolitan Area (SGMA), which comprises Sydney Statistical Division (Sydney SD), Illawarra Statistical Division (Illawarra SD), and Newcastle Statistical Subdivision (Newcastle SSD) as shown in Figure 3.1. The total geographical coverage of the SGMA is summarised in Appendix 1 (Table A1.1). The SGMA is chosen as the study area because it is the geographical coverage of Sydney Strategic Travel Model which is the transport planning model used to project and

¹ A journal paper based on this GWR analysis has been published in Tsai et al. (2012). This paper was previously presented in the 35th Australasian Transport Research Forum and won the David Willis Memorial Prize which is awarded to the best paper conducted by a young professional in the conference. The author wishes to acknowledge the contribution of the co-authors- Corinne Mulley and Geoffrey Clifton.

predict travel patterns with respect to strategic land use planning and transport planning operated by Bureau of Transport Statistics (2011e). Sydney SD, where the Sydney Central Business District (CBD) is located, is the core business and activity centre of the area with highest number of population and employments in NSW. Illawarra SD and Newcastle SSD are located to the south and the north of Sydney SD with local labour markets developed in both regions.



Figure 3.1 The Sydney Greater Metropolitan Area
Source: Bureau of Transport Statistics (2011e)

The population, area size, and population density of each division are summarised in Table 3.1. In 2010, the total population in the SGMA is around 5.56 million with the majority of people residing in Sydney SD. The total area of the SGMA is 24,499 km² with the average population density of 227

(persons/km²). Of the three statistical divisions in the SGMA overall, Sydney SD is the densest area with the highest population density of 377 persons/km², followed by 135 persons/km² in Newcastle SSD and 52.5 persons/km² in Illawarra SD (Australian Bureau of Statistics, 2011b).

Table 3.1 Demographics of the Sydney Greater Metropolitan Area

Division	Population (persons)	Area (km²)	Population Density (persons/km²)
Sydney SD	4,575,532	12,138	377
Newcastle SSD	546,788	4,052	135
Illawarra SD	436,117	8,309	53
Total	5,558,437	24,499	227

Source: Australian Bureau of Statistics (2011a)

3.2.2 Public transport in the Sydney Greater Metropolitan Area

Table 3.2 summarises the general statistics of travel mode split in the SGMA between July 2010 and June 2011 (2010/2011). The major mode of travel in the SGMA is private vehicles including vehicle driver and vehicle passenger trips, which collectively take account of 70.6 percent of the mode share, followed by walk only trips sharing 17.4 percent of the total trips. Train and bus as the two major public transport systems share around 9.8 percent in the SGMA. Of the three divisions, Sydney SD has the highest number of trips as a result of the higher population and employment. Comparing the mode share among the three statistical divisions, Sydney SD has the higher share of train, bus, and walk only trips as compared to the other two divisions and the public transport network is more intensive in Sydney SD. Public transport trips are substantially lower in Newcastle SSD and Illawarra SD, so the total public transport mode share of the SGMA is mostly driven by Sydney SD.

Train and bus form the major public transport system in the SGMA. The train system is solely operated by CityRail which operates 1,595 km of mainline tracks which also provide services between Newcastle SSD and Illawarra SD via Sydney SD. Most areas within the SGMA are served by local buses. The inner Sydney areas in Sydney SD are mainly served by Sydney Buses which is a state-owned agency. In other outer suburbs, bus services are commonly contracted to private bus companies. Other public transport systems, including ferries and light rail, share less than one percent of total trips, so they are not included in this study.

Table 3.2 Statistics of Trip Modes in the SGMA in 2010/2011

Trip Mode		Sydney SD	Newcastle SSD	Illawarra SD	SGMA
Vehicle driver	Trips ('000)	8,062	1,234	938	10,326
	Mode Share	46.9%	59.1%	54.0%	48.9%
Vehicle passenger	Trips ('000)	3,653	489	418	4,575
	Mode Share	21.2%	23.4%	24.1%	21.7%
Train	Trips ('000)	920	14	25	960
	Mode Share	5.3%	0.7%	1.4%	4.5%
Bus	Trips ('000)	1,007	68	46	1,118
	Mode Share	5.9%	3.3%	2.7%	5.3%
Walk only	Trips ('000)	3,153	249	266	3,667
	Mode Share	18.3%	11.9%	15.3%	17.4%
Other modes	Trips ('000)	407	34	43	485
	Mode Share	2.4%	1.6%	2.5%	2.3%

Source: Bureau of Transport Statistics (2012e)

The CityRail network was first opened in 1855. The rail network was substantially completed before 1997, and only three new lines have been added in the network since 1997 as summarised in Table 3.3 with a map of new extension lines in Figure 3.2. The Olympic Park Line was built for the Sydney Olympic Games in 2000 with a five-kilometre track connecting Lidcombe and Olympic Park Stations. The Airport Line is a Public Private Partnership project which provides a rail link between Central Station and Sydney Airports (including the international and domestic terminals) located to the south of Sydney CBD. The Airport Line is operated by a private company named Airport Link which operates Mascot, Green Square, Domestic Airport, and International Airport stations. The latest network extension was in February 2009 when the Epping-Chatswood Line was opened. This 12.5-kilometre rail link connects western north and eastern north of Sydney. Compared to the total 1,595-kilometre network length of CityRail, the three new lines opened after 1997 do not share a large proportion of total network size so that train supply has not changed dramatically since 1997.

Table 3.3 CityRail Network Changes since 1997

Date	Opening	Length
27 April 1999	Olympic Park Line (Lidcombe Station-Olympic Park Station)	5 km
21 May 2000	Airport Line (Central Station- Wollsi Creek Station)	7.3 km
23 February 2009	Epping to Chatswood Rail Link (Epping Station to Chatswood Station)	12.5 km

Source: Summarised from Rail Corporation New South Wales annual reports
(http://www.railcorp.info/publications/annual_reports)

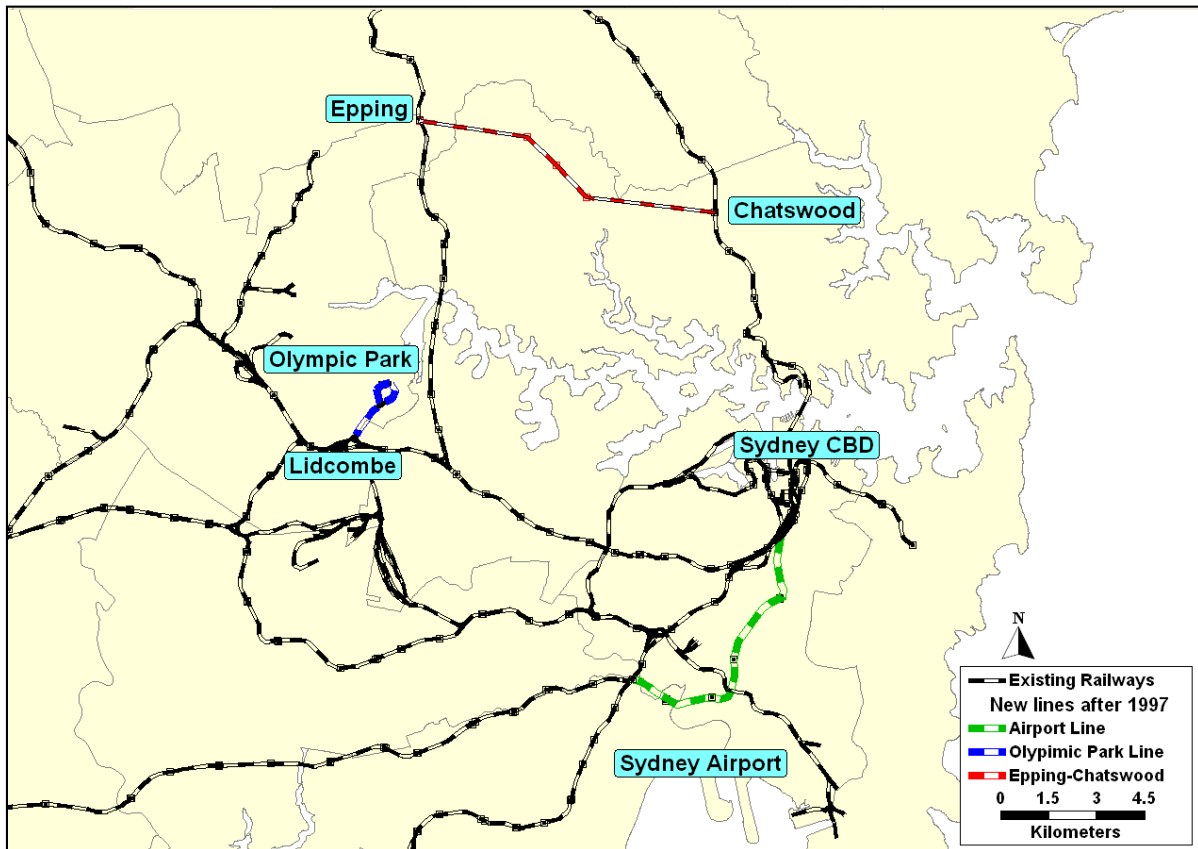


Figure 3.2 Extended Railway Lines of CityRail after 1997

Source: Developed from Sydney GIS layers

The bus services in Sydney are supplied by both public and private operators. The public bus service is operated by State Transit Authority of New South Wales, which is a government owned authority responsible for the operations of Sydney Buses, Newcastle Buses and Ferries, and Western Sydney Buses (known as Liverpool-Parramatta Transitway commenced in February 2003). In other areas apart from STA service coverage, bus services are contracted to private bus companies, and both public and private bus trips are taken into account of this study.

The introductions of the case study presented above highlight the key socio-demographic characteristics, travel patterns, and general urban forms of the SGMA. In general, the urban development in the SGMA is planned based on the CBD as the core of business and activities centres and is gradually expanded to outer areas. The outer areas are supplied by commuter trains to access the CBD and its surrounding business centres, and the outer areas are also served by local buses for the need of local mobility and accessibility to the trip destinations where the train service is not available. As population grows, the outer areas have moderately developed their local labour markets but still with high travel demand to the CBD. This type of urban form and public transport supply is commonly seen in most metropolitan cities in Australia. Although the outcomes of this study cannot be perfectly generalised to other areas, the investigation of the associations between public transport demand and land use characteristics on the SGMA may reasonably provide practical relevance to other cities with similar urban forms and travel patterns.

3.3 Data description

3.3.1 Data sources

The dataset used in this study consists of travel-related data and land use data. The travel-related data including public transport demand and trip price, as well as socio-economic variables of public transport users, are retrieved from the Sydney Household Travel Survey (SHTS) collected by Bureau of Transport Statistics (BTS). The SHTS has been undertaken continuously since 1997/1998, with approximately 8,500 people in 3,500 households recruited annually (Bureau of Transport Statistics, 2011b). The SHTS comprises data about individuals' travel behaviour from a one-day travel diary that records each single trip with related information such as trip modes. The SHTS also includes a household form and personal form to collect the socio-economic information of households and household members. This database provides consistent repeated cross-sectional data with comprehensive travel related information.

The land use data are collected from Australian Census conducted by the Australian Bureau of Statistics (ABS) and Geographical Information System

(GIS) layers of the road network in the SGMA. The Australian Census is conducted at five-year intervals and the most recent data available for this study in 2006 are merged into the SHTS database by matching geographical codes. The road network data used to retrieve other land use variables and accessibility measures are based on the 2010 road network GIS layers provided by BTS. Whilst the most recent census data has typically been used for the land use variables (apart from where noted in Section 3.3.2), sensitivity analysis has been undertaken which compared the change in these variables over all possible census data for the time period covered by the dataset. This sensitivity analysis concluded that changes in land use variables happened only slowly so that the use of 2006 and 2010 data respectively does not introduce significant error. A further discussion about the sensitivity analysis is presented in Section 4.3.2.

The geographical locations of the individual data from the SHTS and Census are only available at the Census Collection District (CD) level as the finest aggregation level. The geographical coordinates of household locations are desirable for this study but not available due to confidentiality issues. This study uses Travel Zones (TZ) as the aggregation level for most of the variables that require geographical information such as land use variables, except for a few variables available at household level provided by BTS as noted in the next section. The CD level is considered to be too disaggregate to have sufficient variations within each CD, and the TZ is more consistent to the strategic transport planning for the SGMA as TZ is designed in a way to create homogeneous areas in terms of travel patterns, land use characteristics, and public transport supply.

3.3.2 Definitions of variables

Dependent variable

In this study, public transport demand as the dependent variable of the demand model is defined by the number of public transport (bus and train) trips made by a traveller per day. Other modes of public transport in the SGMA including ferry, light rail, and monorail are excluded from the public transport sample because they only account for around 2.4 percent of total trips in the SGMA collectively as

shown in Table 3.2, and the service coverage of these modes are restricted to a certain number of Travel Zones and hence the demand and its association with explanatory variables may not be representative of the whole SGMA.

As discussed in Section 2.3.4, the definition of public transport demand must be given careful consideration. Compared to the public transport demand defined in some previous studies, which conventionally defined public transport demand by passenger patronage (Dargay and Hanly, 2002, Cervero, 2006) at an aggregate level, this measure is used because it can be calculated for each individual in the SHTS and thus the associations between public transport demand and the explanatory variables can be investigated at an individual level. For further investigation of the demand at a higher aggregation level, this measure can be aggregated to the SGMA by weighting it to the total population.

Independent variable

As identified in Section 2.1 and 2.2, the literature on travel demand suggests that the determinants of public transport demand should include price, socio-economic factors, public transport supply, and land use factors.

Price

The impact of price on demand provides important policy implications for transport operators and government sectors when setting public transport fare policy. There are several ways of estimating public transport price for public transport demand models. Studies based on panel data across various systems usually calculate public transport trip price by dividing the total fare revenues by total patronage of a system (Dargay and Hanly, 2002, Graham et al., 2009). Studies based on Stated Preference (SP) surveys using disaggregate data typically use public transport fare prices for demand modelling (Hensher, 1998). However, neither of the two common measures is able to represent a specific price for each journey. Considering the importance of individual information for each single public transport trip made in the SGMA, the public transport price in this analysis is calculated for each single public transport trip by dividing the total ticket price reported by the respondents in the SHTS by the total number of

trips provided by this ticket. There are many ticket types in Sydney such as single tickets, return tickets, and periodical tickets, but SHTS respondents only report their purchased ticket prices and ticket types and thus a ticket journey multiplier is employed to assume the average number of trips for periodical tickets (Table 3.4) to approximate the ticket price for each single trip. The use of the reported trip price from the SHTS as opposed to the average public transport fare allows the price variable to vary across the observations rather than being fixed at an aggregate level.

Table 3.4 Ticket Journey Multipliers

Ticket Type	Number of Trips
Single	1
Return	2
Weekly	11
Monthly	48
Quarterly	144
Yearly	585

Source: CityRail (2010)

The alternative cost of a public transport trip, in terms of fuel price, has been investigated in this analysis but was found to be insignificant in the model, and thus it was removed from the dataset to improve the model degrees of freedom.

Socio-economic variables

Socio-economic factors have also been suggested as important explanatory variables of public transport demand. The socio-economic variables in this study include the annual personal income and the age of the public transport user which are both retrieved from the SHTS database. The inclusion of socio-economic factors in the demand model is not only used to explain the variation of public transport demand, but also to mitigate the self-selection problem when attitudinal data are not available (Zhang, 2011).

Car ownership, although suggested as relevant to travel behaviour in the literature, is not included in this study because car ownership is recorded at a household level in the SHTS giving a confounding relationship between individual public transport usage and the number of cars available in a household. For example, young students may only choose to use public transport

regardless the number of cars available for use in their households. Car ownership was investigated in the initial model but found insignificant so it was removed from the dataset as a result.

Public transport supply

Bus frequency is used to control for public transport supply and quality of public transport service in the SGMA. Bus frequency is computed from the number of bus services in each bus stop during peak hours 6am to 10am on a typical weekday, and then the bus frequencies at bus stops are aggregated to a 400-meter buffer of a TZ centroid, which is considered as a reasonable walking distance from household locations to bus stops and the rule of thumb of service planning guidelines. Train frequency is not included in the dataset because it is highly correlated with bus frequency (Pearson's Correlation Coefficient: 0.82) and also because bus services are more accessible than train stations for most of the population and the train network does not cover the whole study area.

Land use variables

The impact of land use characteristics on travel demand can be identified through land use density, diversity, design and accessibility. Land use density in this study is defined by population density in terms of number of populations within 800 meters of a TZ centroid. This measure is used instead of population per squared kilometre of each TZ in order to control for the impact of TZ sizes on density. Of the total 2,742 TZs within the SGMA, the mean TZ area size is 3,623 km² with a large standard deviation of 35,217 and a median of 103 km². This suggests that the TZ boundaries do not necessarily represent the residents' activity areas. Hence, an 800 meter buffer, which is assumed to be a maximum walking distance for a traveller to access business and recreational activities, is used to standardise the population density measurement. Employment density was initially investigated but eventually removed from consideration because of its strong correlation with population density.

The entropy of land use types is used to measure land use diversity. The entropy derived from Equation (3.1) represents the diversity of land use in a TZ. Entropy

has a value between zero and one with zero representing extremely homogenous land use, and an entropy of one indicates the land use is equally heterogenous across all land use types. The entropy of the land use mix is measured from four land use types including agricultural and parkland, commercial, residential, and other.

$$ENTROPY = -\sum_{i=1}^n M_i * (\ln M_i / \ln Q) \quad \text{Equation (3.1)}$$

where M_i =proportion of land use type i in a TZ

Q =total number of land use types

Land use design can be observed from the connectivity and walkability of a local built environment. This study uses the number of pseudo nodes within 800 meters of a TZ centroid to measure land use design. Pseudo nodes are retrieved from the GIS layers of the road network by measuring the curvature and the number of dead ends in the built environment, and the denser the pseudo nodes, the curvier and more disconnected are the roads. Figure 3.3 and Figure 3.4 illustrate the composition of pseudo nodes in the GIS layer which represent two contrasting walking environments. In general, a built environment with more curvy roads and more cul-de-sacs has more pseudo nodes (Figure 3.3) than an area with a grid network (Figure 3.4). The hypothesis of the relationship between public transport demand and land use design is that the relationship would be negative for pseudo nodes, that is, public transport demand is higher in areas with fewer pseudo nodes. This hypothesis is based on previous studies of the relationship between travel behaviour and built environment which found that people tend to drive less and walk or use public transport more in areas with fewer cul-de-sacs (Cervero and Kockelman, 1997, Rajamani et al., 2003). The higher degree of road curvature is expected to have a negative impact on walkability and connectivity because travellers will need to walk further and more indirectly to public transport stations or trip destinations.

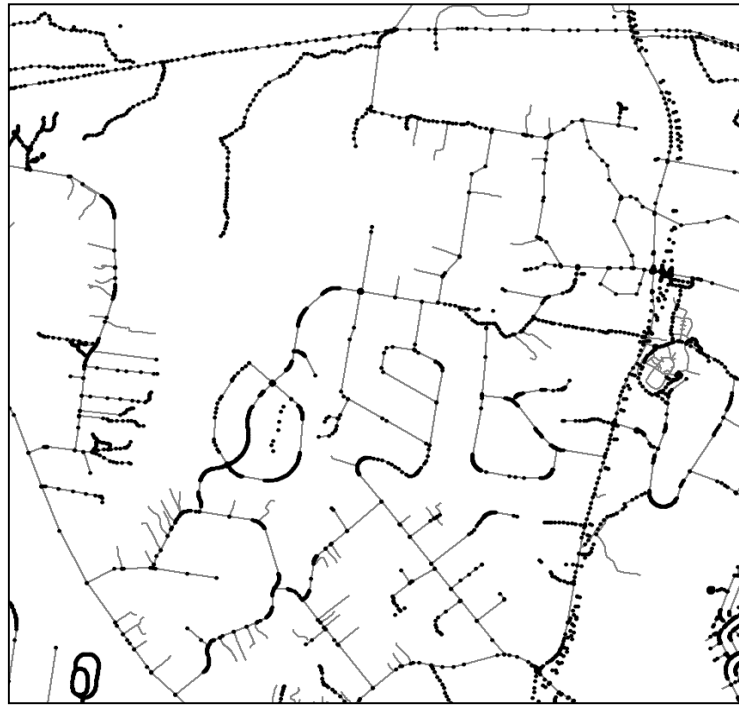


Figure 3.3 Pseudo Nodes in a Cul-de-sac Built Environment

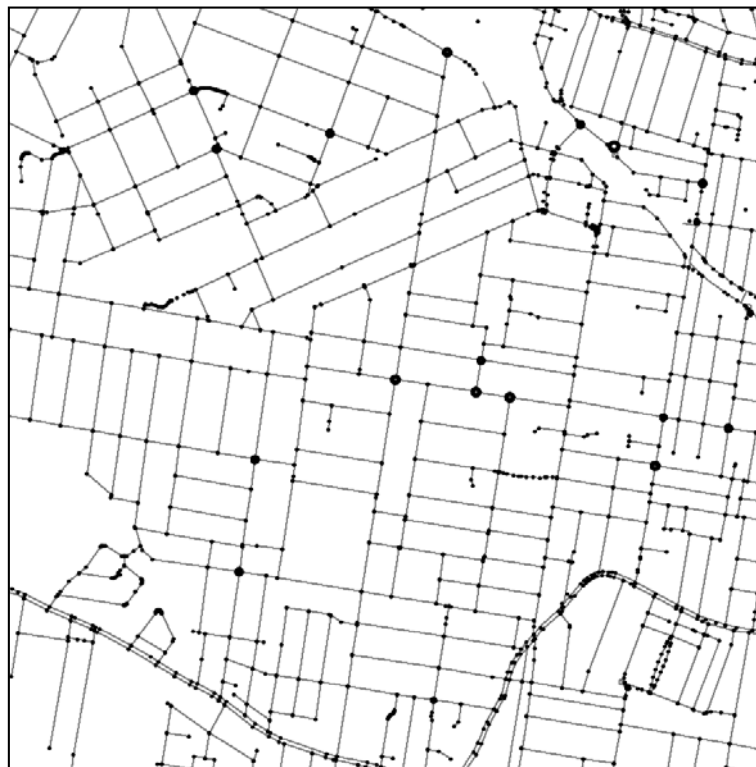


Figure 3.4 Pseudo Nodes in a Built Environment with a Grid Road Network

A broader definition of accessibility is the accessibility to trip destinations or major activity attractions, and access to local public transport stations. The former measure is captured by the distance to Sydney CBD measured by the road distance between the centroid of a TZ to the CBD. This measure is considered to

be important in explaining public transport demand in the context of Sydney, because the major public transport network is focused on accessing the Sydney CBD. For a local access measure, the walk distance between household locations and the nearest train station or bus stop has been suggested influential for public transport demand. This variable is provided by Bureau of Transport Statistics for every household recruited in SHTS. In addition, the number of train stations and bus stops is also used as a proxy of local public transport access. This is calculated as the total number of train stations and bus stops within 800 meters of the household location.

The dataset including price, socio-economic variables, public transport supply, and land use variables covering land use 3D and accessibility with their hypothesised relationships to public transport demand are summarised in Table 3.5. The next section presents a preliminary analysis on the relationship between public transport demand and the explanatory variables at a TZ level. This micro-level analysis allows for the identification of a global relationship between public transport demand and the explanatory variables in the SGMA using a multiple regression as well as the spatial variation in space within the study area using a Geographically Weighted Regression (GWR) method.

Table 3.5 Summary and Descriptive Statistics of Variables

Variable	Description	Unit	Hypothesis	Source
<i>Dependent variable</i>				
PTTRIP	No. of public transport trips per person	Trips/person	n/a	SHTS
<i>Price variable</i>				
PRICE	Public transport trip price	Dollars (AUD)	Negative	SHTS
<i>Socio-economic factors</i>				
INCOME	Annual personal income	Thousand dollars (AUD)	Negative	SHTS
AGE	Age	Years	Negative	SHTS
<i>Public Transport Supply</i>				
BUS FREQUENCY	Number of buses serving a bus stop between 6am and 10am on Tuesday within 400 meters of a TZ centroid	Thousands	Positive	BTS
<i>Land use density</i>				
POPULATION DENSITY	Population within 800 meters of a TZ centroid	Thousands	Positive	Census
<i>Land use diversity</i>				
LANDMIX	Entropy of land use mix	n/a	Positive	Census
<i>Land use design</i>				
PSEUDO NODES	Number of pseudo nodes within 800 meters of a travel zone centroid	Thousands	Negative	Road network
<i>Accessibility</i>				
DISTACNE TO CBD	Distance between CBD and travel zone centroids	meter	Negative	Road network
DISTANCE TO PT STOP	Distance between households and the nearest train station or bus stop	meter	Negative	Road network
PT STOPS	Number of train stations and bus stops within 800 meters of a household	n/a	Positive	Road network

3.4 Exploratory analysis

This section presents an exploratory analysis on the relationship between public transport demand and its explanatory variables including price, socio-economic factors, public transport supply, and land use factors, using data introduced in Section 3.3. This exploratory analysis is conducted to identify the variation of

public transport demand with respect to the geographical information in the SGMA, with preliminary findings being used to underpin the pseudo panel data construction and analysis in Chapter 4.

Although the relationship between public transport demand and land use has been identified in the existing literature, there is a lack of micro-analysis which comprehensively incorporates all the land use factors (3D and accessibility) in a public transport demand model, together with other key determinants such as price and socio-economic factors. Previous studies, based on the regional level (e.g., cities, states, or countries), have not been able to provide insights into the spatial variation of different variables across local communities on public transport demand within a specific study area. Moreover, the relationship between demand and land use has been conventionally examined at an average level assuming a homogenous parameter across all observations, without taking the spatial variability of land use variables into consideration. Spatial variability is important because if it exists it indicates a heterogeneous association between public transport demand and its determinants in the local areas, which provides important policy implications for local transport and urban planning. Spatial variability is particularly considered relevant when analysing land use data as shown by Wang et al. (2011) with land use factors tending to be correlated across space. The use of a Geographically Weighted Regression (GWR) methodology to identify the spatial variation addresses the issue.

The GWR approach which can effectively capture the spatial variability of parameter estimates was developed by Fotheringham et al. (2002). GWR consists of a global model and a local model. The global model is essentially a multiple regression model, whereas the local model takes account of the spatial dependency in the estimation process by weighting the observations according to their geographical locations. The first application of GWR in transport research was by Du and Mulley (2006) who investigated the association between land use value and public transport demand in the Tyne and Wear region in the United Kingdom. In modelling travel demand, Mulley and Tanner (2009) applied the GWR approach to model household vehicle kilometres travelled (VKT) in Sydney.

In terms of modelling public transport demand, Chow et al. (2006) used a mixed GWR approach to predict public transport demand in Broward County of Florida using accessibility to employment, car ownership, employment density, and the composition of population.

This analysis applies the GWR methodology to model public transport demand in the SGMA at a TZ level. A global public transport demand model incorporating comprehensive land use variables as well as price and socio-economic factors is constructed to identify the average relationship in the SGMA, and GWR local models are estimated to investigate the spatial variability of these relationships.

3.4.1 Introduction of Geographically Weighted Regression

The global public transport demand model constructed for this analysis hypothesises that public transport demand (Y) in a TZ (i) is determined by public transport trip price (P_i), a vector of socio-economic factors (E'_i), a measure of public transport supply and quality of service (S_i), and a vector of land use variables (L'_i) and an independent error term (ε_i) as specified in Equation (3.2). A linear functional form is chosen in this analysis as it is an exploratory investigation. A linear model implies that elasticities vary with the corresponding public transport demand and the explanatory variable concerned. The elasticity of variable k , evaluated at the mean, can be derived from the average demand (\bar{Y}) and explanatory variables (\bar{X}) by Equation (3.3).

$$Y_i = \beta_0 + \beta_1 P_i + \beta_2 E'_i + \beta_3 S_i + \beta_4 L'_i + \varepsilon \quad \text{Equation (3.2)}$$

$$\bar{e}_k = \frac{dY}{dX_k} \cdot \frac{\bar{X}_k}{\bar{Y}} = \beta_k \cdot \frac{\bar{X}_k}{\bar{Y}} \quad \text{Equation (3.3)}$$

A drawback of the global multiple regression model is that all observations in the study area are equally weighted in the estimation process, and thus the relationship between the dependent variable and independent variables is homogenous. This assumption is over-simplified when there is spatial heterogeneity across the observations. Therefore, the local model of GWR is used

to accommodate spatial effects by allowing for heterogenous parameters for observations located in different geographical coordinates (u_i, v_i) as shown in Equation (3.4).

$$Y_i(u_i, v_i) = \beta_0(u_i, v_i) + \beta_1(u_i, v_i)P_i + \beta_2(u_i, v_i)E_i + \beta_3(u_i, v_i)S_i + \beta_4(u_i, v_i)L_i + \varepsilon \quad \text{Equation (3.4)}$$

The local model employs a kernel weighting scheme in the estimation process as defined in Equation (3.5). This analysis uses an adaptive weighting approach which allows a same number of observations for each estimation when the observations are not regularly distributed in geographical space (Figure 3.5), so the bandwidth of a kernel (Figure 3.6) is larger when the observations are sparser and is smaller when the observations are densely clustered (Charlton and Fotheringham, 2009). The estimated parameters for a location (i) are more influenced by its surrounding areas than areas further away within a given bandwidth. As a result, the estimated parameters in different locations vary over geographical space to capture spatial heterogeneity.

$$w_i(u_i, v_i) = \left(1 - \left(\frac{d_i(u_i, v_i)}{h}\right)^2\right)^2 \quad \text{Equation (3.5)}$$

where

w_i = geographical weight for an observation i

d_i = distance between the i^{th} observation and the location (u_i, v_i)

h = bandwidth

In short, the global model provides the general relationship between the dependent variables and its determinants without taking the spatial variation into consideration, so the results can only be interpreted as an average value for the study area. In contrast, the local model investigates the spatial heterogeneity of this relationship for each observation, and the results can be projected to GIS maps to visualise this effect. The performance of the local model as compared to the global model is identified through the Akaike Information Criterion (AIC) and adjusted R-squared value. The AIC is used to optimise the bandwidth in the estimation, where the bandwidth with the lowest AIC is used for estimating the

models, and a model with a lower AIC represents a better goodness of fit (Charlton and Fotheringham, 2009).

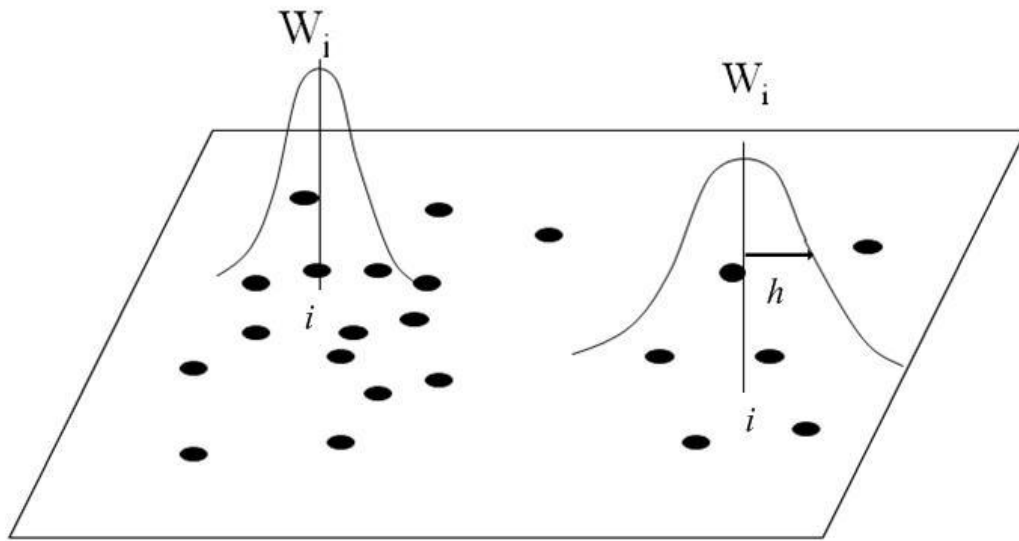


Figure 3.5 Adaptive Kernels in Local Model Estimation
 Source: <http://ncg.nuim.ie/ncg/gwr>

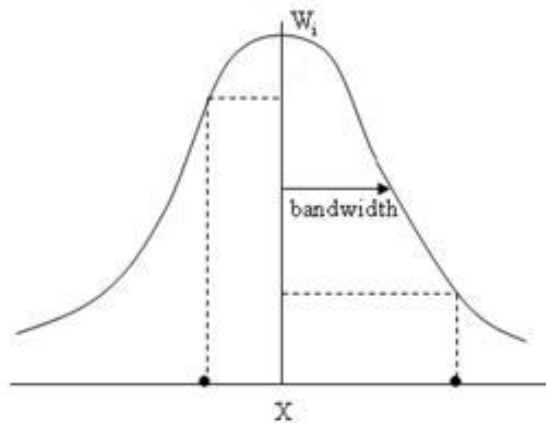


Figure 3.6 Bandwidth of a Kernel
 Source: <http://ncg.nuim.ie/ncg/gwr>

3.4.2 Global model estimation

The descriptive statistics of the variables in the dataset are presented in Table 3.6. This analysis only considers TZs where public transport trips are identified between 1997 and 2009 so the number of observations is 1824 TZs. Around 900 TZs with no public transport identified in the SHTS are excluded from this dataset because there is no information on trip price, age, and income. Some

variables show very high standard deviations which indicate that spatial variation appears to be present across TZs since each observation represents a TZ, and this effect is more substantial in the land use variables such as pseudo nodes, distance to CBD, and distance to the nearest public transport stops. This corresponds to Wang et al. (2011) as identified earlier, in which spatial variability is more significant in land use analysis. The entropy of land use mix appears to be small given a maximum entropy of 0.26 with a mean value of 0.13, suggesting that the land use types are generally homogenous within TZs. This is possibly because the TZ level of the land use mix measurement, which is aggregated from mesh blocks as the finest statistical level in Census, is not large enough to generate sufficient variation in land use types within TZs. Using a larger aggregation level may create more variation but is not applied in this analysis to ensure the geographical aggregation level of the land use variables is consistent in the dataset.

Table 3.6 Descriptive Statistics of Variables

Variable	Obs	Mean	Std. Dev.	Min	Max
PTTRIP	1824	0.40	0.32	0.01	3.00
PRICE	1824	2.19	1.02	0.25	9.00
AGE	1824	43.95	5.80	22.00	81.80
INCOME	1824	40.04	13.87	6.90	153.64
POPULATION DENSITY	1824	19.22	9.44	0.25	65.45
LAND MIX	1824	0.13	0.07	0.00	0.26
PSEUDO NODES	1824	1.60	1.72	0.08	29.87
DISTANCE TO CBD	1824	29.64	29.07	0.16	152.70
DISTANCE TO PT STOP	1824	274.77	304.42	0.36	4622.26
BUS FREQUENCY	1824	0.20	0.47	0.00	4.40
PT STOPS	1824	37.69	19.64	0.00	121.00

The correlation matrix in Table 3.7 is presented to identify potential collinearity among independent variables, with correlation coefficient higher than 0.500 highlighted in bold texts. The highest correlation occurs between population density and number of bus stops at 0.655, and this is expected since public transport supply is usually higher in more populated areas. The correlation between population density and distance to CBD is also relatively high at -0.519 because in Sydney as elsewhere, the urban development originated from the CBD as a core so population density is higher in areas closer to the CBD. Another

strong correlation is identified between bus frequency and number of bus stops at 0.578 as might be expected since the bus frequency is the sum of bus services for each bus stop around 400 meters of the TZ centroid.

Table 3.7 Correlation Matrix

	PTTRIP	PRICE	AGE	INCOME	DENSITY	LANDMIX	PSEUDO	CBD	DISTANCE	FREQ	STOPS
PTTRIP	1										
PRICE	-0.186	1									
AGE	-0.153	-0.117	1								
INCOME	0.054	0.013	-0.050	1							
DENSITY	0.384	-0.252	-0.244	0.119	1						
LANDMIX	-0.040	0.036	0.010	-0.006	-0.075	1					
PSEUDO	-0.226	0.161	-0.001	-0.081	-0.208	0.065	1				
CBD	-0.427	0.163	0.101	-0.328	-0.519	0.114	0.239	1			
DISTANCE	-0.107	0.132	0.053	0.005	-0.255	-0.063	0.225	0.135	1		
FREQ	0.255	-0.123	-0.290	0.077	0.423	-0.183	-0.140	-0.273	-0.121	1	
STOPS	0.334	-0.219	-0.187	0.001	0.655	-0.142	-0.264	-0.392	-0.348	0.578	1

The estimation results of the global model using pooled Ordinary Least Squares (OLS) estimation are displayed in Table 3.8. The adjusted R-squared is 0.252 which suggests that 25.2 percent of the variation in the dependent variable is explained by the explanatory variables. The global model does not have a good model goodness-of-fit with omitted variables identified from the Ramey's RESET test. However, the F-test of the global regression model confirms the relationship between dependent and independent variables is statistically significant and this is the focus of this exploratory analysis. The Variance Inflation Factor (VIF) indicators are minimal suggesting that there is not a significant multicollinearity problem.

Table 3.8 Global Model Estimation Results

Dependent Variable: PTTRIP	Coef.	S.E.	t	P>t	[95% C.I.]	VIF
PRICE***	-0.026	0.007	-3.84	0.000	-0.039 -0.013	1.13
AGE***	-0.005	0.001	-3.76	0.000	-0.007 -0.002	1.17
INCOME***	-0.002	0.000	-3.68	0.000	-0.003 -0.001	1.16
POPULATION DENSITY***	0.004	0.001	3.93	0.000	0.002 0.006	2.16
LAND MIX*	0.173	0.101	1.71	0.088	-0.026 0.372	1.06
PSEUDO NODES***	-0.020	0.004	-4.91	0.000	-0.027 -0.012	1.14
DISTANCE TO CBD***	-0.003	0.000	-12.45	0.000	-0.004 -0.003	1.58
DISTANCE TO PT STOP	0.036	0.023	1.55	0.121	-0.009 0.081	1.20
BUS FREQUENCY**	0.036	0.018	2.05	0.041	0.001 0.071	1.64
PT STOPS *	0.001	0.001	1.82	0.069	0.000 0.002	2.45
_CONS	0.718	0.073	9.84	0.000	0.575 0.861	n/a
Observations	1824					
Prob>F	0.00					
R-squared	0.256					
Adj R-squared	0.252					
Ramsey RESET Test (Ho: model has no omitted variables)						
F(3, 222)	42.51					
Prob > F	0.000					

* P<0.10, ** P<0.05, *** P<0.01

As this is a linear regression model, the interpretation of the estimated coefficients relates to the units of variables. The model estimation results suggest that most variables are significant at 95 percent confidence level with the expected signs. According to the estimated coefficients, public transport demand is expected to be higher in TZs with lower average trip price, lower income, and lower age. For public transport supply, increasing bus frequency is expected to increase public transport demand given the positive sign of its coefficient. In terms of land use variables, public transport demand is expected to increase with higher population density, but decrease with the increase of distance to CBD and pseudo nodes at 95 percent confidence level. This confirms the hypothesis that a built environment with fewer curvy roads and cul-de-sacs such as a grid network provides better connectivity and walkability for public transport users and thus increases public transport demand.

The land use mix entropy and number of bus stops are only significant at 90 percent confidence level but with expected signs. This is possibly a result of the fine aggregation level of land use mix measurement and the correlation between number of public transport stops and population density as well as bus frequency which have already captured the similar effects on public transport demand.

The distance to the nearest bus stop, although suggested as relevant to travel behaviour in some previous studies, is not significant in this exploratory analysis. This may be because people living closer to a bus stop or a train station are not necessarily more likely to use public transport services, and instead, the frequency of public transport services is more important. People may not take public transport in a low-frequency station or stop although they live close to the services. This finding highlights the importance of including a public transport supply measure to control for the quality of public transport service.

The average elasticity derived from Equation (3.3) is a convenient way to interpret the proportional change of public transport demand caused by the changes in the explanatory variables as shown in Table 3.9. All the elasticities are less than one in absolute value indicating public transport demand is inelastic to each of the variables individually. The highest elasticity is the age elasticity at -0.49 suggesting that a one-hundred percent increase in age is expected to decrease public transport demand by 49 percent. The second highest elasticity based on the absolute value is distance to CBD at -0.25, followed by population density, income, and price. Other land use variables have relatively smaller elasticities but the joint effect can be considerably influential. Given the two highest elasticities of age and distance to CBD, public transport demand in the SGMA is expected to have more variation with respect to these two measures. This result is used in the construction of the pseudo panel data analysis presented in Chapter 4.

Table 3.9 Average Elasticity to Public Transport Demand

Variable ¹	Elasticity
PRICE	-0.14
AGE	-0.49
INCOME	-0.18
POPULATION DENSITY	0.19
LAND MIX	0.05
PSEUDO NODES	-0.08
DISTANCE TO CBD	-0.25
BUS FREQUENCY	0.02
PT STOPS	0.09

¹Distance to PT stops is not included because it is not significant in the global model

3.4.3 Local model estimation

The local models of GWR are estimated by taking account of the spatial variability with results being compared to the global model estimation discussed above. The goodness of fit of the GWR local model can be compared to the global model using the AIC and adjusted R-square values. The local model estimation results suggest that the AIC and adjusted R-square are 247.21 and 0.40 respectively, and both are improved as compared to the global model (AIC: 470.96 and adjusted R-square: 0.252). This confirms that the local model has better model explanatory power and model goodness-of-fit by taking account of the spatial heterogeneity of the observations. The Monte Carlo test can be used to examine the significance of the spatial variability of parameters identified in the local model. The results of this, shown in Table 3.10, suggest that the spatial variability is significantly evident in the price, pseudo nodes, distance to CBD, distance to the nearest bus stop, and bus frequency variables.

The GWR local model estimates the parameters for each observation, and the results can be displayed using GIS layers to visualise the spatial variation. The spatial variation of this analysis is only apparent in the urban area close to the Sydney CBD, so only the Sydney urban area as highlighted in Figure 3.7 is discussed in this analysis.

Table 3.10 Results of the Monte Carlo Test for Spatial Variability

Variable	P-value
PRICE	0.000
AGE	0.050
INCOME	0.210
POPULATION DENSITY	0.190
LAND MIX	0.810
PSEUDO NODES	0.000
DISTANCE TO CBD	0.000
DISTANCE TO PT STOP	0.000
BUS FREQUENCY	0.000
PT STOPS	0.960

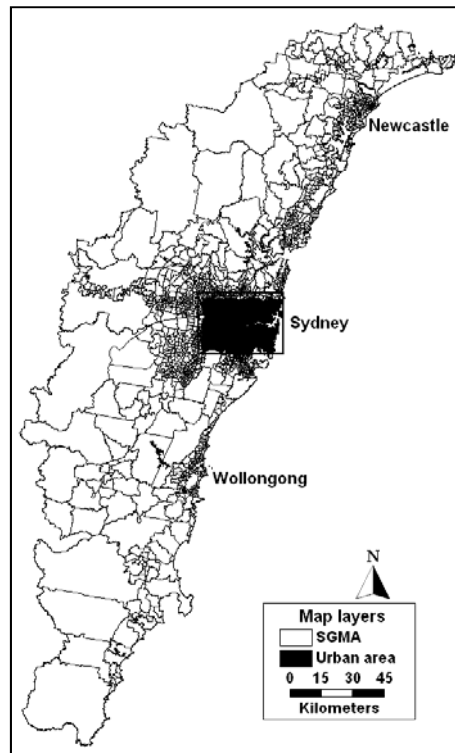


Figure 3.7 Location of the Sydney Urban Area in the Sydney Greater Metropolitan Area

(Source: developed from GIS maps)

The estimated parameters of price, pseudo nodes, bus frequency, and distance to CBD are presented in Figure 3.8 to 3.11. TZs with significant parameters are highlighted in colour, with positive signs coloured in green and negative signs coloured in red. The grey areas and white areas are TZs with insignificant

parameters at 95 percent confidence level or no public transport observations respectively.

As shown in Figure 3.8, price has a significantly negative impact on public transport demand in the North Sydney areas. The estimated parameters of price in these negative areas are larger in absolute terms than the global parameter of price at -0.026, suggesting that public transport users in these TZs are more sensitive to the price change than the average in the SGMA. It is also important to note that in the south of the CBD, there is an area which has a significantly positive relationship between public transport demand and price. This is possibly because the “Airport Link” service between the CBD and the Airports operated by a private sector charged an access fee for passengers using train stations between Sydney Airport and the CBD, even when the airport itself was not being accessed² giving higher train trip prices in this area as compared to surrounding areas whilst the public transport demand here is also higher.

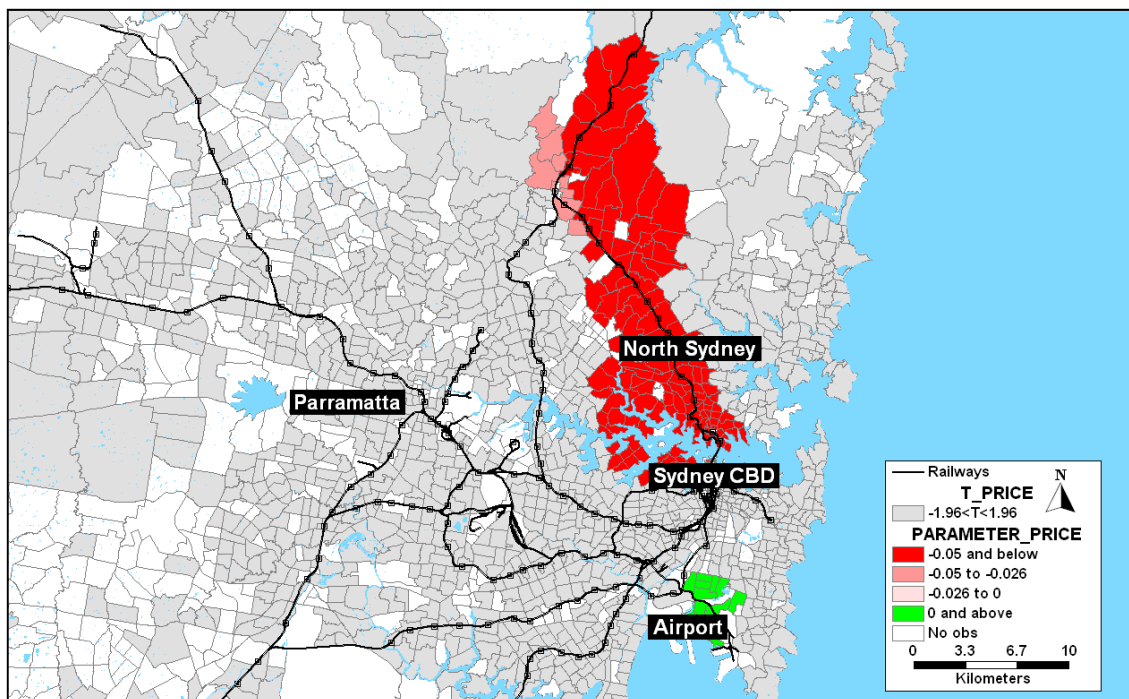


Figure 3.8 Map of the Local Model Estimates of Price in the Sydney Urban Area

² This access fee for using stations other than the domestic and international terminals was cancelled in March 2011, but the fee remains in place for accessing the terminals.

For the local estimates of pseudo nodes shown in Figure 3.9, TZs with statistically significant parameters mostly have a negative relationship between public transport demand and the number of pseudo nodes, with larger parameters in absolute terms than in the global model at -0.020. It is interesting to note that both Manly and Watsons Bay in the East have a particularly strong relationship between public transport demand and pseudo nodes. These two areas are popular attractions for tourists and leisure activities which have more public transport usage than their surrounding areas so this relationship stands out in this area. In the North Sydney areas close to the Sydney CBD, the number of pseudo nodes has a positive impact on the public transport demand. This is because although the local road network here consists of more curves and cul-de-sacs because of the topological features in these areas close to Sydney harbour, residents in this area highly rely on the public transport to access the CBD due to the congestion and tolls on Sydney Harbour Bridge connecting North Sydney and the CBD and thus results in an inverse relationship between public transport demand and the number of pseudo nodes.

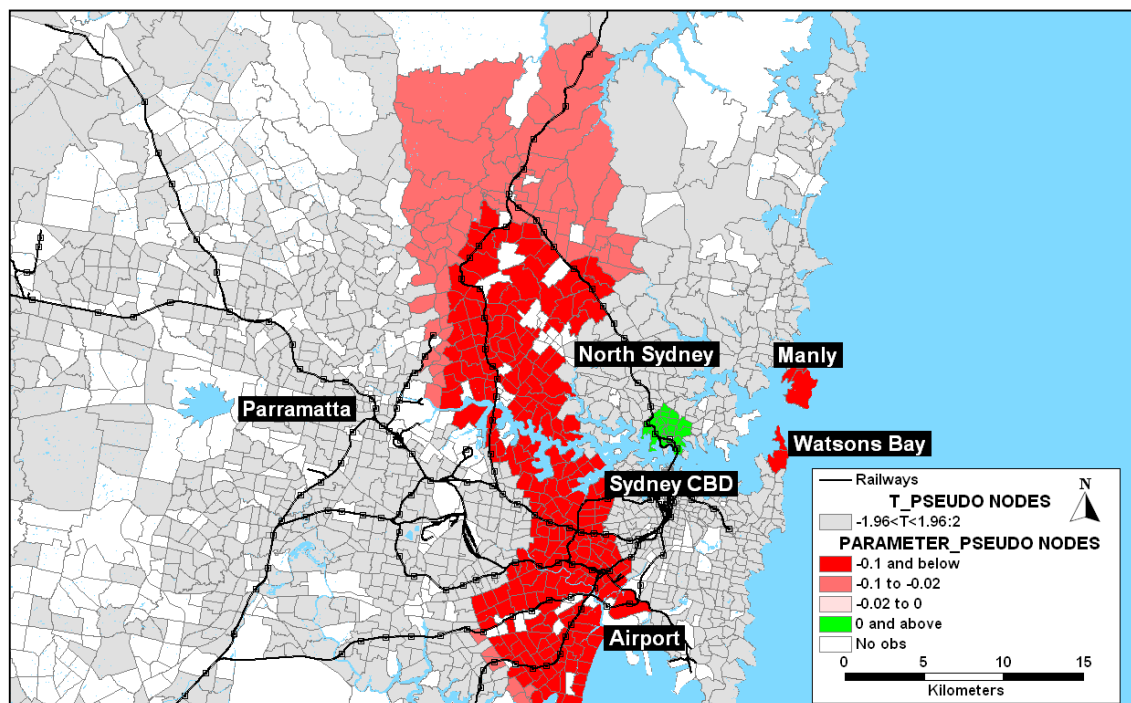


Figure 3.9 Map of the Local Model Estimates of Pseudo Nodes in the Sydney Urban Area

The local parameter of bus frequency is displayed in Figure 3.10. The inner Sydney and the west of Sydney show two distinctive patterns in the relationship between bus frequency and public transport demand. Public transport demand is higher in areas with higher bus frequency in the west of Sydney, but an inverse relationship is identified in inner Sydney. In inner Sydney this maybe because there are more short trips which can be made by walking given the higher land use density. This is supported by the way that the proportion of walk trips in inner Sydney (40 percent) is considerably larger than the total average in the SGMA (18.3 percent), and the average trip distance in inner Sydney (4.6 km) is shorter than the average in the SGMA (8.5 km) in 2010/2011 (Bureau of Transport Statistics, 2011d). As a result, the higher frequency of bus services does not guarantee higher public transport use in inner Sydney. In contrast, the relationship is positive in the west of Sydney suggesting an increase in bus frequency is expected to raise public transport demand as identified in the global model estimation results. This distinctive difference between the two regions has important policy implications. Increasing bus frequency is expected to encourage more public transport use in the outskirts of Sydney, particularly in the western areas, as compared to inner Sydney where the bus services are already frequent.

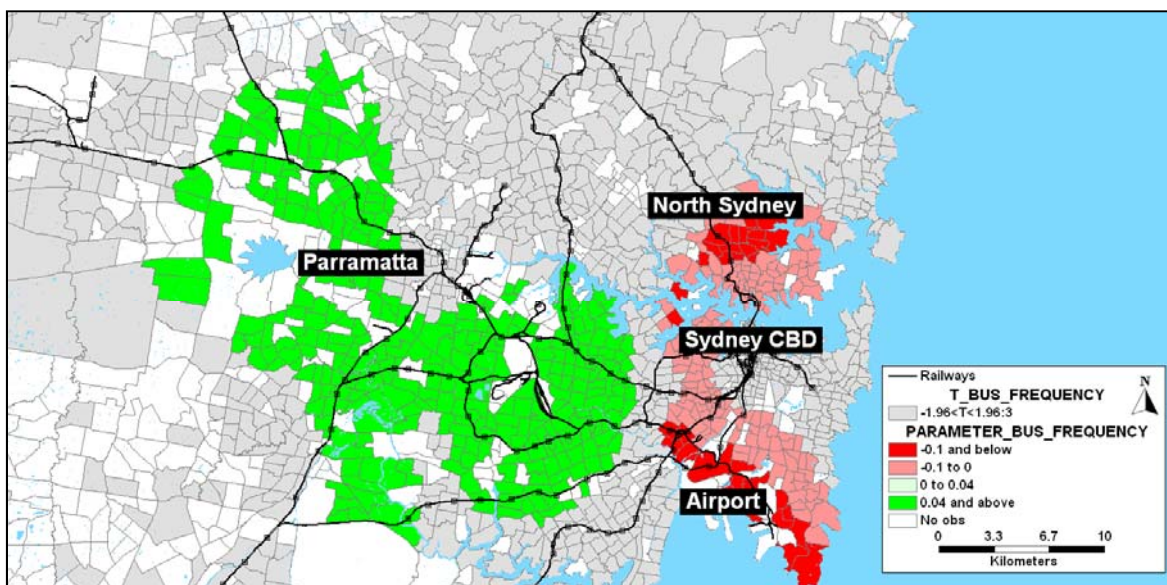


Figure 3.10 Map of the Local Model Estimates of Bus Frequency in the Sydney Urban Area

For distance to CBD, the negative relationship to public transport demand is more obvious in TZs closer to the CBD than in the outskirts of the city as shown in Figure 3.11. There is no TZ with a positive parameter and all the local parameters are larger in absolute terms than the global parameter suggesting that public transport users residing in inner Sydney are more sensitive to the travel distance to the CBD, as opposed to people living in the outer Sydney who may travel more frequently to local business centres instead of the CBD so this relationship is less significant. Distance to CBD appears to have the most consistent and strongest relationship to public transport demand, given its high elasticity in the global model estimation and this local model evidence which shows that the magnitude of the impact of distance to CBD on public transport demand gradually decreases from the city centre to the outskirts. This indicates that variation in public transport demand in the SGMA can be fairly distinguished according to the distance to CBD and households, and this finding provides a rationale for the pseudo panel data analysis presented in Chapter 4.

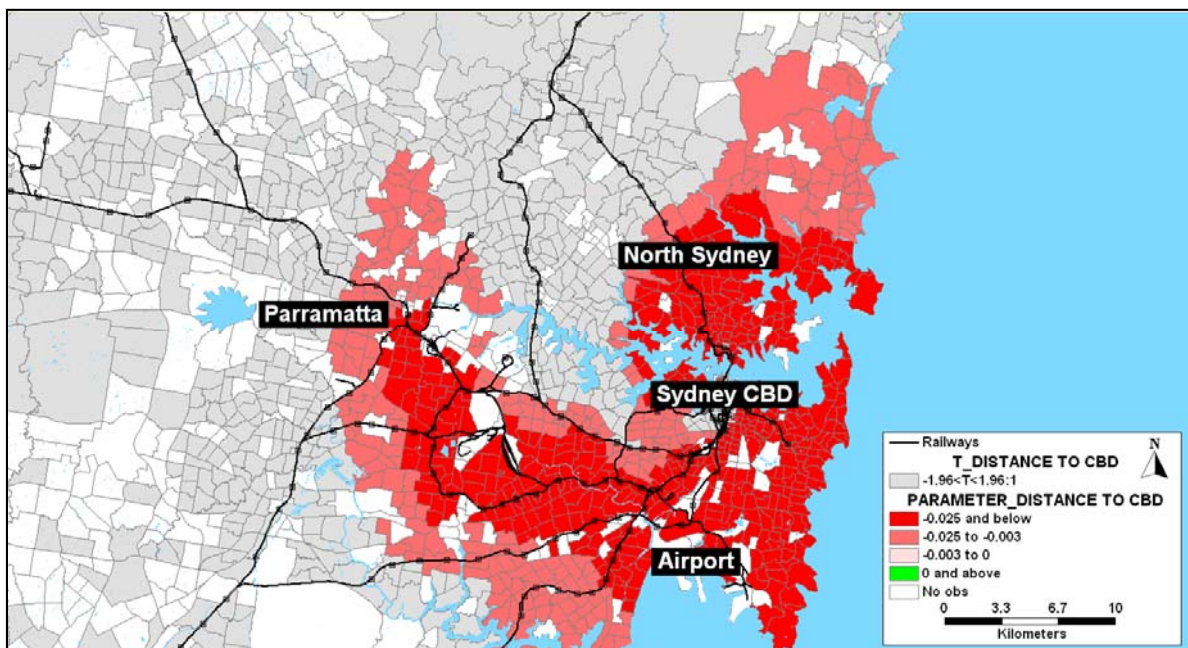


Figure 3.11 Map of the Local Model Estimates of Distance to CBD in the Sydney Urban Area

In GWR local model estimation, it is possible that the model exhibits spatial dependency and thus introduces spatial autocorrelation in the error term. Spatial dependency can be investigated by mapping the residuals of the local model on the GIS map as shown in Figure 3.12. The spatial dependency may exist if the

signs and values of residuals are similar in the neighbouring geographical zones, but this is not strongly evident in Figure 3.12 which indicates that the potential spatial autocorrelation is not substantial. More advanced models and statistical tests have been developed to address the spatial autocorrelation (Charlton and Fotheringham, 2009) but are not further discussed in this study, since the aim of the GWR model estimation is an exploratory analysis to investigate the relationship between public transport demand and selected independent variables.

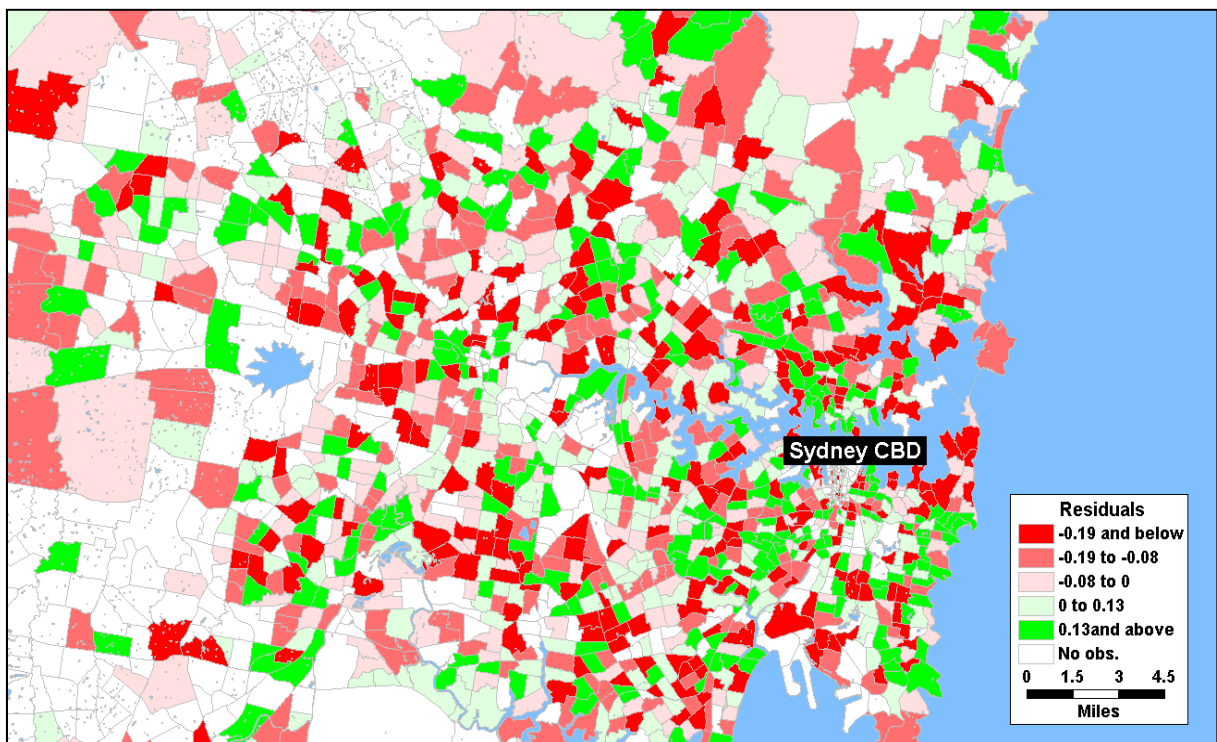


Figure 3.12 The Residuals of the Local Model Estimation in the Sydney Urban Area

3.5 Summary

This chapter starts with the introduction of the SGMA with its key geographical and demographical information, followed by a summary of travel-related statistics and public transport network in the SGMA. This introduction shows how Sydney is a city highly dependent on vehicle use rather than public transport which only take accounts of around ten percent of total trips. Although public transport is not the major means of travel, urban development and the strategic public transport planning in Sydney are aimed to provide better and easier accessibility to and from the urban area through public transport. This

highlights the importance of understanding public transport demand and its explanatory variables with respect to their relationship to the Sydney CBD which is the core of business and trip attractions in Sydney.

Section 3.3 presents and defines the variables in the dataset of this study. This study aims to investigate the variation in public transport demand with respect to various aspects of factors, including public transport price, public transport supply, socio-economic factors, and most importantly, a comprehensive set of land use variables comprising 3D of land use measures and accessibility. A public transport demand model integrated with these multiple types of variables at a micro-level has not been commonly identified in the literature of conventional public transport demand modelling as reviewed in Section 2.1 and 2.2.

The exploratory analysis conducted in Section 3.4 confirms the relationship between public transport demand and the explanatory variables as hypothesised, with expected signs identified in the global model estimation which can be interpreted as an average relationship for the SGMA as a whole. The GWR local model gives more insight into the relationship at a disaggregate level, showing how the variation in public transport demand is related to the geographical location of the observations. This exploratory analysis not only identifies that there is a relationship between public transport demand and the selected variables, but also provides evidence for a pseudo panel data analysis presented next by giving an understanding of the variation of public transport demand across geographical locations.

However, this exploratory analysis is not able to take account the temporal effect of public transport demand. Using pooled data from 1997 to 2009, this analysis can only be considered as a cross-sectional analysis across all TZs and does not capture the changes in public transport demand and its explanatory variables over time. Moreover, the global model does not suggest good model fit with omitted variable bias being identified. This may not be a serious issue if only the relationship between the dependent variable and independent variables is of interest, but the validity of the model will be questionable if the model is used for

demand forecasting. Therefore, the following chapters use a pseudo panel data approach as reviewed in Section 2.3 to construct a public transport demand model which is capable of taking account of the dynamics of travel behaviour, with a potentially more rigorous modelling approach and better model goodness-of-fit for demand forecasting purpose.

CHAPTER 4 PSEUDO PANEL DATA APPROACH

4.1 Introduction

The preliminary analysis presented in Section 3.4 is conducted to investigate the relationship between public transport demand and its explanatory variables using pooled data in the Sydney Household Travel Survey (SHTS) from 1997 to 2009. This analysis identifies this relationship across geographical locations without taking account of the temporal effect of travel demand. The investigation of the temporal effect of public transport demand is one of the key research questions of this study as discussed in Section 1.1 as the presence of the temporal effect will lead to a difference between short-run and long-run demand.

A longitudinal analysis is required to investigate the temporal effect of public transport demand change. As discussed in Section 2.3, when genuine panel data are not available, a pseudo panel data approach with sound theory developed in the literature can be used to address this research question and thus is employed in this study. This chapter explicitly introduces the process of pseudo panel data construction for this study in Section 4.2 and Section 4.3, followed by an exploratory analysis on the created cohort data in Section 4.4. The general form of the pseudo panel data models and a discussion on the estimation techniques are presented in Section 4.5.

4.2 Grouping criteria for pseudo panel data

Pseudo panel data, introduced by Deaton (1985), are created from repeated cross-sectional data by classifying individuals into analyst-defined cohorts based on time-invariant criteria such as birth year. A collection of cohorts which share some common characteristic across time are defined as a “group” in this study. The pseudo panel dataset of this study is constructed from individual records repeatedly collected in the Sydney Household Travel Survey (SHTS). The SHTS database contains all modes of trips made by respondents recruited in the survey. As the focus of this study is public transport, only public transport trips, constituted of train and bus trips as the two major public transport systems in the Sydney Greater Metropolitan Area (SGMA), are selected for this study.

Respondents under 18 years old are excluded because their travel mode choices are considered to be constrained by their ineligibility of driving.

Public transport demand in the pseudo panel data analysis is defined as same as the Geographically Weighted Regression (GWR) analysis in Chapter 3, which is the average number of public transport trips made by a traveller (where a traveller refers to a respondent making at least one trip using any trip mode) in a day. The historical total number of public transport tips and travellers recorded in the SHTS are summarised in Table 4.1.

Table 4.1 Historical Statistics of Public Transport Trips from the Sydney Household Travel Survey data

	No. of Public Transport Trips	No. of Travellers	Average Public Transport Trip per Traveller
1997	2,100	6,053	0.35
1998	2,020	5,697	0.35
1999	1,763	5,044	0.35
2000	1,776	5,323	0.33
2001	1,708	4,979	0.34
2002	1,813	5,173	0.35
2003	1,540	4,782	0.32
2004	1,441	4,883	0.30
2005	1,613	4,885	0.33
2006	1,476	5,208	0.28
2007	1,577	5,097	0.31
2008	1,697	5,275	0.32
2009	1,798	5,304	0.34
AVG	1,717	5,208	0.33

From the aggregate data, the annual public transport demand does not change substantially from 1997 to 2009 with an average of 0.33 public transport trips made by a traveller per day. Thus, at an aggregate level, there is little evidence of public transport demand changing over time which limits the identification of long-run travel demand change if an aggregate approach is in use, and thus the differentiation of short-run and long-run demand remains under-researched in the literature. Therefore, a pseudo panel dataset is constructed to investigate the

long-run travel demand change whilst incorporating more individual information collected from household travel survey data.

The principles of constructing a pseudo panel dataset have been reviewed in Section 2.3.2. A pseudo panel dataset is constituted of analyst-formed cohorts using grouping criteria that aim to increase the inter-group heterogeneity of the cohorts. The previously used grouping criteria in the existing pseudo panel data literature are summarised in Table 4.2 and are evaluated next for this study.

Table 4.2 An Evaluation of Grouping Criteria

Grouping Criteria	Pros	Cons	Scale
Birth Year	<ul style="list-style-type: none"> • Time-invariant • Related to travel behaviour • Identification of life-cycle and generation effects • Has been commonly used 	<ul style="list-style-type: none"> • Correlated to age which is one of the explanatory variables 	<ul style="list-style-type: none"> • 5-year band • 10-year band • Variable band
Gender	<ul style="list-style-type: none"> • Time-invariant • Independent of exogenous variables 	<ul style="list-style-type: none"> • Not substantially related to public transport use in Sydney 	<ul style="list-style-type: none"> • Male • Female
Household Location	<ul style="list-style-type: none"> • Highly related to travel behaviour • Allows geographical analysis 	<ul style="list-style-type: none"> • Correlated to land use characteristics • Time-varying 	<ul style="list-style-type: none"> • SD (3 regions) • SSD (18 regions) • Other aggregation levels
Household Structure	<ul style="list-style-type: none"> • Related to travel behaviour 	<ul style="list-style-type: none"> • Time-varying • Correlated to car ownership and household income 	<ul style="list-style-type: none"> • Single • Couple • Couple and children
Education Level	<ul style="list-style-type: none"> • Independent of exogenous variables 	<ul style="list-style-type: none"> • Not related to travel behaviour • Time-varying 	<ul style="list-style-type: none"> • High school • Vocational college • University • Postgraduate

From the list in Table 4.2, birth year is the most commonly used variable to create cohorts because it is time-invariant and it also allows the identification of the “life-cycle effect” and “generation effect” of travel behaviour as demonstrated in Dargay and Vythoulkas (1999). In the context of Sydney, age appears to have a strong relationship to public transport demand as demonstrated in the exploratory analysis in Chapter 3 which shows that the age elasticity of public transport demand is the highest among all the explanatory variables at -0.49 (see Table 3.9). In Figure 4.1 which presents the relationship between public transport mode share and age in Sydney, people age between 11 and 30 years old have higher train and bus mode share as compared to other generations. The

mode shares of train and bus both drop when people reach their middle age. The bus mode share then increases again for age groups over 60 years old, whereas train mode share remains as low as the middle-age groups. This clearly shows that the public transport use in the SGMA is related to travellers' life-cycle and age generation, and hence birth year is selected as one of the grouping criteria to form the cohorts for this study in order to increase the inter-group heterogeneity of the created pseudo panel data in terms of public transport demand.

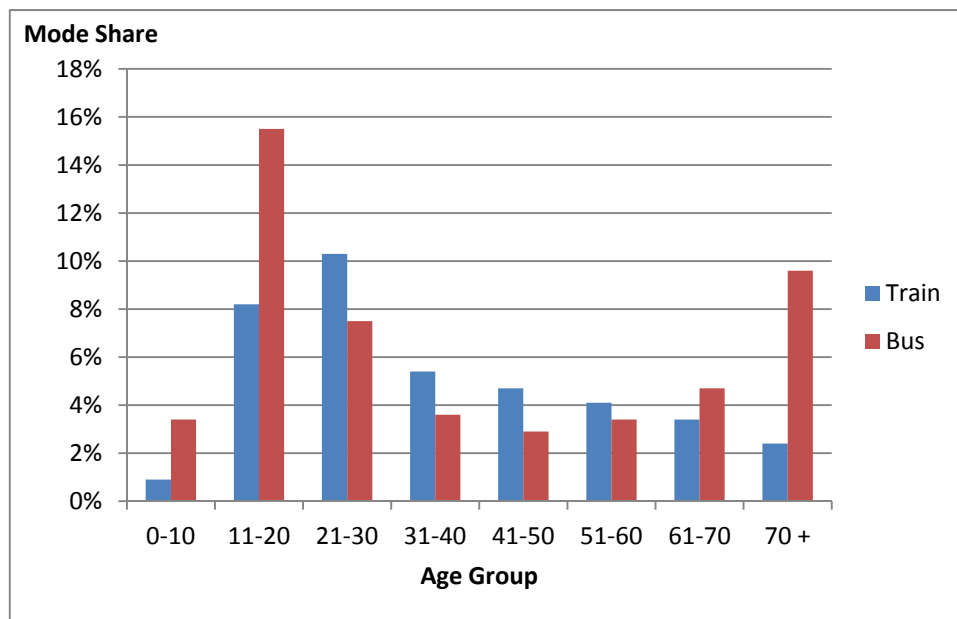


Figure 4.1 Train and Bus Mode Share by Age Group in 2009/10
 Source: Bureau of Transport Statistics (2011b)

Gender, although time-invariant, is not strongly related to public transport demand in the context of Sydney. According to the 2009/10 household travel summary report by Bureau of Transport Statistics (2011b), the mode share of train is 5.6 percent for males and 4.9 percent for females, and bus mode share is 5.6 percent for males and 5.9 percent for females. The total public transport mode share, combining train and bus, is also not very different between male and female travellers, so gender is not considered as an appropriate grouping criterion for constructing heterogeneous cohorts for this study.

The household location can be defined in various ways and at different geographical aggregation levels. Although household locations may be variable over time, the advantage of using the geographical information of households is

that it is expected to affect travel behaviour and it also allows for geographical analysis. In Sydney, the major public transport network is designed for accessing the Sydney Central Business District (CBD) and its surrounding areas. People living closer to the CBD have better accessibility to public transport than people in suburban areas and those living close to the CBD also have higher public transport demand. The demand elasticity with respect to the distance to CBD is the second highest at -0.25 among all the variables examined in Chapter 3 (see Table 3.9). In addition, as shown in Figure 4.2, the average number of public transport trips made by a traveller per day from 1997 to 2010 is higher in the areas closer to the Sydney CBD, and gradually declines with the distance to CBD. Therefore, using the household distance to CBD to group public transport users is expected to generate more homogenous groups and greater inter-group heterogeneity across the created groups in the pseudo panel dataset, and hence this variable is used to create the pseudo panel data although it is not time-invariant.

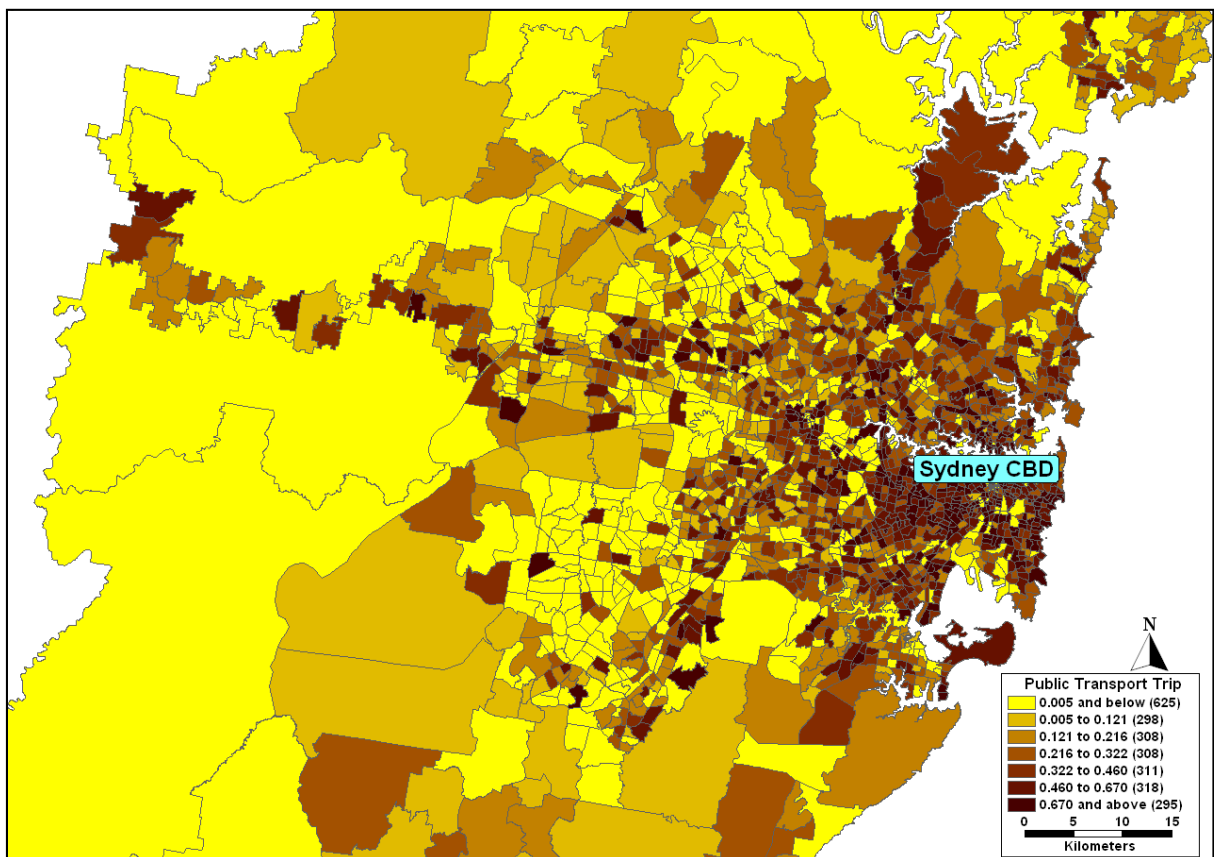


Figure 4.2 Average Public Transport Trips by Travel Zones
 Source: Summarised from the Sydney Household Travel Survey database

Household structure and education level are considered as inappropriate for this research because these variables are time-varying and not highly related to public transport demand in Sydney. Therefore, to ensure the individuals within cohorts display a relatively homogenous pattern of public transport demand, birth year and household distance to CBD are chosen for their capability of creating distinctive groups.

4.3 Pseudo panel data construction

4.3.1 Forming the cohorts

The next step of pseudo panel data construction is to assign individual records from the SHTS to distinctive groups according to the two grouping criteria (i.e., birth year and household distance to CBD), and to create groups consisting of similar cohorts created for each year of survey. Only public transport (train and bus) trips from the SHTS are selected for the pseudo panel data construction. Trips made by other modes are excluded because they do not contain the information about the public transport price, which is the key variable of the public transport demand model. Moreover, in the SGMA, around 90 percent of trips are not made by public transport, and hence including all modes of trips is likely to confound the demand elasticity estimated from the demand models. Thus, this study focuses on public transport users only and the research outcomes such as demand elasticities in the following chapters are applied only to the current public transport users.

To create groups by birth years, most pseudo panel studies have applied a fixed-range band to each birth year group (for example: five years or ten years). The shortcoming of this approach is that it generates large variation in cohort sizes across all the created cohorts. For example, the older groups tend to contain fewer individuals than middle-age groups. As a result, this grouping method generates more cohorts with insufficient individuals in a cohort. This may not be a serious issue when the overall sample size is sufficiently large as in most car travel studies. In this case, the average number of public transport trips from the SHTS is only around 1,717 trips annually (as shown in Table 4.1) which may not

be sufficient for cohort construction if a fixed range of birth year is adopted. Therefore, variable ranges are applied to the grouping process in a way to equalise the number of individuals assigned to each birth year group. This approach to define the scales of groups has not been evident in previous pseudo panel data research in the literature.

The other grouping criterion is the household distance to CBD. The distance to CBD is determined by measuring the distance between the centroid of the Local Government Area (LGA) where a household is located and the centroid of the LGA containing Sydney CBD. This grouping process also aims to equalise the number of individuals included in each of the distance-to-CBD groups created by equally allocating all the individual records to the defined distance-to-CBD groups.

The other issue in constructing the pseudo panel dataset is the number of cohorts to be created. As discussed in Section 2.3.2, given the number of total individual records is fixed, having a larger number of cohorts gives more observations which results in the better estimation efficiency, whereas increasing the cohort size (implying decreasing the number of cohorts) reduces measurement errors from the population means but with a lower statistical power.

From the SHTS, two pseudo panel datasets with different numbers of cohorts are first created as shown in Table 4.3. Other combinations were considered but did not show as much heterogeneity across the created groups than these two combinations. The first pseudo panel dataset consists of four birth year groups and three distance-to-CBD groups across 13 years ($T = 13$) yielding 156 cohorts in total. The second dataset reduces the cohort sizes and increases the number of groups (G) which generate 256 cohorts, after excluding four cohorts with an average age below 18 years old. The average cohort size of the second dataset has fallen from 143 trips to 86 trips, with 74 percent of the cohorts having less than one hundred cohort members.

Table 4.3 A Comparison between Two Different Pseudo Panel Datasets

	4 Birth Year Group 3 Distance-to-CBD Group	5 Birth Year Group 4 Distance-to-CBD Group
Number of Cohorts	156 (G=12, T=13)	256 (G=20, T=13) (4 cohorts have an average age below 18 years old)
Cohort Size	Avg: 143 trips Cohort size<100: 16% Cohort size<80: 5%	Avg: 86 trips Cohort size<100: 74% Cohort size<80: 49%

A drawback of the first dataset is the small number of cohorts, which potentially leads to inefficient model estimation given the small number of panel units (G) and the short time period (T). In contrast, the second dataset is likely to induce more measurement errors as a result of small cohort size which may generate larger estimation bias. With regard to this trade-off, a Monte Carlo experiment is conducted to examine the estimation efficiency and bias from the two types of datasets by simulating data with the similar properties. The experiment is presented in Chapter 5, and one of the key findings suggests that increasing the number of groups (G) substantially improves the estimation efficiency at a lower cost of bias. The overall Root Mean Square Error (RMSE) is also reduced and thus the second dataset is suggested as the preferred dataset for this study.

The final pseudo panel dataset is constituted of 20 groups with corresponding birth years and distances to CBD as shown in Table 4.4.

A table summarising the number of individual records in each cohort is presented in Table A2.1 in Appendix 2. Each group has thirteen cohorts from 1997 to 2009 except Group 5, Group 10, Group 15, and Group 20 with only 12 cohorts. These are young groups which have an average age of less than eighteen years in 1997 and thus are excluded from the dataset. As a result, there are 256 cohorts in total with an average cohort size of 86 members. The average cohort sizes still vary across the created groups although the grouping process aimed to equalise the numbers. This is because the distribution of respondents across birth year groups and distance-to-CBD groups can not be simultaneously controlled in the grouping process.

Table 4.4 Results of Pseudo Panel Construction

Group	Birth Year	Distance to CBD ¹	Average Cohort Size
1	1907-1945	Zone 1	100
2	1946-1959	Zone 1	92
3	1960-1971	Zone 1	124
4	1972-1979	Zone 1	129
5	1980-1991	Zone 1	80
6	1907-1945	Zone 2	78
7	1946-1959	Zone 2	68
8	1960-1971	Zone 2	79
9	1972-1979	Zone 2	71
10	1980-1991	Zone 2	65
11	1907-1945	Zone 3	99
12	1946-1959	Zone 3	83
13	1960-1971	Zone 3	88
14	1972-1979	Zone 3	77
15	1980-1991	Zone 3	86
16	1907-1945	Zone 4	80
17	1946-1959	Zone 4	82
18	1960-1971	Zone 4	90
19	1972-1979	Zone 4	71
20	1980-1991	Zone 4	77

¹Household distance to CBD- Zone 1: within 7.26km; Zone 2: 7.26-12.81km; Zone 3: 12.81-28.07km; Zone 4: over 28.07km.

As public transport demand in this study is defined by the number of trips made by a traveller per day, around 16 percent of SHTS respondents report no trips made on the reporting day and are excluded. The distribution of these non-travellers across cohorts was investigated as shown in Table A2.2 (Appendix 2) and found to be similar to the distribution of travellers, so these non-travellers are not expected to distort the representativeness of the selected data.

4.3.2 Variables in the pseudo panel dataset

The variables in the pseudo panel dataset are based on the dataset used for the Geographically Weighted Regression (GWR) analysis introduced in Section 3.3, which comprises public transport demand as the dependent variable and public

transport price, travellers' socio-economic factors, public transport supply, and land use characteristics at household locations as explanatory variables. The only difference is that household distance to Sydney CBD is excluded from the pseudo panel dataset to avoid endogeneity problems which would arise because it has been used as a grouping criterion to create groups. On the other hand, age is retained in the pseudo panel dataset because it is a time-varying measurement which is different from birth year as a time-invariant grouping criterion.

In a panel data analysis, it is essential to have historical data available to investigate the effect of changes in explanatory variables over time on the dependent variable. Variables collected from the SHTS have been recorded consistently since 1997, so trip price and socio-economics variables including age and income are available with annual records between 1997 and 2009. Trip price and personal annual income are deflated by using the Australian Consumer Price Index (CPI=100 in 1997) to compute the real values of both variables.

Data collected from Australian Census are not available continuously for every year because the Census is conducted only every five years. For this study, recent Census years of 1996, 2001, and 2006 are used to identify population density. To investigate how population density changes over time between 1996 and 2006 at a cohort level, a simple moving average method is employed to smooth data between 1996 and 2006. The moving-averaged data in 2006 and the actual data collected in 2006 are compared in Figure 4.3. The result suggests that there is a strong linear relationship between these two measures, and this is because the population density in Sydney has not changed substantially over the 13 years of data in the SHTS database after it is aggregated to the cohort level. Thus, this study uses the 2006 data for population density since there is no evidence to show that the moving average method is superior and there is no population density data available between Census years.

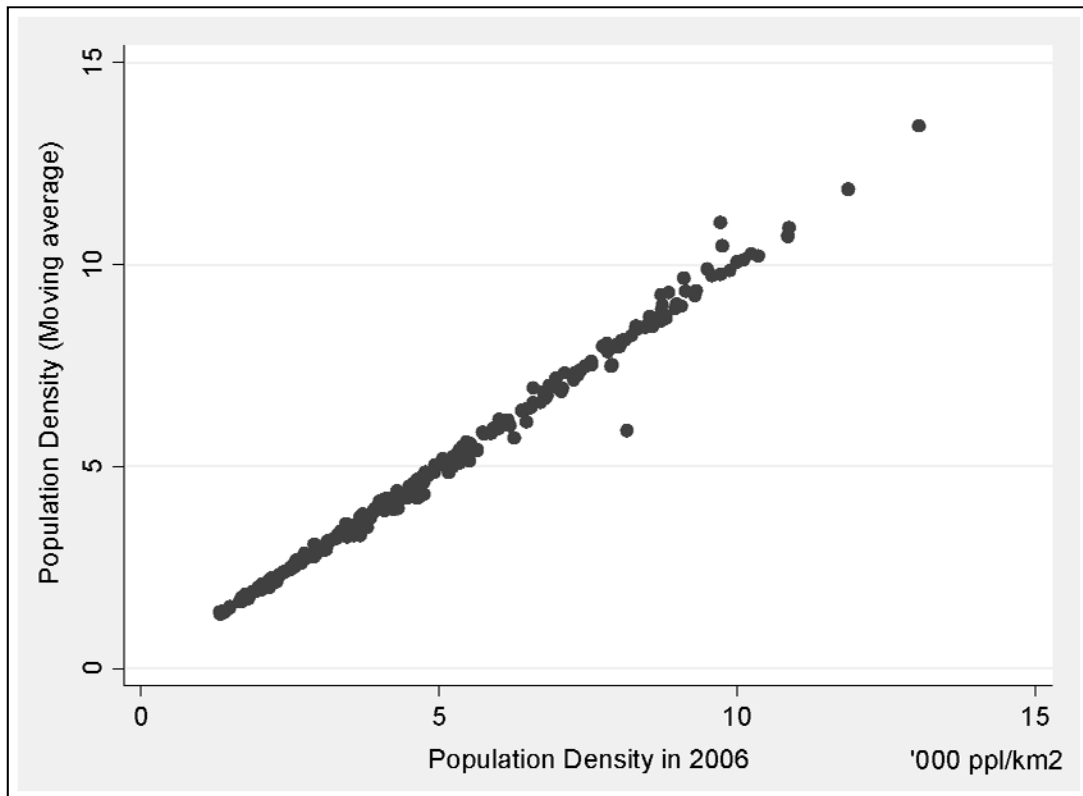


Figure 4.3 Moving-averaged Population Density versus Population Density in 2006

Other variables are collected at a single point of time because of the unavailability of historical data, such as land use mix (which is only available from 2011 Census), frequencies of train and bus in 2006, and number of road links, pseudo nodes, public transport stops, as well as the walk distance to the nearest public transport stop derived from the 2010 road network GIS layer. These variables are assumed to be time-invariant between 1997 and 2009 because the land use type, public train supply, and road network have not substantially changed in the past ten years. Nevertheless, these variables are still essential in the dataset because their cross-sectional variation are expected to have an impact on public transport demand across geographical space, and the capture of the cross-sectional variations can also be used as a reference for long-term planning. A summary of the variables in the pseudo panel dataset with their timeframes is presented in Table 4.5.

Table 4.5 A Summary of Variables in the Pseudo Panel Dataset

Variable	Description	Unit	Timeframe	Source
<i>Dependent variable</i>				
PTRIP	No. of public transport trips per person	Trips/person	Annual data 1997-2009	SHTS
<i>Price variable</i>				
PRICE	Public transport trip price	Dollars (AUD)	Annual data 1997-2009	SHTS
<i>Socio-economic factors</i>				
INCOME	Annual personal income	Thousand dollars (AUD)	Annual data 1997-2009	SHTS
AGE	Age	Years	Annual data 1997-2009	SHTS
<i>Public Transport Supply</i>				
BUS FREQUENCY	Number of buses serving a bus stop between 6am and 10am on Tuesday within 400 meters of a TZ centroid	Thousands	2006	BTS
<i>Land use density</i>				
POPULATION DENSITY	Number of population within 800 meters of a TZ centroid	Thousands	2006	Census
<i>Land use diversity</i>				
LANDMIX	Entropy of land use mix	n/a	2011	Census
<i>Land use design</i>				
PSEUDO NODES	Number of pseudo nodes within 800 meters of a travel zone centroid	Thousands	2010	Road network
<i>Accessibility</i>				
DISTACNE TO CBD	Distance between CBD and travel zone centroids	meter	2010	Road network
DISTANCE TO PT STOP	Distance between households and the nearest train station or bus stop	meter	2010	Road network
PT STOPS	Number of train stations and bus stops within 800 meters of a household	n/a	2010	Road network

4.4 Preliminary analysis

4.4.1 Group-specific effects

After constructing the pseudo panel dataset, it is important to investigate whether there is sufficient inter-group variation with group-specific patterns of travel behaviour in the pseudo panel dataset. The effectiveness of the grouping criteria to create cohorts can be examined by comparing the between-group standard deviations and within-group standard deviations of the variables. The between-group standard deviation represents the differences between the group means and the overall mean, whereas the within-group standard deviation is derived from the differences between each value and the mean of its group. From Table 4.6, which summarises the between-group and within-group variances of

all variables in the dataset, shows that the between-group standard deviations are larger than the within-group standard deviations for all variables except for land use mix and distance to the nearest bus stop which show a similar magnitude of between-group and within-group variation. The between-group variation of land use mix is small, possibly as a result of the low aggregation level of this measure (TZ level), and the distance to bus stop also shows little variation because the NSW planning guidelines specify that 90 percent of households in the metropolitan bus contract regions should be within 400 m of a rail line and/or bus route during the day, to ensure a minimum accessibility to local public transport (NSW Ministry of Transport, 2006). Nevertheless, the key variables such as public transport demand and trip price as well as other socio-economic and land use variables have demonstrated sufficient inter-group heterogeneity, which confirms that the grouping method has created sufficient between-group variation not only in the public transport demand but also in most of the explanatory variables.

For the time-invariant variables such as number of pseudo nodes and bus frequency, there are still some time-varying variations in the pseudo panel dataset although they are time-invariant variables. This is because the cohorts are constituted of different members, even when they are within the same group. Hence, the within-group variance of the time-invariant variables comes from the composition of cohort members rather than their actual changes over time. Despite the unavailability of historical data, the analysis on these time-invariant variables also provides information about their relationships with public transport demand at a cross-sectional basis, and these relationships can be used as guidance for strategic policy and planning such as transforming the land use characteristics or increasing bus frequency to encourage public transport use in the long-term plan.

Table 4.6 Between-group and Within-group Variances of all Variables

Variable	Unit		Mean	S.D.
PTTRIP	Trips	overall	0.45	0.28
		between		0.26
		within		0.11
PRICE	AU\$	overall	1.73	0.59
		between		0.56
		within		0.21
INCOME	AU\$(000')	overall	28.64	12.98
		between		11.64
		within		6.37
AGE	Years	overall	41.32	17.64
		between		17.80
		within		3.33
BUS FREQUENCY	Buses (000')	overall	0.19	0.15
		between		0.14
		within		0.05
POPULATION DENSITY	Populations(000')	overall	22.08	5.59
		between		5.50
		within		1.54
LAND MIX	Entropy	overall	0.13	0.01
		between		0.01
		within		0.01
PSEUDO NODES	Nodes (000')	overall	1.36	0.62
		between		0.59
		within		0.23
DISTANCE TO PT STOP	Kilometre	overall	0.24	0.08
		between		0.05
		within		0.06
PT STOPS	Stops	overall	41.45	7.58
		between		6.91
		within		3.44

Dargay and Vythoulkas (1999) demonstrated that the group-specific patterns of travel behaviour can be identified from the generation effect and the life-cycle effect. The former refers to the variation between birth year groups, and the later effect is investigated by considering the cohort variation over time for a specific group.

The life-cycle effect in this example is shown in Figure 4.4. Each line in Figure 4.4 represents a birth year group, with the corresponding average numbers of public transport trips and average ages of cohorts from 1997 to 2009. For example, the group “1972-1979” has an average age of 21 years in 1997 and 34 years in

2009 with an average 0.69 public transport trips in 1997 and 0.44 public transport trips in 2009. There is an age gap between birth group “1907-1945” and “1946-1959”. The reason is that there are many more public transport users over 65 years old in the group “1907-1945” than people between 55 and 65 years old and thus the average age is largely weighted by the older people in those cohorts. Figure 4.4 shows that the average number of public transport trips decreases over time for younger groups, whereas it is more stable for middle-age groups. This pattern confirms that the life-cycle effect of people’s travel behaviour in terms of their public transport demand is evident from the constructed pseudo panel dataset.

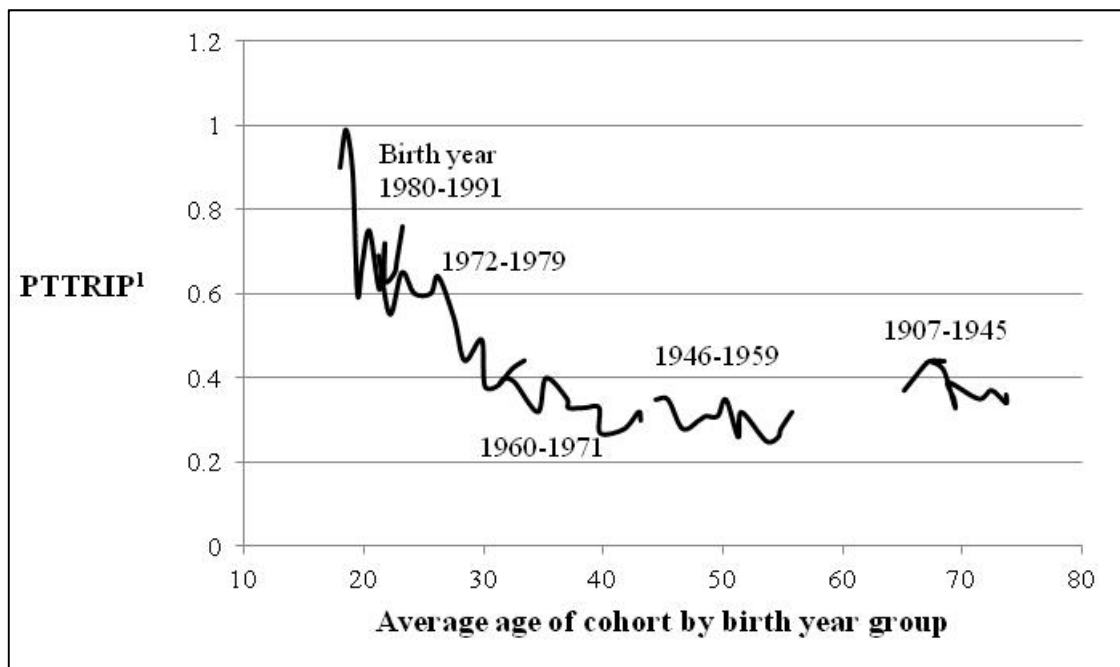


Figure 4.4 Number of Public Transport Trips by Age for Different Birth Year Groups
¹PTTRIP: Number of public transport trips per person per day

Figure 4.5 and Figure 4.6 show the box plot of average number of public transport trips by birth year groups and by distance-to-CBD groups respectively. In Figure 4.5, the average number of public transport trips is larger in the oldest group “1907-1945” and the younger groups “1972-1979” and “1980-1991”, which suggests that the level of public transport usage for people in middle-age is relatively lower. This generation effect is considered to be a result of the travellers’ socio-economic factors and the concession public transport fares for

students and pensioners in Sydney. In Figure 4.6, the relationship between the average number of public transport trips and household location is also evident. People living closer to CBD have a higher level of public transport use than people in the suburban areas. This is thought to be related to the public transport price and public transport supply and urban development characteristics.

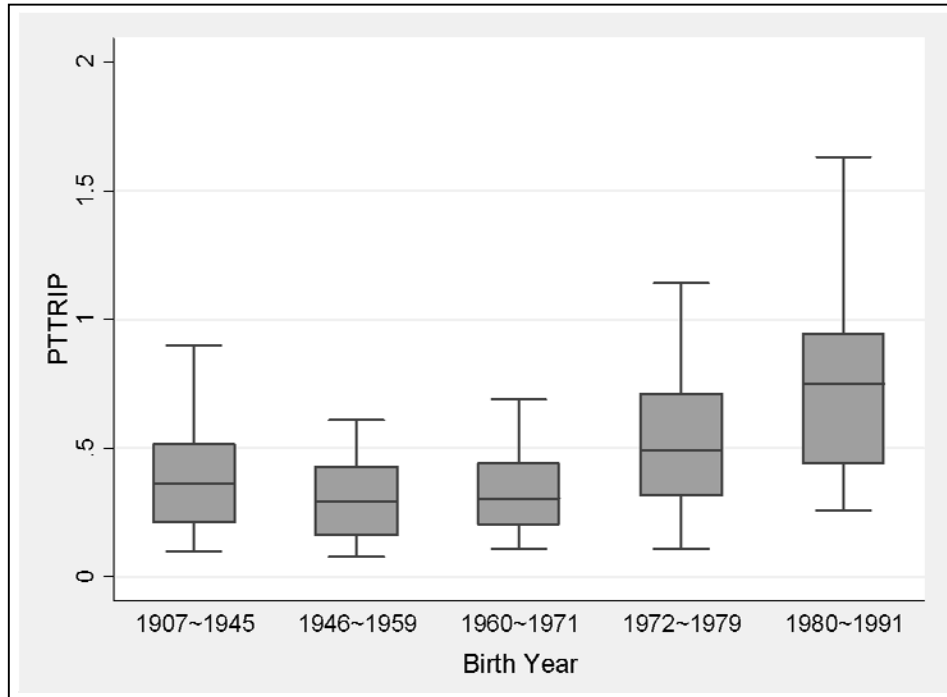


Figure 4.5 Box Plot of Number of Public Transport Trips by Birth Year Groups

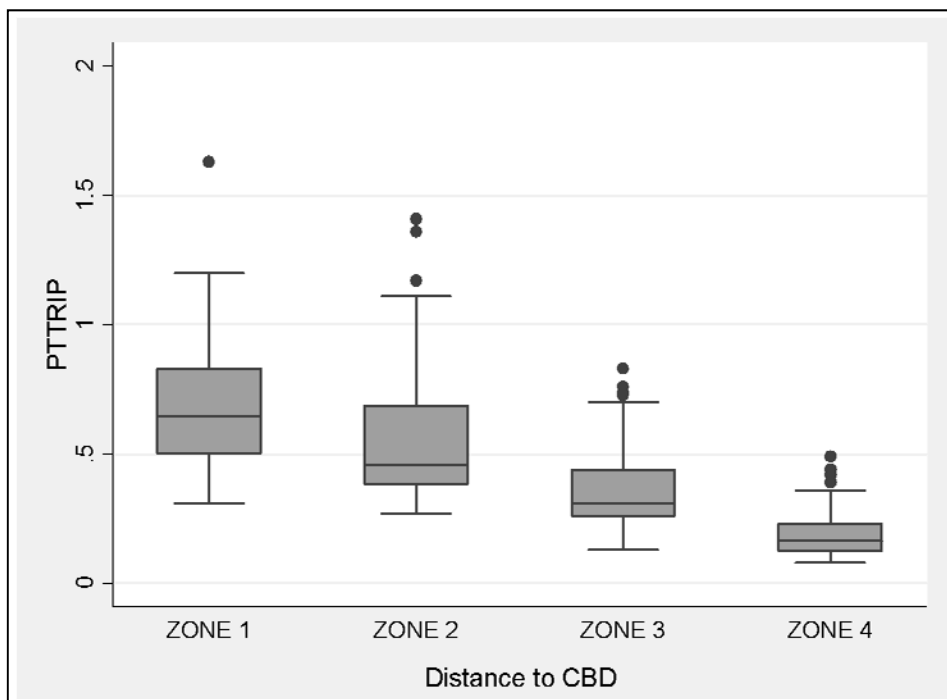


Figure 4.6 Box Plot of Number of Public Transport Trips of Distance-to-CBD Groups

This exploratory analysis confirms that the grouping method employed to construct the pseudo panel creates distinctive groups which demonstrate more individual characteristics than the aggregate data as shown in Table 4.1. This corresponds to the purpose of a pseudo-panel approach: providing an alternative way to conduct a panel data analysis that allows a certain degree of micro-economic information to be identified.

4.4.2 Historical trends of variables by groups

A panel data analysis is used to investigate the time-series changes (i.e., within-group variation) and cross-sectional changes (i.e., between-group variation). The between-group variation of the pseudo panel dataset has been presented in Section 4.4.1, and the within-group variation of time-varying variables in the dataset is introduced in this section.

Figure 4.7 shows the historical trend of number of public transport trips per person per day (i.e., *PTTRIP*) from the 256 cohorts out of the 20 groups created for the pseudo panel dataset of this study. Each plot represents a defined group of 12 to 13 cohorts over 13 waves of surveys from 1997 to 2009. The definition of group numbers is summarised in Table 4.4. The scatter plots in Figure 4.7 show that the younger groups born between 1972 and 1979 (Group 4, 9, 14, 19) and between 1980 and 1991 (Group 5, 10, 15, 20) have more substantial changes over time with a decreasing trend, whereas older groups and middle-age groups appear to be more stable over time. This is because of the life-cycle effect shown in Section 4.4.1 where younger generations tend to reduce their public transport use as they become older.

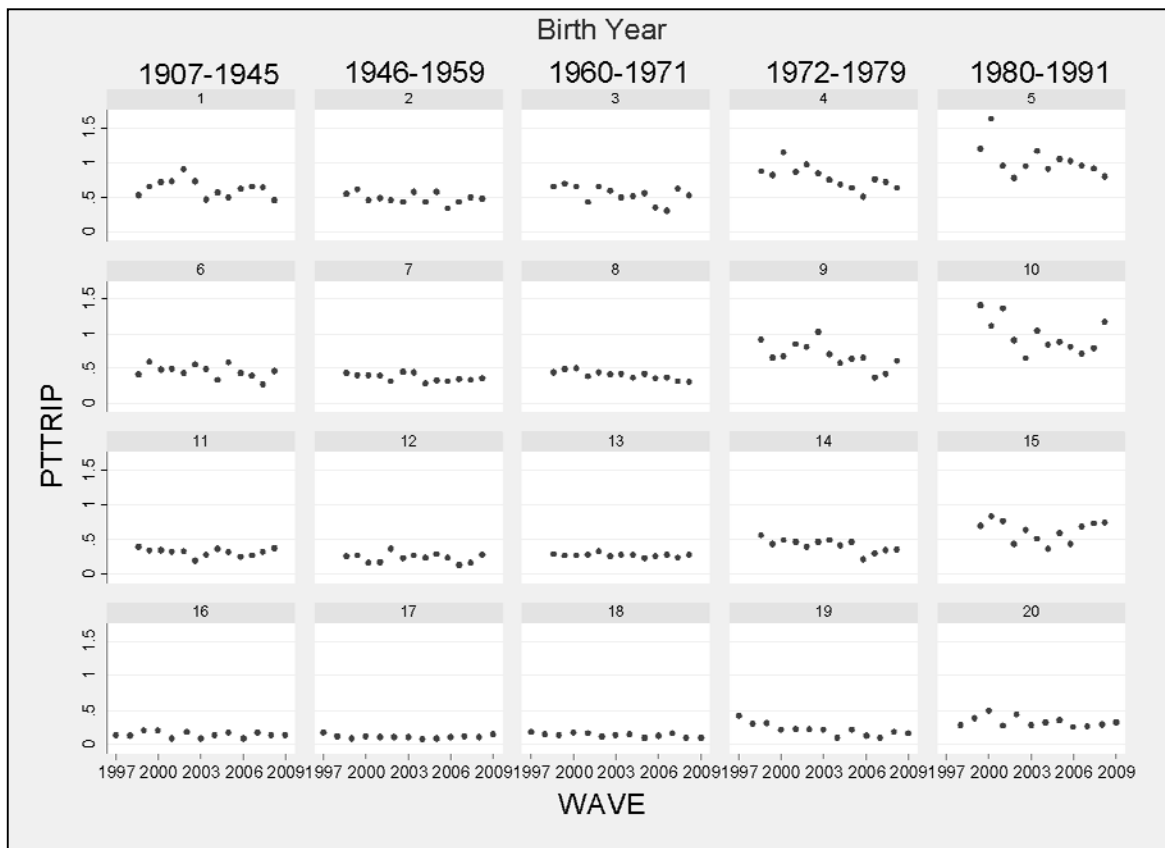


Figure 4.7 Time Trends of Number of Public Transport Trips from 1997-2009 by Group

The historical changes in public transport trip price are displayed in Figure 4.8 for the groups. Apart from the oldest birth year group 1907-1945 (Group 1, 6, 11, 16), and Group 7, 12, and 17 in birth year groups 1946-1959, public transport trip price in real terms has been increasing since 1997, with the most substantial increase being in the youngest group 1980-1991 where some public transport users are eligible for student concession prices on tickets. The trip price fluctuates more in older birth year groups as older people become eligible to pensioner concession price over the period.

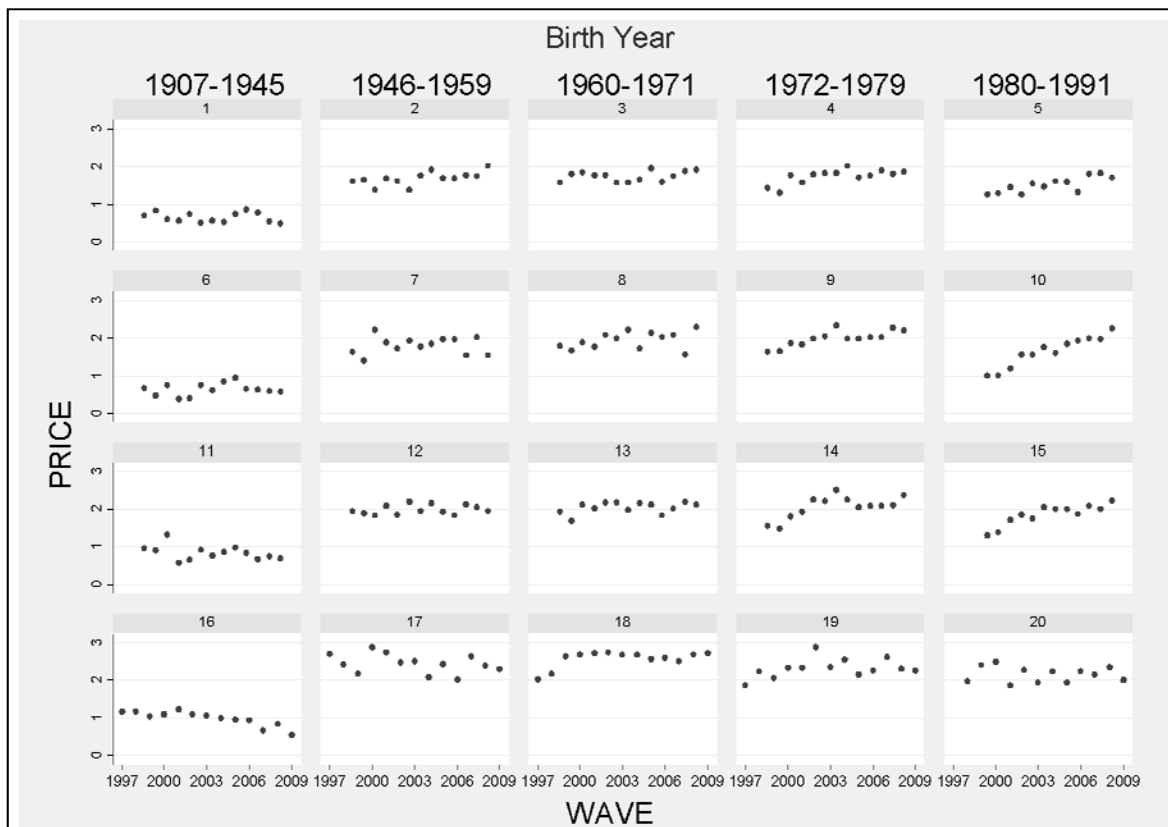


Figure 4.8 Time Trends of Public Transport Trip Price from 1997-2009 by Groups

The personal income change over time in Figure 4.9 shows that the average personal income has been increasing for the younger groups largely as a result of changes of social status as they become older. In comparison, the income of middle-age groups fluctuates over time with no substantial increase or decrease identified, whereas the income of older groups appears to be more stable as a result of most people in the older groups being of retirement age.

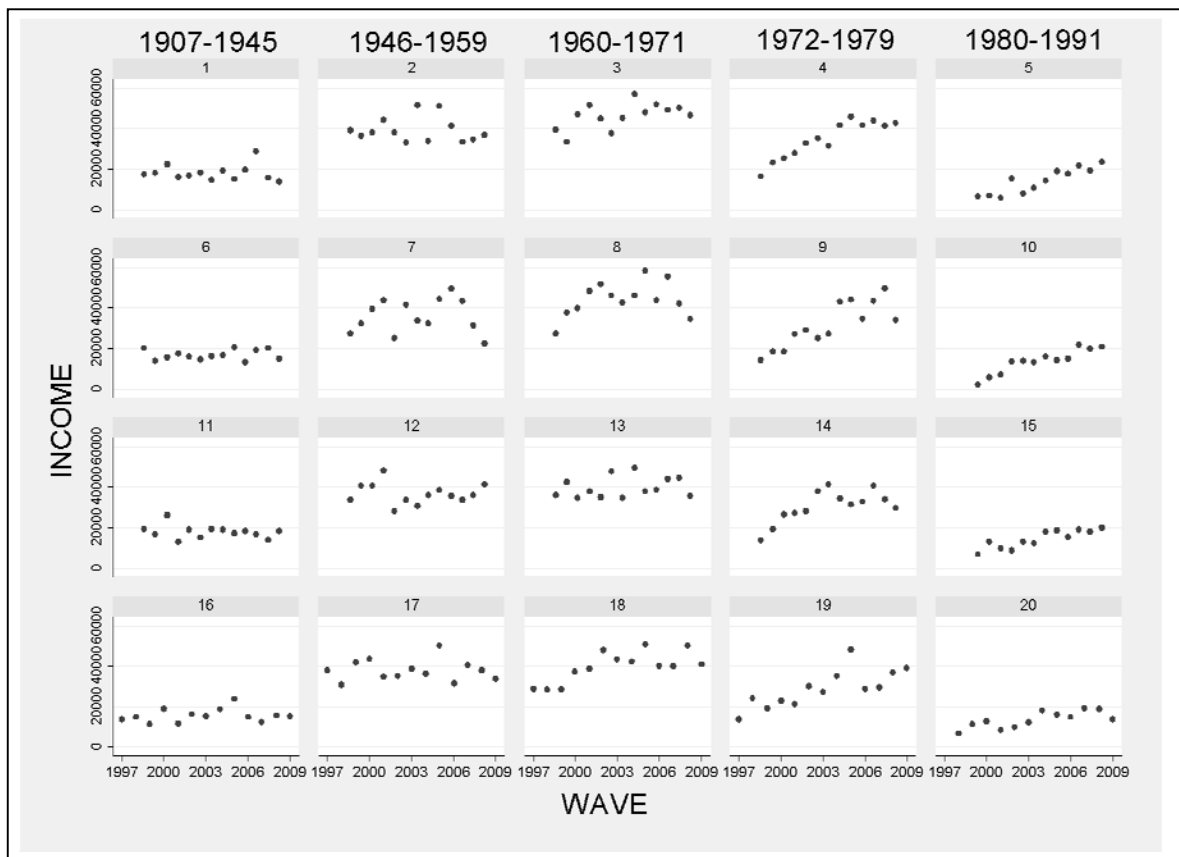


Figure 4.9 Time Trends of Personal Income from 1997-2009 by Groups

The exploratory analysis on these time-varying variables of the pseudo panel dataset suggests that the time-series variation is not consistent across the created groups, and these variations over time are not as substantial as variations across groups. This indicates that the inter-group variation needs to be taken into account in the estimation process because it is substantially larger than within-group variation. This finding requires further discussion of pseudo panel estimation techniques presented which is in the next section.

4.5 Pseudo panel data model

The general form of the public transport demand model (Equation (4.1)) defines the public transport demand $D_{i,t}$ for an individual i in period t is determined by public transport price $P_{i,t}$, a vector of travellers' socio-economic factors $E'_{i,t}$, public transport supply $S_{i,t}$ and a vector of land use characteristics $L'_{i,t}$.

$$D_{i,t} = f(P_{i,t}, E'_{i,t}, S_{i,t}, L'_{i,t}) \quad \text{Equation (4.1)}$$

To capture the dynamics of public travel behaviour adjustments, a partial adjustment model is employed to take account of the effect of previous behaviour on current behaviour as specified in Equation (4.2), where the public transport demand for an individual i in period $t - 1$ ($D_{i,t-1}$) is assumed to have an impact on the current demand in period t , with the coefficient λ representing the speed of adjustments.

$$D_{i,t} = f(P_{i,t}, E'_{i,t}, S_{i,t}, L'_{i,t}) + \lambda * D_{i,t-1} \quad \text{Equation (4.2)}$$

Assuming a linear relationship between public transport demand and its explanatory variables, the static and dynamic models are expanded as Equation (4.3) and Equation (4.4) respectively, where β_0 is the constant and $u_{i,t}$ is the combined error term constituted of the unobserved fixed individual effect α_t and the independent error term $\varepsilon_{i,t}$.

$$D_{i,t} = \beta_0 + \beta_P P_{i,t} + \beta_E E'_{i,t} + \beta_S S_{i,t} + \beta_L L'_{i,t} + u_{i,t}, \quad u_{i,t} = \alpha_t + \varepsilon_{i,t} \quad \text{Equation (4.3)}$$

$$D_{i,t} = \beta_0 + \lambda D_{i,t-1} + \beta_P P_{i,t} + \beta_E E'_{i,t} + \beta_S S_{i,t} + \beta_L L'_{i,t} + u_{i,t}, \quad u_{i,t} = \alpha_t + \varepsilon_{i,t} \quad \text{Equation (4.4)}$$

The pseudo panel data model introduced by Deaton (1985) uses average cohort data aggregated from individuals with the static and dynamic pseudo panel data models written as the following forms:

$$\bar{D}_{g,t} = \beta_0 + \beta_P \bar{P}_{g,t} + \beta_E \bar{E}'_{g,t} + \beta_S \bar{S}_{g,t} + \beta_L \bar{L}'_{g,t} + \bar{u}_{g,t}, \quad \bar{u}_{g,t} = \bar{\alpha}_{g,t} + \bar{\varepsilon}_{g,t} \quad \text{Equation (4.5)}$$

$$\bar{D}_{g,t} = \bar{\beta}_0 + \lambda \bar{D}_{g,t-1} + \beta_P \bar{P}_{g,t} + \beta_E \bar{E}'_{g,t} + \beta_S \bar{S}_{g,t} + \beta_L \bar{L}'_{g,t} + \bar{u}_{g,t}, \quad \bar{u}_{g,t} = \bar{\alpha}_{g,t} + \bar{\varepsilon}_{g,t} \quad \text{Equation (4.6)}$$

Compared to the genuine panel data model (Equation (4.3) and Equation (4.4)), Equation (4.5) and Equation (4.6) use the subscript g instead of i to denote the created groups in the pseudo panel data instead of individuals in the genuine panel data. These variables represent the way in which the observation of each variable is the mean value for all individuals classified into group g in period t .

As reviewed in Section 2.4.3, the main difference between pseudo panel data models (Equation (4.5) and Equation (4.6)) and genuine panel data models (Equation (4.3) and Equation (4.4)) in terms of model estimation is that because the cohorts within the same group are constituted of different members, the average unobserved group effect $\bar{\alpha}_{gt}$ is time-varying in contrast to the unobserved individual effect (α_i) which is fixed in a genuine panel data model. The result is that the time-varying group effects will not be eliminated through the demeaned transformation in the standard fixed effect estimation, so the conventional fixed effect estimator will be problematic for both in the static or dynamic pseudo panel models.

Most pseudo panel data studies reviewed in Section 2.3 have adopted the Fixed Effect (FE) estimator as the preferred estimator (Gassener, 1998; Gardes et al., 2005; Huang, 2007; Weis and Axhausen, 2009; Warunsiri and McNown, 2010) for static model estimation. This follows Deaton (1985) and Verbeek and Nijman (1992) who found that, with a sufficiently large cohort size (n_c) with sufficient inter-group variation, the time-varying $\bar{\alpha}_{gt}$ can be treated as constant over time as $\bar{\alpha}_g$, so that the pseudo panel data can be treated as genuine panel data using conventional estimation techniques.

Dynamic models can also be estimated on pseudo panel data. As with genuine panel data, the lagged dependent variable is likely to be correlated with the error term which causes estimation bias (Nickell, 1981). Some pseudo panel studies have employed Instrumental Variable (IV) estimators address the endogeneity problem (Dargay and Vythoukcas, 1999, Bernard et al., 2011) and results show that the IV estimator should be chosen over the FE estimator. However, the IV estimator has been criticised for its inefficiency as a consequence of using instruments (Beck and Katz, 2011) and bias when the number of panel units is not sufficiently large (Bruno, 2005a).

From the pseudo panel data literature, there has not been a "superior" estimator suggested for either the static model or the dynamic model. The common practice is to employ the FE estimator for static models by ignoring the measurement

errors if the cohort sizes are considered to be sufficiently large (i.e., one hundred members as a rule of thumb) to ensure the time-varying unobserved group effect is a serious issue in model estimation. However, Plümper and Troeger (2007) demonstrated that the FE estimator will be inefficient if panel data have much larger between-group variation than within-group variation. This indicates that the FE estimator typically used for pseudo panel data model estimation is likely to be inefficient and thus generate unreliable statistical inference if the ratio of between-group variation and within-group variation is large. This property of pseudo panel data has not been well acknowledged by previous applied pseudo panel data studies. Indeed, the unique properties of pseudo panel datasets, and the way in which between-group and within-group variation can differ mean that there is no defaulted rule for which in the best estimation process. This issue is investigated in Chapter 5.

4.6 Summary

This chapter details the construction process of the pseudo panel dataset for this study. It discusses selection of grouping criteria and the determination of cohort sizes. The grouping criteria used to create cohorts are first evaluated to ensure the grouping approach is able to create sufficient inter-group variation. The discussion on the scale of grouping bands has also addressed the issue of a limited number public transport observations from the survey data. The adoption of a variable range for defining groups by age and by distance to CBD appears to be a good approach to approximately equalise the number of individual records in each cohort to reduce the number of small cohorts in the pseudo panel dataset. This discussion on constructing pseudo panel datasets for limited sample observations has not yet been identified in published pseudo panel data research.

The effectiveness of the grouping approach is investigated in Section 4.4. The statistics for the between-group and within-group variances show that the heterogeneity between the created groups is fairly substantial, which meets the aim of the pseudo panel data construction. The variations in public transport demand as well as the explanatory variables are also identified through generation effects, life-cycle effects, and location effects, which demonstrate that

the pseudo panel dataset shows more individual information, as compared to aggregate data.

In the final section, the theoretical foundations for pseudo panel estimation are outlined. Whilst the literature suggests pseudo panel models may be empirically as genuine panel data models, the discussion of pseudo panel data model and the estimation techniques suggests that the pseudo panel data model should not be simply treated as genuine panel data models due to its unique properties such as minimal within-group variation in some explanatory variables. These unique properties of pseudo panel data mean that simply adapting panel models without rigorously examining the performance of applied estimators could potentially lead to estimation bias or inefficiency, as commonly applied in previous pseudo panel data research. With regard to this, Chapter 5 presents a Monte Carlo simulation experiment to examine the performance of various estimators for pseudo panel data models, whilst incorporating the properties of pseudo panel data that are expected to influence the estimation results.

CHAPTER 5 MONTE CARLO SIMULATION

5.1 Introduction

Applied pseudo panel data research reported in the literature (Gassner, 1998, Dargay and Vythoukas, 1999, Gardes et al., 2005, Dargay, 2007, Huang, 2007, Weis and Axhausen, 2009, Warunsiri and McNown, 2010, Bernard et al., 2011) have empirically estimated pseudo panel data as if they were genuine panel data, and conventional estimation techniques such as pooled Ordinary Least Squares (OLS), Fixed Effect (FE), Random Effect (RE) and Instrumental Variable (IV) estimators are commonly applied in pseudo panel data research.

However, as discussed in Section 4.5, some unique properties of pseudo panel data need to be taken into consideration in the estimation process. Evaluating the various estimators for pseudo panel data models require a Monte Carlo simulation experiment, and there has been no discussion in the literature as to why a particular estimation technique might be better and hence there is a need to investigate this.

The unique properties of pseudo panel data that may lead to problematic estimation results when using conventional panel data estimators are discussed below.

As a consequence of how cohorts are created, the cross-sectional variation of the exogenous variables among cohorts across different groups (between-group variance) is usually larger than the variation of the exogenous variables among cohorts in the same group across time (within-group variance). Variables with relatively small within-group variance known as “rarely changing variables” may lead to inefficient estimation for some estimators in panel data analysis, especially for the FE estimator that only takes account of within-group variation as recently discussed in Plümer and Troeger (2007, 2011).

With pseudo panel data, the unobserved group effect is time-varying because each cohort, even within the same group, is composed of different individuals

over time as demonstrated in Section 4.3. Hence, non-spherical errors such as heteroscedasticity are likely to be introduced which cannot be controlled through conventional estimation techniques that incorporate only fixed individual or group effects. Moreover, since repeated cross-sectional surveys are not primarily concerned with understanding longitudinal questions at a disaggregate level, a rather small number of groups (G) and short time periods (T) are normally obtained as compared to aggregate genuine panel data, resulting in finite-sample properties being embedded in pseudo panel data. Therefore, as seen in Section 4.3, pseudo panel data construction has a trade-off between the cohort size (number of individuals in a cohort) and the number of groups (G). Increasing the cohort size reduces the estimation bias, but it also decreases the estimation efficiency because the number of groups being estimated is reduced (Verbeek and Nijman, 1992).

Given the features identified above, applying estimation techniques developed for genuine panel data to finite-sample pseudo panel data without recognising its unique properties, as is commonly practised in the literature, may lead to problematic estimation results and invalid policy interpretations. This chapter³ investigates the performance of various estimators in static and dynamic pseudo panel models, whilst taking account the properties of pseudo panel data including: time-varying unobserved group effect; larger between-group variance than within-group variance; a small total number of cohorts; and the trade-off between cohort size and number of groups. The performance of estimators is measured by the degree of bias, efficiency, and Root Mean Square Error (RMSE) for each estimator. The property of consistency which is typically used to examine the asymptotic behaviour of estimators in large samples is not examined in this experiment, because the main contribution of this exercise is to provide empirical suggestions for estimating the pseudo panel model of this study, which has a finite sample property. The purpose of this investigation is to identify benchmarks to analyse the suitability of different estimation techniques (as

³ The work presented in this chapter has been published in Tsai et al. (2013). The author wishes to acknowledge the contribution of the co-authors: Waiyan Leong, Corinne Mulley, and Geoffrey Clifton.

reviewed in Section 2.4), under the properties of pseudo panel data as highlighted above.

5.2 Experiment design

The following static model in Equation (5.1) and dynamic Partial Adjustment Model (PAM) in Equation (5.2) for pseudo panel data are designed for the data generating process of the simulation experiment. The simulation models simplify the pseudo panel data models introduced in Section 4.5 with one single exogenous variable (\bar{x}_{gt}) and one lagged dependent variable (\bar{y}_{gt-1}) being employed. This simplification is to avoid the confounding results from possible interactions between multivariate exogenous variables.

$$\bar{y}_{gt} = \beta_0 + \beta_1 \bar{x}_{gt} + \bar{u}_{gt} \quad \text{Equation (5.1)}$$

$$\bar{y}_{gt} = \lambda \bar{y}_{gt-1} + \beta_1 \bar{x}_{gt} + \bar{u}_{gt} \quad \text{Equation (5.2)}$$

where

$$\bar{u}_{gt} = \bar{\alpha}_g + \bar{\omega}_{gt} + \bar{\varepsilon}_{gt}$$

$$\bar{\alpha}_g \sim N(0, \sigma_{\alpha}^2)$$

$$\bar{\omega}_{gt} \sim N(0, (\sigma_{\alpha} / \sqrt{n_c})^2), \text{ and } n_c \sim N(\bar{n}_c, \sigma_{n_c}^2)$$

$$\bar{\varepsilon}_{gt} \sim N(0, 1^2)$$

The composite error term \bar{u}_{gt} includes three elements: the unobserved group effect ($\bar{\alpha}_g$) for a given created group in the pseudo panel data set, the time varying cohort effect within groups ($\bar{\omega}_{gt}$), and independent identically distributed (*i.i.d.*) disturbances ($\bar{\varepsilon}_{gt}$). σ_{α}^2 is an experimentally controlled variable to simulate the variance of the unobserved group effect. Allowing $\bar{\omega}_{gt}$ to be drawn from a random distribution allows time variation in the cohort effect since within-group cohorts across time are not created from the same individuals in contrast to genuine panel data. $\bar{\omega}_{gt}$ is assumed to be normally distributed with a mean of zero, and its variance is positively related to σ_{α}^2 but negatively related to the square root of cohort size (n_c). This assumption is a convenient way of allowing the variance of $\bar{\omega}_{gt}$ to be smaller than the variance of $\bar{\alpha}_g$ and to allow a larger n_c to reduce the variance of $\bar{\omega}_{gt}$. To simulate the unequal cohort sizes found in real

life pseudo panel data, n_c is assumed to be normally distributed across all the created cohorts, with a mean and variance bounded in the experiment by values computed from the pseudo panel data constructed from the Sydney Household Travel Survey (SHTS) data.

As highlighted in Section 5.1, the features of pseudo panel data requiring examination include: time-varying unobserved group effects ($\bar{\omega}_{gt}$); larger between-group variance in the exogenous variable ($\sigma_{B,x}^2$) than within-group variance ($\sigma_{W,x}^2$); small total number of cohorts (C); and the trade-offs between cohort size (n_c) and number of groups (G). $\bar{\omega}_{gt}$ has been incorporated in the simulation models as specified in Equation (5.1) and (5.2). The other properties are examined through seven scenarios (Table 5.1). Each scenario is replicated for one thousand times in the simulation experiment. The experiment uses Stata 12.0 package to program the simulation models. Parts of the programming codes are provided in Appendix 3.

Table 5.1 Scenario Design for Monte Carlo Experiments

Scenario	Variance in exogenous variable (\bar{x}_{gt})	Distribution of unobserved group effects ($\bar{\alpha}_g$)	Size of data ($G; T$)	Cohort Size (n_c)
1	$\sigma_{B,x}^2 = \sigma_{W,x}^2 = 1$; $E(\bar{x}_{gt}) = 0$	$\bar{\alpha}_g \sim N(0, 0.5^2)$	$G=12; T=13$	$n_c \sim N(150, 50^2)$
2	$\sigma_{B,x}^2 = \sigma_{W,x}^2 = 1$; $E(\bar{x}_{gt}) = 0$	$\bar{\alpha}_g \sim N(0, 0.2^2)$	$G=12; T=13$	$n_c \sim N(150, 50^2)$
3	$(\sigma_{B,x}^2, \sigma_{W,x}^2) = (0.5^2, 0.2^2)$; $E(\bar{x}_{gt}) = 0$	$\bar{\alpha}_g \sim N(0, 0.5^2)$	$G=12; T=13$	$n_c \sim N(150, 50^2)$
4	$(\sigma_{B,x}^2, \sigma_{W,x}^2) = (0.5^2, 0.2^2)$; $E(\bar{x}_{gt}) = 0$	$\bar{\alpha}_g \sim N(0, 0.2^2)$	$G=12; T=13$	$n_c \sim N(150, 50^2)$
5	$(\sigma_{B,x}^2, \sigma_{W,x}^2) = (0.2^2, 0.5^2)$; $E(\bar{x}_{gt}) = 0$	$\bar{\alpha}_g \sim N(0, 0.5^2)$	$G=12; T=13$	$n_c \sim N(150, 50^2)$
6	$(\sigma_{B,x}^2, \sigma_{W,x}^2) = (0.5^2, 0.2^2)$; $E(\bar{x}_{gt}) = 0$	$\bar{\alpha}_g \sim N(0, 0.5^2)$	$G=36; T=13$	$n_c \sim N(50, 15^2)$
7	$(\sigma_{B,x}^2, \sigma_{W,x}^2) = (0.5^2, 0.2^2)$; $E(\bar{x}_{gt}) = 0$; $\text{corr}(\bar{x}_{gt}, \bar{\alpha}_g) = 0.5$	$\bar{\alpha}_g \sim N(0, 0.5^2)$	$G=12; T=13$	$n_c \sim N(150, 50^2)$

Scenario 1 and Scenario 2 are designed as the base scenarios which assume that the between-group variance and the within-group variance are the same in order to investigate the impact of unobserved group effects ($\bar{\alpha}_g$) on the estimation performance. Scenarios 3/4/6 are designed to allow for the commonly observed feature of pseudo panel datasets of larger between-group variance than within-group variance ($\sigma_{B,x}^2 > \sigma_{W,x}^2$), as compared to $\sigma_{B,x}^2 = \sigma_{W,x}^2$ in Scenario 1/2. Scenario 5 reverses Scenarios 3/4/6 and assumes that $\sigma_{B,x}^2 < \sigma_{W,x}^2$. The magnitude of $\sigma_{\bar{\alpha}}^2$ is likely to have an impact on the estimation results because it is the factor that causes endogeneity and non-spherical errors, so a larger (i.e., $\sigma_{\bar{\alpha}} = 0.5$) and a smaller effect (i.e., $\sigma_{\bar{\alpha}} = 0.2$) are tested between Scenario 1 and 2 as well as Scenario 3 and 4.

Scenario 6 is designed for a larger number of groups (G), with a correspondingly smaller cohort size (n_c), as compared to Scenario 3. G and n_c will be inversely related as in the situation, as for the situations where the total number of sampled individuals is fixed. The simulation data in the experiment is made more similar to the real data used in this study by conditioning the values of the between-group variance, the within-group variance, the number of cohorts and the cohort size on the SHTS pseudo panel data.

Scenario 7 allows for correlation between the explanatory variable and the unobserved group effect. Investigating the impact of the correlation on the estimation results is important because this is the main difference between the theoretical assumptions of the FE and RE estimators. This correlation is only introduced in Scenario 7 to avoid confounding other results due to the other experimental conditions of Scenarios 1 to Scenario 6.

In all scenarios the number of time periods (T) is kept constant at thirteen, with the number of groups (G) being changed only in Scenario 6 for the main purpose of investigating the trade-off between cohort size and number of groups. The consistency of an estimator is not the primary focus in this analysis because the aim of this simulation experiment is to examine estimator properties within the constraints of real data estimation where there is little flexibility in expanding

the number of panel units (whether in pseudo or genuine panel data) or number of time periods whilst keeping the cohort size constant.

5.3 Estimators and performance measurements

The properties of estimators commonly used in pseudo panel studies are examined in this Monte Carlo experiment. The theoretical foundations of these panel data estimators are reviewed in Section 2.4 and their key properties are summarised in Table 5.2.

Table 5.2 Summary of Estimator Properties

Estimator	Theoretical Properties
Pooled Ordinary Least Squares (pooled OLS)	<ol style="list-style-type: none"> 1. Does not control for unobserved individual effects 2. Biased in the presence of unobserved individual effects 3. Inefficient in the presence of non-spherical errors
Fixed Effect (FE)	<ol style="list-style-type: none"> 1. Controls for unobserved individual effects 2. Allows correlation between explanatory variables and unobserved individual effects 3. Biased in dynamic model estimation 4. Efficient with time-varying variables; inefficient when variables rarely change over time
Random Effect (RE)	<ol style="list-style-type: none"> 1. Controls for unobserved individual effects 2. Assumes no correlation between explanatory variables and unobserved individual effects 3. Biased in dynamic model estimation 4. Weights between-group variances and within-group variances
Panel-Corrected Standard Error (PCSE)	<ol style="list-style-type: none"> 1. Accounts for non-spherical errors 2. Corrects serial correlation and cross-sectional dependency in estimation 3. Biased in the presence of unobserved individual effects
Instrumental Variable (IV, including GMM)	<ol style="list-style-type: none"> 1. Controls for unobserved individual effects by using instrumental variables 2. Inefficient when number of panel units is small

For static models, the pooled OLS estimator is used as a benchmark to be compared to FE, RE and PCSE estimators. For dynamic models, the System Generalized Method of Moments (GMM) (Blundell and Bond, 1998) is included because of its ability to incorporate the endogeneity between the lagged dependent variable and error terms, and because it has been suggested as an appropriate estimator for pseudo panel data from previous Monte Carlo experiments under certain conditions (McKenzie, 2004, Inoue, 2008). The GMM

method applied in this analysis employs the second lag of the dependent variable (\bar{y}_{gt-2}) as an instrumental variable.

Two fundamental criteria to measure an estimator's performance are bias and efficiency. Bias refers to the expectation of the difference between the value of the parameter estimate and its assumed value in the experiment. An efficient estimator is an unbiased estimator with the least variance. If none of the evaluated estimators have minimum variance among all possible estimators, a measurement of relative efficiency can be used to compare the performance of the applied estimators. A more efficient estimator requires fewer observations to achieve the same statistical power, has smaller standard errors of estimates when the same number of observations is applied to the estimation procedures being compared, and thus generates more reliable statistical inferences.

The choice of estimators often depends on both bias and efficiency considerations. An unbiased but inefficient estimator is not necessarily superior to a biased but efficient estimator. When a "best" estimator needs to be determined, a common approach is to use the RMSE as specified in Equation (5.3) as an overall performance measure (Judson and Owen, 1999) which equally weights bias and efficiency.

$$\text{RMSE} = \sqrt{\text{Bias}(\tilde{\beta})^2 + \text{Var}(\tilde{\beta})} \quad \text{Equation (5.3)}$$

5.4 Analysis of Monte Carlo simulation

5.4.1 Simulation results for static models

Scenario 1 to 3 are analysed for the static model and the results are presented in Table 5.3. The simulation results from Scenario 1, which assumes an identically distributed \bar{x}_{gt} across time and groups and a large variance for the unobserved group effect ($\sigma_{\bar{\alpha}} = 0.5$), show that there is no substantial bias in the small pseudo panel data set ($G=12, T=13$). The FE and RE estimators, as expected, are more efficient than pooled OLS and PCSE in the presence of unobserved group effects $\bar{\alpha}_g$. In Scenario 2 where $\sigma_{\bar{\alpha}}^2$ is reduced, it can be seen that the FE and RE estimators are not necessarily more efficient than pooled OLS and PCSE

estimators. This is because $\bar{\alpha}_g$ which makes pooled OLS and PCSE estimators inefficient now has a much smaller impact on the estimation process. In these two scenarios, using PCSE to correct the non-spherical errors does not substantially improve the efficiency of pooled OLS, indicating that the simulation data generated for pseudo panel data do not possess strong non-spherical errors when \bar{x}_{gt} is identically distributed.

Table 5.3 Simulation Results for Static Models

Model: $\bar{y}_{gt} = 0.2 + 0.8 * \bar{x}_{gt} + \bar{u}_{gt}$				
	Pooled OLS	FE	RE	PCSE
Scenario 1: $\sigma_{B,x}^2 = \sigma_{W,x}^2 = 1; \bar{\alpha}_g \sim N(0, 0.5^2)$				
β_1	0.803	0.805	0.805	0.803
$\beta_1_SE^1$	0.090	0.084	0.083	0.089
β_1_BIAS	0.003	0.005	0.005	0.003
β_1_RMSE	0.090	0.084	0.083	0.089
Scenario 2: $\sigma_{B,x}^2 = \sigma_{W,x}^2 = 1; \bar{\alpha}_g \sim N(0, 0.2^2)$				
β_1	0.798	0.798	0.798	0.798
$\beta_1_SE^1$	0.082	0.084	0.082	0.081
β_1_BIAS	-0.002	-0.002	-0.002	-0.002
β_1_RMSE	0.082	0.084	0.082	0.082
Scenario 3: $(\sigma_{B,x}^2, \sigma_{W,x}^2) = (0.5^2, 0.2^2); \bar{\alpha}_g \sim N(0, 0.5^2)$				
β_1	0.791	0.783	0.789	0.791
$\beta_1_SE^1$	0.181	0.419	0.265	0.162
β_1_BIAS	-0.009	-0.017	-0.011	-0.009
β_1_RMSE	0.181	0.420	0.266	0.162

¹ Standard errors

Scenario 3 simulates data with a larger between-group cross-sectional variance ($\sigma_{B,x}^2$) for the exogenous variable \bar{x}_{gt} , compared to its within-group time variance ($\sigma_{W,x}^2$). Comparing Scenario 3 with Scenario 1, the most noticeable difference is that the standard error of the FE estimator substantially increases from 0.084 to 0.419, whilst the standard errors of other estimators have a relatively minor increase. This result shows that when there is larger between-group variation in \bar{x}_{gt} , the FE estimator, which only takes account of within-group variation, is inefficient. Comparing the results from pooled OLS, RE, and PCSE estimators, it can be seen that the PCSE estimator has improved the efficiency of the pooled OLS estimator as the standard error of β_1 drops from 0.181 to 0.162, showing that non-spherical errors are more influential in this case

than in Scenarios 1 and 2 where \bar{x}_{gt} is identically distributed. It is also important to note that the bias in Scenario 3, for all estimators, is increased by more than two hundred percent as compared to Scenario 1, although the magnitude of the bias remains small.

In summary, there is no one superior estimator for static models under the scenario where \bar{x}_{gt} is identically distributed (i.e., $\sigma_{B,x}^2 = \sigma_{W,x}^2$). In contrast, when larger between-group cross-sectional variance relative to within-group time variance is present ($\sigma_{B,x}^2 > \sigma_{W,x}^2$), the FE estimator is particularly inefficient because the explanatory variable is rarely changing over time which confirms the finding in previous research (Plümper and Troeger, 2007, Plümper and Troeger, 2011). In this case, the PCSE estimator is suggested as the preferred estimator because of its lower RMSE.

5.4.2 Simulation results for dynamic models

The dynamic model simulation results for Scenarios 1/2/3/4 are summarised in Table 5.4. For Scenario 1 and Scenario 2 where \bar{x}_{gt} is identically distributed across time and groups, λ_1 clearly shows an upward bias when using pooled OLS and a downward bias when using the FE estimator: this identifies the Nickell bias (Nickell, 1981) as reviewed in Section 2.4.2. The pooled OLS bias in λ_1 is reduced in Scenario 2 when $\sigma_{\bar{\alpha}}$ is lowered to 0.2 but the bias remains the same in FE. This is because the bias of pooled OLS comes from the interaction of λ and $\sigma_{\bar{\alpha}}$, whereas the bias of FE results from the correlation between the transformed lagged dependent variable and the transformed error terms (Baltagi, 2008). In contrast, β_1 does not show obvious bias for all estimators in either scenario. This suggests that the endogeneity problem in dynamic models only makes the estimate of λ_1 problematic but does not have a strong impact on β_1 . In these two scenarios, FE performs better than other estimators when $\sigma_{\bar{\alpha}}$ is large, but FE may not be the favoured estimator when $\sigma_{\bar{\alpha}}$ is small. On the other hand, GMM appears to be inefficient given the relatively large standard errors for both parameters. This concurs with Kiviet's (1995) finding that IV estimation methods may lead to small sample bias and large standard errors.

It is also important to note that the RE estimator gives almost identical results to the pooled OLS estimation results in the dynamic model. This effect has been identified by Baltagi (2008, p. 20) and the reason is that the variance of the unobserved group effect ($\sigma_{\bar{\alpha}}^2$) may be negative and replaced by zero from the RE estimation process when it is minimal in the composite error term⁴. This effect is not identified in Scenario 1 and Scenario 3 of the static model estimation where $\sigma_{\bar{\alpha}} = 0.5$ but is evident in Scenario 2 where $\sigma_{\bar{\alpha}} = 0.2$, showing that the smaller variance of unobserved individual (or group) effects may degenerate the RE estimator into the pooled OLS estimator. In the dynamic model estimation the degeneration happens either when $\sigma_{\bar{\alpha}} = 0.5$ or $\sigma_{\bar{\alpha}} = 0.2$ possibly because the inclusion of the lagged dependent variable increases the explanatory power and thus reduces the impact of the unobserved individual effect on the estimation results.

Table 5.4 Simulation Results (Scenarios 1 – 4) for Dynamic Models

Model: $\bar{y}_{gt} = 0.2 * \bar{y}_{gt-1} + 0.8 * \bar{x}_{gt} + \bar{u}_{gt}$											
	Pooled OLS	FE	RE	PCSE	GMM		Pooled OLS	FE	RE	PCSE	GMM
Scenario 1: $\sigma_{B,x}^2 = \sigma_{W,x}^2 = 1; \bar{\alpha}_g \sim N(0, 0.5^2)$											
λ_1	0.328	0.134	0.328	0.328	0.282	β_1	0.798	0.795	0.798	0.798	0.790
λ_{1_SE}	0.065	0.069	0.065	0.085	0.249	β_{1_SE}	0.093	0.088	0.093	0.091	0.106
λ_{1_BIAS}	0.128	-0.066	0.128	0.128	0.082	β_{1_BIAS}	-0.002	-0.005	-0.002	-0.002	-0.010
λ_{1_RMSE}	0.143	0.096	0.143	0.154	0.262	β_{1_RMSE}	0.093	0.089	0.093	0.091	0.107
Scenario 2: $\sigma_{B,x}^2 = \sigma_{W,x}^2 = 1; \bar{\alpha}_g \sim N(0, 0.2^2)$											
λ_1	0.219	0.134	0.219	0.219	0.215	β_1	0.799	0.795	0.799	0.799	0.787
λ_{1_SE}	0.065	0.069	0.065	0.081	0.242	β_{1_SE}	0.086	0.088	0.086	0.084	0.103
λ_{1_BIAS}	0.019	-0.066	0.019	0.019	0.015	β_{1_BIAS}	-0.001	-0.005	-0.001	-0.001	-0.013
λ_{1_RMSE}	0.068	0.095	0.068	0.084	0.243	β_{1_RMSE}	0.086	0.088	0.086	0.084	0.104
Scenario 3: $(\sigma_{B,x}^2, \sigma_{W,x}^2) = (0.5^2, 0.2^2); \bar{\alpha}_g \sim N(0, 0.5^2)$											
λ_1	0.366	0.098	0.365	0.366	0.319	β_1	0.659	0.816	0.660	0.659	0.688
λ_{1_SE}	0.077	0.087	0.077	0.109	0.293	β_{1_SE}	0.198	0.440	0.198	0.197	0.365
λ_{1_BIAS}	0.166	-0.102	0.165	0.166	0.119	β_{1_BIAS}	-0.141	0.016	-0.140	-0.141	-0.112
λ_{1_RMSE}	0.183	0.134	0.183	0.199	0.316	β_{1_RMSE}	0.243	0.441	0.243	0.243	0.382
Scenario 4: $(\sigma_{B,x}^2, \sigma_{W,x}^2) = (0.5^2, 0.2^2); \bar{\alpha}_g \sim N(0, 0.2^2)$											
λ_1	0.221	0.097	0.221	0.221	0.208	β_1	0.778	0.790	0.778	0.778	0.782
λ_{1_SE}	0.081	0.087	0.081	0.111	0.303	β_{1_SE}	0.187	0.441	0.187	0.187	0.325
λ_{1_BIAS}	0.021	-0.103	0.021	0.021	0.008	β_{1_BIAS}	-0.022	-0.010	-0.022	-0.022	-0.018
λ_{1_RMSE}	0.083	0.135	0.083	0.113	0.303	β_{1_RMSE}	0.188	0.441	0.188	0.188	0.326

⁴ The variance of unobserved individual effect (Baltagi, 2008, p.20): $\hat{\sigma}_{\bar{\alpha}}^2 = [(T \sum_{g=1}^N \bar{\alpha}_g^2 / G) - \hat{\sigma}_{\bar{\epsilon}}^2] / T$

In Scenario 3 and Scenario 4 where \bar{x}_{gt} has a larger between-group variance ($\sigma_{B,x}^2$) than within-group variance ($\sigma_{W,x}^2$), the bias and standard errors of λ_1 are both increased as compared to Scenarios 1 and 2. β_1 also becomes more biased for all estimators suggesting that the large between-group variation scenarios are likely to induce bias in exogenous variable estimates which were unbiased when \bar{x}_{gt} was assumed to be identically distributed. As with the static model simulations, the standard errors of β_1 obtained from the FE estimator are increased by more than four hundred percent suggesting severe inefficiency in the FE estimator. Although FE is the least biased estimator, the inefficient standard errors will enlarge the confidence intervals and make the statistical inference unreliable. Therefore, the pooled OLS or RE estimator is favoured given the lowest combined RMSE of λ_1 and β_1 estimates in these scenarios.

The simulation results from Scenario 3 and Scenario 4 show that there is no one absolutely unbiased estimator for a dynamic model when the exogenous variable has a larger between-group variance than within-group variance. The degree of bias and efficiency of an estimator must be both taken into consideration when determining a preferred estimator. Figure 5.1 to Figure 5.4 visualise the density distributions of the parameter estimates of λ_1 and β_1 by comparing the pooled-OLS and FE estimation results in Scenario 3 and 4. Only Scenario 3 and 4 are discussed here because these two scenarios are closer to the data property of the pseudo panel dataset of this study. The bias of parameter estimates can be identified from the difference between the mean of the distribution and zero (given the assumptions of experimental design), whereas the efficiency can be observed from the bandwidth of the distribution.

For Scenario 3, Figure 5.1 shows that both the pooled OLS estimator and FE estimator have a similar degree of bias in λ_1 but in different directions, with the same level of efficiency. On the other hand, Figure 5.2 shows that the pooled OLS estimator is much more efficient than the FE estimator for β_1 , with minor bias identified in both estimators. Therefore, the pooled OLS estimator should be chosen over the FE estimator when both parameter estimates (λ_1 and β_1) are taken into consideration.

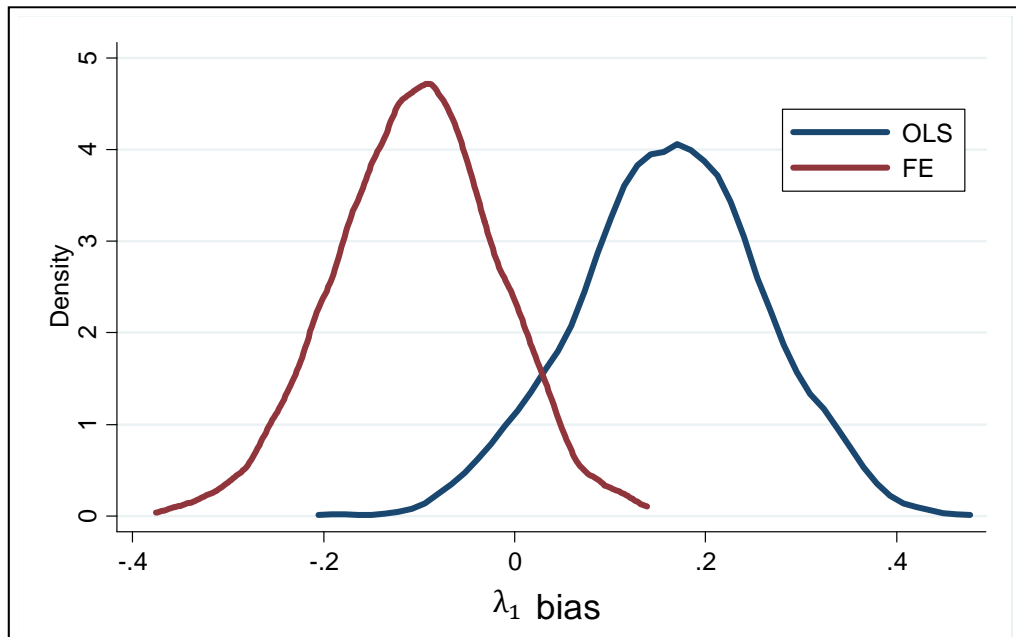


Figure 5.1 Density Plots of λ_1 Estimates from Pooled OLS and FE Estimators (Scenario 3)

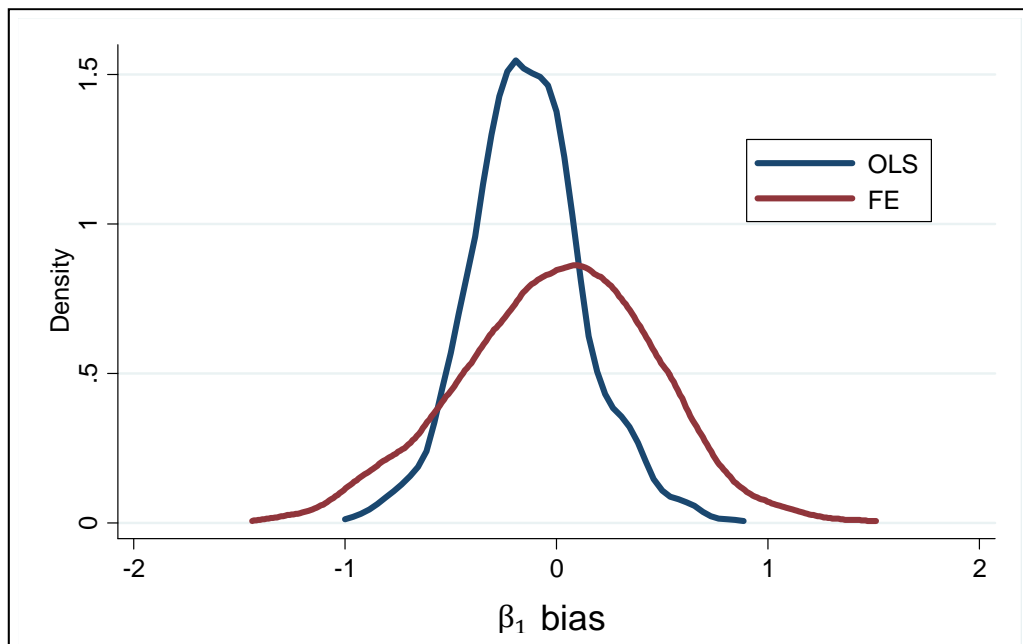


Figure 5.2 Density Plots of β_1 Estimates from Pooled OLS and FE Estimators (Scenario 3)

In Scenario 4 (Figure 5.3 and Figure 5.4) where the variance of the unobserved group effect is reduced, it can be seen that the bias of λ_1 in the pooled OLS estimation is substantially decreased, whereas the bias of λ_1 from the FE estimation remains at the same level. The bias of β_1 from both estimators is also reduced, but the FE estimator is still more inefficient than the pooled OLS estimator. This suggests that if the unobserved group effect can be minimised,

possibly through better model specification, the pooled OLS estimator has the potential to be an unbiased and efficient estimator for dynamic model estimation. This comparison highlights the importance of weighting both bias and efficiency when evaluating estimation performance using the visualised graphs from the simulation results.

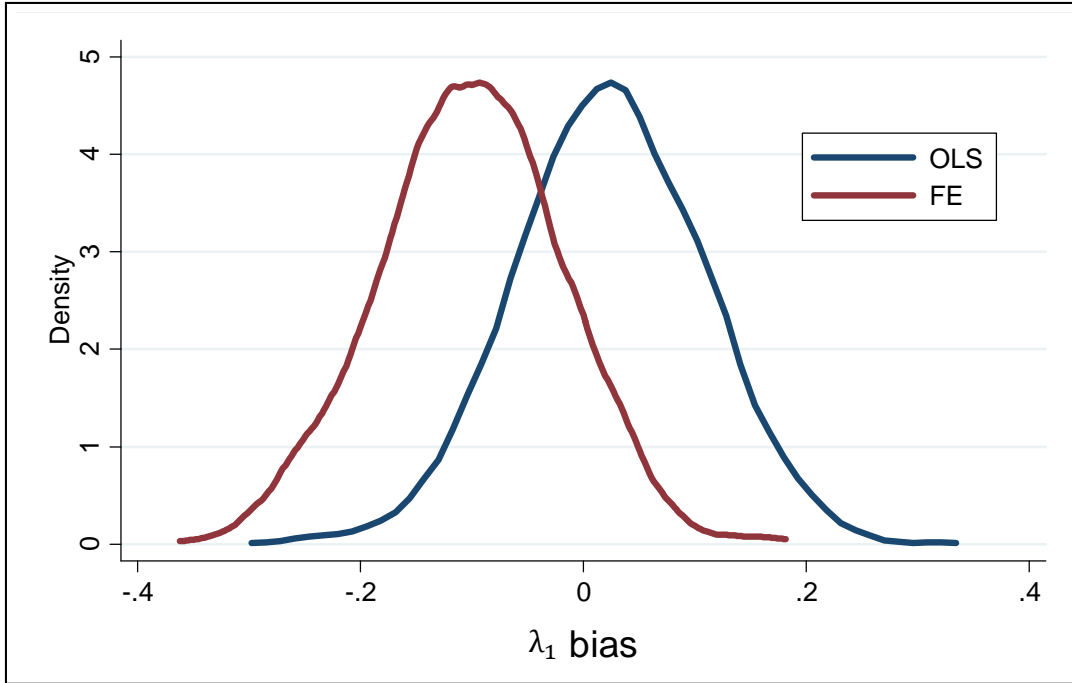


Figure 5.3 Density Plots of λ_1 Estimates from Pooled OLS and FE Estimators (Scenario 4)

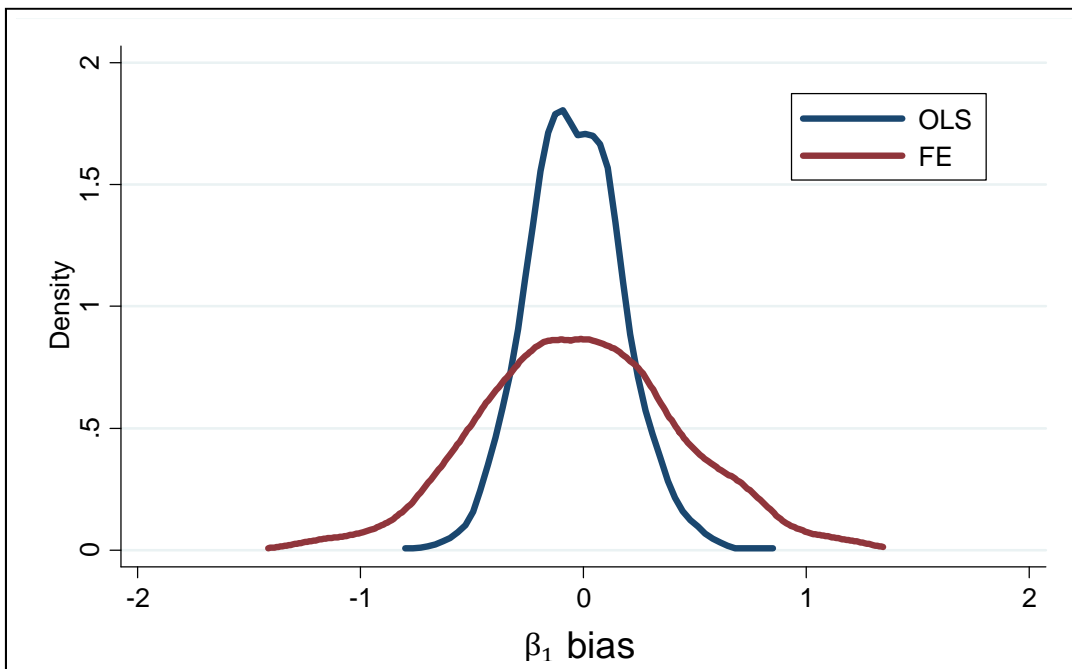


Figure 5.4 Density Plots of β_1 Estimates from Pooled OLS and FE Estimators (Scenario 4)

To further investigate the relationship between the distribution of \bar{x}_{gt} and estimators' performance, Scenario 5 assumes that $\sigma_{W,x}^2$ is larger than $\sigma_{B,x}^2$. In Table 5.5, the bias of β_1 in Scenario 5 is moderately reduced for all estimators as compared to scenario 3, whilst the bias of λ_1 does not change noticeably. Furthermore, all the standard errors are decreased, particularly for the FE estimator where the standard error drops from 0.440 to 0.176 for β_1 . Given the lowest combined RMSE, FE is the preferred estimator in this case.

Scenario 6 is used to evaluate the trade-off between cohort sizes (n_c) and number of groups (G). In this scenario, n_c is specified as $n_c \sim N(50, 15^2)$ as opposed to $n_c \sim N(150, 50^2)$ in previous scenarios. In Table 5.5, comparing Scenario 6 to Scenario 3 with the same variance of unobserved group effects and the same distribution of \bar{x}_{gt} , the results show that the biases are moderately increased for the pooled OLS, RE, and PCSE estimators, but are less evident for the FE and GMM estimators in Scenario 5. However, similar to the results of Verbeek and Nijman (Verbeek and Nijman, 1992), standard errors are reduced for all estimators by around 40 percent to 45 percent as a result of the increase in the number of groups. If the RMSE for λ_1 and β_1 are added up to form an overall measurement of error, then using a smaller n_c and a larger G as in Scenario 6 results in better results across all estimators as compared with Scenario 3. Therefore, reducing the average cohort size improves the overall statistical inference at a relatively low cost of increasing the bias. This results supports the choice of the two constructed pseudo panel datasets as discussed in Section 4.3, where the dataset with more groups but smaller average cohort size is chosen based on this simulation finding.

Table 5.5 Simulation Results (Scenarios 3/5/6) for Dynamic Models

Model: $\bar{y}_{gt} = 0.2 * \bar{y}_{gt-1} + 0.8 * \bar{x}_{gt} + \bar{u}_{gt}$

	Pooled						Pooled				
	OLS	FE	RE	PCSE	GMM		OLS	FE	RE	PCSE	GMM
Scenario 3: $(\sigma_{\bar{b},x}^2, \sigma_{\bar{w},x}^2)=(0.5^2, 0.2^2)$; $\bar{\alpha}_g \sim N(0,0.5^2)$											
λ_1	0.366	0.098	0.365	0.366	0.319	β_1	0.659	0.816	0.660	0.659	0.688
λ_{1_SE}	0.077	0.087	0.077	0.109	0.293	β_{1_SE}	0.198	0.440	0.198	0.197	0.365
λ_{1_BIAS}	0.166	-0.102	0.165	0.166	0.119	β_{1_BIAS}	-0.141	0.016	-0.140	-0.141	-0.112
λ_{1_RMSE}	0.183	0.134	0.183	0.199	0.316	β_{1_RMSE}	0.243	0.441	0.243	0.243	0.382
Scenario 5: $(\sigma_{\bar{b},x}^2, \sigma_{\bar{w},x}^2)=(0.2^2, 0.5^2)$; $\bar{\alpha}_g \sim N(0,0.5^2)$; G=12; $n_c \sim N(150, 50^2)$											
λ_1	0.364	0.108	0.364	0.364	0.301	β_1	0.771	0.788	0.771	0.771	0.785
λ_{1_SE}	0.073	0.082	0.073	0.102	0.275	β_{1_SE}	0.173	0.176	0.173	0.170	0.201
λ_{1_BIAS}	0.164	-0.092	0.164	0.164	0.101	β_{1_BIAS}	-0.029	-0.012	-0.029	-0.029	-0.015
λ_{1_RMSE}	0.180	0.123	0.179	0.193	0.293	β_{1_RMSE}	0.175	0.177	0.175	0.172	0.202
Scenario 6: $(\sigma_{\bar{b},x}^2, \sigma_{\bar{w},x}^2)=(0.5^2, 0.2^2)$; $\bar{\alpha}_g \sim N(0,0.5^2)$; G=36; $n_c \sim N(50, 15^2)$											
λ_1	0.392	0.098	0.392	0.392	0.306	β_1	0.646	0.805	0.646	0.646	0.677
λ_{1_SE}	0.044	0.050	0.044	0.090	0.206	β_{1_SE}	0.108	0.252	0.108	0.124	0.215
λ_{1_BIAS}	0.192	-0.102	0.192	0.192	0.106	β_{1_BIAS}	-0.154	0.005	-0.154	-0.154	-0.123
λ_{1_RMSE}	0.197	0.113	0.197	0.212	0.232	β_{1_RMSE}	0.188	0.252	0.188	0.198	0.248

5.5 Investigation of correlation

Scenarios 1 to 6 are designed for models with no correlation between the explanatory variable \bar{x}_{gt} and the fixed group effect $\bar{\alpha}_g$. The results do not favour the use of the FE estimator when \bar{x}_{gt} has a larger between-group variation. However, it is well known that the pooled OLS and RE estimators are biased and inconsistent if this correlation is present although this is hard to identify in applied research because $\bar{\alpha}_g$ is unobserved. Scenario 7 is designed to be compared with Scenario 3 by adding a correlation coefficient of 0.5 between \bar{x}_{gt} and $\bar{\alpha}_g$. The comparison is conducted for both static and dynamic models and addresses the important question of estimator performance when there is correlation between \bar{x}_{gt} and $\bar{\alpha}_g$.

The simulation results of Scenario 7 for the static model are shown in Table 5.6. Comparing Scenario 7 to Scenario 3, the bias of the pooled OLS, RE, and PCSE estimators increases substantially from around -0.010 to 0.336 or 0.251, whereas the bias of the FE estimator only slightly changes from -0.017 to -0.037, demonstrating the FE estimator's ability to control for the correlation between \bar{x}_{gt} and $\bar{\alpha}_g$. However, the inefficiency of the FE estimator still remains the same in Scenario 7 as in Scenario 3, and this is larger than the other estimators. This

suggests a trade-off between an unbiased but inefficient estimator (i.e., FE) and a biased but efficient estimator (i.e., pooled OLS, RE, PCSE) when a preferred estimator needs to be chosen. If RMSE is used as the measure to determine the best estimator, the RE estimator would be preferred.

Table 5.6 of Scenario 3 and Scenario 7 in the Static Model

Static Model: $\bar{y}_{gt} = 0.2 + 0.8 * \bar{x}_{gt} + \bar{u}_{gt}$

	Pooled			
	OLS	FE	RE	PCSE
Scenario 3: $(\sigma_{B,x}^2, \sigma_{W,x}^2)=(0.5^2, 0.2^2); \bar{\alpha}_g \sim N(0,0.5^2)$				
β_1	0.791	0.783	0.789	0.791
β_1 _SE	0.181	0.419	0.265	0.162
β_1 _BIAS	-0.009	-0.017	-0.011	-0.009
β_1 _RMSE	0.181	0.420	0.266	0.162
Scenario 7: $(\sigma_{B,x}^2, \sigma_{W,x}^2)=(0.5^2, 0.2^2); \bar{\alpha}_g \sim N(0,0.5^2); \text{corr}(\bar{x}_{gt}, \bar{\alpha}_g)=0.5$				
β_1	1.136	0.763	1.051	1.136
β_1 _SE	0.161	0.419	0.235	0.145
β_1 _BIAS	0.336	-0.037	0.251	0.336
β_1 _RMSE	0.373	0.421	0.344	0.366

Table 5.7 summarises a comparison of Scenario 3 and Scenario 7 in the dynamic model estimation. The introduction of correlation between \bar{x}_{gt} and $\bar{\alpha}_g$ does not change the estimation results of λ_1 noticeably, with standard errors for all estimators remaining almost the same. Although the absolute bias of β_1 estimates does not change between Scenario 3 and Scenario 7, the sign is reversed for all estimators but with standard errors remaining about the same. As a result, the RMSE of all estimators between the two scenarios do not vary substantially. The FE and GMM estimators are still not favoured given their large combined RMSE of λ_1 and β_1 , despite the presence of the correlation coefficient of 0.5 between \bar{x}_{gt} and $\bar{\alpha}_g$.

Table 5.7 Comparisons of Scenario 3 and Scenario 7 in the Dynamic Model

Dynamic Model: $\bar{y}_{gt} = 0.2 * \bar{y}_{gt-1} + 0.8 * \bar{x}_{gt} + \bar{u}_{gt}$											
	Pooled						Pooled				
	OLS	FE	RE	PCSE	GMM		OLS	FE	RE	PCSE	GMM
Scenario 3: $(\sigma_{\bar{b},x}^2, \sigma_{\bar{w},x}^2)=(0.5^2, 0.2^2)$; $\bar{\alpha}_g \sim N(0,0.5^2)$											
λ_1	0.366	0.098	0.365	0.366	0.319	β_1	0.659	0.816	0.660	0.659	0.688
λ_1 -SE	0.077	0.087	0.077	0.109	0.293	β_1 -SE	0.198	0.440	0.198	0.197	0.365
λ_1 -BIAS	0.166	-0.102	0.165	0.166	0.119	β_1 -BIAS	-0.141	0.016	-0.140	-0.141	-0.112
λ_1 -RMSE	0.183	0.134	0.183	0.199	0.316	β_1 -RMSE	0.243	0.441	0.243	0.243	0.382
Scenario 7: $(\sigma_{\bar{b},x}^2, \sigma_{\bar{w},x}^2)=(0.5^2, 0.2^2)$; $\bar{\alpha}_g \sim N(0,0.5^2)$; $\text{corr}(\bar{x}_{gt}, \bar{\alpha}_g)=0.5$											
λ_1	0.346	0.099	0.346	0.346	0.285	β_1	0.942	0.784	0.942	0.942	0.987
λ_1 -SE	0.077	0.087	0.077	0.106	0.294	β_1 -SE	0.194	0.439	0.194	0.204	0.431
λ_1 -BIAS	0.146	-0.101	0.146	0.146	0.085	β_1 -BIAS	0.142	-0.016	0.142	0.142	0.187
λ_1 -RMSE	0.165	0.133	0.165	0.181	0.306	β_1 -RMSE	0.240	0.439	0.240	0.249	0.470

5.6 Summary

This chapter examines the estimation performance of pooled OLS, FE, RE, PCSE, and GMM estimators, using Monte Carlo simulation experiments based on the typical properties of pseudo panel data, and where appropriate, the calibration of the SHTS pseudo panel dataset used in this study. These experimental results illustrate the importance of measuring bias as well as efficiency when comparing various estimators. As emphasised in Plümper and Troeger (2011), bias represents an expected deviation from the true value of the coefficient in finite sample econometrics, and both bias and inefficiency increase the probability that a point estimate differs from its true value in applied research. It is possible that an optimal estimator is not available, and in this case the RMSE that equally weighs bias and variance as suggested in the literature can be adopted to determine a second best estimator which potentially generates estimates closest to the true parameter value.

The static model simulation results suggest that the variance of unobserved group effects does not lead to substantial bias on the exogenous variable and the FE estimator performs slightly better than other estimators under this circumstance. However, when the exogenous variable has a larger between-group variance than the within-group variance, the PCSE estimator is the preferred estimator after taking account of both bias and efficiency.

For dynamic models, there is likewise no unambiguously superior estimator when the exogenous variable has a larger between-group variation. The FE estimator appears to be the least biased estimator but with the largest standard errors, whereas the pooled OLS, RE, and PCSE estimators are more efficient but with larger biases, even when the correlation between the explanatory variable and unobserved group effect is present at the correlation coefficient of 0.5. However, the bias of the pooled OLS, RE, and PCSE estimators can be decreased by reducing the variance of unobserved group effects. This suggests that the pooled OLS, RE and PCSE estimators are potentially unbiased and efficient if the variance of unobserved group effects can be minimised, possibly through better model specifications.

The trade-off between cohort size n_c and total number of groups (G) in pseudo panel data construction is also investigated. The findings indicate that using a dataset with a smaller n_c but a larger G effectively improves the estimation efficiency, with just a slight increase in bias. This result justifies the selection between the two pseudo panel datasets constructed in Section 4.3 where the dataset with a larger number of groups ($G = 20$) is chosen over the other dataset with a smaller number of groups ($G = 12$).

This data generating process of the simulation model presented in this chapter is based on one single exogenous variable because of the technical difficulties of extending to multiple explanatory variables. Its simulation results may not perfectly transfer to empirical panel data models with multiple exogenous variables, since the correlation and interaction of the multiple exogenous variables may cause a certain degree of estimation bias or inefficiency. However, the simulation analysis in this chapter identifies the causes of estimation bias and inefficiency by controlling the data conditions in various scenarios. The results still provide important information for guiding the empirical pseudo panel data estimation in the following chapters.

Rather than proposing a best estimator for pseudo panel data, this exercise highlights the necessity of understanding the nature and properties of pseudo

panel data before deciding which estimator to use in empirical applications. The findings of this experiment provide a reference point for empirical pseudo panel data estimation, which are presented in Chapter 6 and Chapter 7.

CHAPTER 6 STATIC MODEL ESTIMATION

6.1 Introduction

This chapter presents the relationship between public transport demand and its determinants whilst not taking the temporal effect of demand changes into account in the form of a static model. It presents a static pseudo panel data model to analyse public transport demand in the Sydney Greater Metropolitan Area (SGMA) with respect to its determinants. Section 6.2 first discusses the various functional forms of regression models with their associated economic theory and implications before defining the static public transport demand model for this study. The dataset used to estimate the static model is summarised in this section together with the descriptive statistics of the variables. The estimation techniques are discussed building on the simulation findings presented in Chapter 5. Section 6.3 presents the estimation results for the static model and compares parameters estimated from various functional forms and estimators, together with an analysis of the model diagnostics. A summary of research findings from the static pseudo panel data model is discussed in Section 6.4.

6.2 Static public transport demand model

6.2.1 *Functional forms*

The general form of the pseudo panel data model in this study is introduced in Section 4.5. In econometric analysis, there are a number of alternative functional forms, each of which assumes a certain relationship between the dependent variable and the explanatory variables. Evaluating these various functional forms is important because a misspecification of functional form may result in biased or inconsistent estimated parameters. This section discusses the economic theory underpinning the various functional forms and model specification tests for evaluating model functional forms when estimated.

The most basic functional form is a linear regression model which assumes a linear relationship between the dependent variable (Y) and explanatory variables (X) as defined in Equation (6.1).

$$Y = \beta_1 + \beta_2 X + \varepsilon \quad \text{Equation (6.1)}$$

The coefficient β_2 represents the impact of a unit change in X on Y . The elasticity derived from this linear model is defined by Equation (6.2). The elasticity from a linear model is not constant but will vary with the values of X and Y , with a larger ratio of X to Y giving a larger elasticity.

$$e = \frac{dY}{dX} \cdot \frac{X}{Y} \quad \text{Equation (6.2)}$$

Another functional form is the double-logarithmic (abbreviated as double-log) model which is derived from a non-linear model as in Equation (6.3).

$$Y = KX^{\beta_2} \quad \text{Equation (6.3)}$$

The double-log model can be transformed from Equation (6.4) by taking logs on both sides of the equation:

$$\ln Y = \beta_1 + \beta_2 \ln X, \quad \text{where } \beta_1 = \ln K \quad \text{Equation (6.4)}$$

The double-log model is commonly employed because of its convenience in interpreting the relationship between Y and X . The coefficient β_2 represents the impact of percentage changes in X on percentage changes in Y . The elasticity derived from the model is constant as shown in Equation (6.5), and is represented by β_2 . In the context of this study this would imply that the elasticity is the same across all individuals regardless the values of Y or X .

$$e = \frac{dY}{dX} \cdot \frac{X}{Y} = K \beta_2 X^{\beta_2-1} \cdot \frac{X}{KX^{\beta_2}} = \beta_2 \quad \text{Equation (6.5)}$$

The other non-linear functional form is an exponential function (Equation (6.6)) which yields the log-linear function shown in Equation (6.7) by taking logs on both sides of the equation.

$$Y = Ke^{X\beta_2} \quad \text{Equation (6.6)}$$

$$\ln Y = \beta_1 + \beta_2 X, \quad \beta_1 = \ln K \quad \text{Equation (6.7)}$$

In the log-linear function, β_2 represents the effect of a unit change in X on percentage changes in Y , with the elasticity defined by Equation (6.8). The elasticity from the log-linear function is not constant but varies with X .

$$e = \frac{dY}{dX} \cdot \frac{X}{Y} = K \beta_2 e^{X\beta_2} \cdot \frac{X}{Ke^{X\beta_2}} = \beta_2 X \quad \text{Equation (6.8)}$$

The other alternative of the exponential function is Equation (6.9):

$$Y = Ke^{X\beta_2} \quad \text{Equation (6.9)}$$

$$Y = \beta_1 + \beta_2 \ln X, \quad \beta_1 = \ln K \quad \text{Equation (6.10)}$$

Equation (6.10) is a linear-log function derived from Equation (6.9), in which β_2 measures the effect of percentage changes in X on unit changes in Y . The elasticity as defined by Equation (6.11) is not constant but varies with the value of Y .

$$e = \frac{dY}{dX} \cdot \frac{X}{Y} = \frac{\beta_2}{X} \cdot \frac{X}{Y} = \beta_2 \frac{1}{Y} \quad \text{Equation (6.11)}$$

The preferred model among these various functional forms can be evaluated by comparing the model goodness of fit in terms of adjusted R-squared values amongst the models that have the same dependent variable. However, models with different dependent variables, for example, Y and $\ln Y$, cannot be compared at the basis of R-squared values. In this case, the Regression Specification Error Test (RESET) developed by Ramsey (1969) can be employed to evaluate the functional forms based on their relative specification errors.

The RESET model specification test is derived from the following regression model (Equation (6.12)):

$$Y_i = \beta_1 + \beta_2 X_i + \gamma_1 \hat{Y}_i^2 + \gamma_2 \hat{Y}_i^3 + v_i \quad \text{Equation (6.12)}$$

where \hat{Y}_i^2 and \hat{Y}_i^3 are the predicted values of dependent variables for higher-order models and they are used as test variables to examine the significance of the explanatory variable (X_i) in higher orders and their cross-products. If the joint effect of γ_1 and γ_2 is significant, then the assumption that the relationship between Y_i and X_i is linear in the parameters is violated because of potential omitted variables. The RESET test can be used as a general test for model misspecification with the null hypothesis being there are no omitted variables existing in the regression model.

The static public transport demand model presented in this chapter does not presume a preferred functional form. Instead, this analysis examines the four functional forms introduced above by comparing their model goodness-of-fit and their RESET test results. The linear model is estimated as a base model with the estimation results being compared to the Geographically Weighted Regression (GWR) global model in Chapter 3 which is also a static linear model but with a different modelling approach.

The static linear pseudo panel data model expanded from the general model (Equation (4.3)) is specified as Equation (6.13).

$$\begin{aligned} \bar{D}_{g,t} = & \beta_0 + \beta_1 \overline{PRICE}_{g,t} + \beta_2 \overline{INCOME}_{g,t} + \beta_3 \overline{AGE}_{g,t} + \beta_4 \overline{FREQ}_{g,t} \\ & + \beta_5 \overline{DENSITY}_{g,t} + \beta_6 \overline{LANDMIX}_{g,t} + \beta_7 \overline{PSEUDO}_{g,t} + \beta_8 \overline{DISTANCE}_{g,t} \quad \text{Equation (6.13)} \\ & + \beta_9 \overline{STOPS}_{g,t} + \bar{u}_{g,t}, \quad \bar{u}_{g,t} = \bar{\alpha}_{g,t} + \bar{\epsilon}_{g,t} \end{aligned}$$

where public transport demand ($\bar{D}_{g,t}$) of a cohort in a group g at time t is predicted by a set of explanatory variables as reviewed in Section 2.1.1, which includes: the average public transport trip price ($\overline{PRICE}_{g,t}$), socio-economic factors including average personal income ($\overline{INCOME}_{g,t}$) and average age ($\overline{AGE}_{g,t}$),

average bus frequency ($\overline{FREQ}_{g,t}$) as a measure of public transport supply, land use factors including average population density ($\overline{DENSITY}_{g,t}$), average entropy of land use mix ($\overline{LANDMIX}_{g,t}$), average number of pseudo nodes ($\overline{PSEUDO}_{g,t}$), average walk distance to the nearest public transport stop ($\overline{DISTANCE}_{g,t}$), and average number of public transport stops within 800 meters of a traveller's household location ($\overline{STOPS}_{g,t}$), with a composite error term ($\bar{u}_{g,t}$) comprising the unobserved time-varying group effects ($\bar{\alpha}_{g,t}$) and an *i.i.d.* error term ($\bar{\epsilon}_{g,t}$).

The static linear public transport demand model in Equation (6.13) has the same dependent variable and explanatory variables as the GWR global model (Equation (3.2)) except for the variable of distance to CBD. Distance to CBD is included in the GWR model to investigate the relationship between public transport demand and its potential predictors. After the strong relationship between public transport demand and distance to CBD is identified in the GWR analysis, this variable is used as one of the grouping criteria for the pseudo panel data construction as presented in Chapter 4. Therefore, distance to CBD is not included in the pseudo panel data model because its influence on public transport demand variation has been captured by the way in which the groups are created for the pseudo panel dataset. Apart from the variable of distance to CBD, the other explanatory variables in the GWR model of Chapter 3 are included in this static pseudo panel data model with the same units, using the mean values of the variables in the pseudo panel data model as computed for each of the cohorts.

6.2.2 Descriptive statistics

The descriptive statistics of the pseudo panel dataset are displayed in Table 6.1. As discussed in Section 4.4.1, the variation of each variable can be observed either as a between-group standard deviation or as a within-group standard deviation as well as an overall standard deviation. The between-group standard deviations represent the variations across the 20 groups in the pseudo panel dataset which can be interpreted as cross-sectional variations without taking time-varying changes into consideration. In contrast, within-group standard deviations are the average variation within each defined group over time, which can also be referred as time-series variation.

Table 6.1 Descriptive Statistics of Variables

Variable	Unit		Mean	S.D.	Min	Max
PTTRIP	Trips	overall	0.45	0.28	0.08	1.63
		between		0.26	0.12	1.02
		within		0.11	0.12	1.06
PRICE	Dollars (AUD)	overall	1.73	0.59	0.39	2.88
		between		0.56	0.64	2.57
		within		0.21	1.09	2.35
INCOME	Thousand dollars (AUD)	overall	28.64	12.98	2.08	58.38
		between		11.64	13.48	46.32
		within		6.37	10.51	47.78
AGE	Years	overall	41.32	17.64	18.00	75.65
		between		17.80	20.25	70.10
		within		3.33	34.35	48.24
BUS FREQUENCY	Thousands	overall	0.19	0.15	0.02	0.77
		between		0.14	0.05	0.52
		within		0.05	-0.04	0.44
POPULATION DENSITY	Thousands	overall	22.08	5.59	11.45	33.15
		between		5.50	13.76	30.61
		within		1.54	17.21	26.63
LAND MIX	Entropy	overall	0.13	0.01	0.09	0.17
		between		0.01	0.11	0.14
		within		0.01	0.09	0.17
PSEUDO NODES	Thousands	overall	1.36	0.62	0.76	4.14
		between		0.59	0.90	2.46
		within		0.23	0.71	3.04
DISTANCE TO PT STOP	Kilometre	overall	0.24	0.08	0.12	0.59
		between		0.05	0.19	0.35
		within		0.06	0.12	0.52
PT STOPS	Stops	overall	41.45	7.58	25.60	60.77
		between		6.91	29.50	50.27
		within		3.44	27.24	51.95

Most variables in the dataset have substantially higher between-group standard deviations than within-group standard deviations as a result of the pseudo panel data construction process which aimed to produce sufficient inter-group variations. It is important to note that these time-series variations not only come from the time-varying changes in variables, but also from the composition of cohort members who are different individuals over time although within the same group. Hence, some variables that are collected at a single point of time, such as land use variables, still display a certain degree of within-group variations which comes from the different composition of cohort members. Although it would be ideal to have all variables with true historical values, most

land use variables do not show sufficient time-varying variations at the cohort level, as shown by the sensitivity analysis presented in Section 4.3.2. However, estimating the within-group variations of these time-invariant variables in the pseudo panel dataset can also provide information for long-term transport planning and policies by capturing the current cross-sectional variations in these variables and the elasticity of public transport demand with respect to changes in these rarely-changing over time variables.

As the variables in the pseudo panel dataset are the mean values of the individuals in each cohort, this level of aggregation mitigates a certain degree of measurement error from extreme values in the individual data. This is shown by the overall standard deviations being relatively smaller than the anticipated mean for most variables.

According to the mean values of variables, public transport demand in terms of the average number of public transport trips per person per day is low at 0.45 trips per day, but this is expected given the low overall usage of public transport in the Sydney Greater Metropolitan Area (SGMA) as discussed in Section 3.2.2 which shows that the overall mode share of train and bus trips is around ten percent in the SGMA.

Average trip price and average person income appear to be lower than expectation with an average public trip price of 1.73 dollars and an average annual person income of 28.64 thousand Australian dollars. This is partly because these two variables are adjusted to real terms based on 1997 CPI, and partly because there is a considerable number of students and pensioners with lower incomes who are entitled to free school buses and concession tickets and thus lower the average income and age of public transport users in the dataset. However, these generation and life-cycle effects are captured by the way in which the individuals are allocated to groups according to their birth years.

The entropy of land use mix also has a low average value at 0.13, with a maximum value of 0.17 in which suggests a very homogenous land use mix with small variation across Travel Zones (TZs). As discussed in Section 3.4.2, this is

due to the fine aggregation level of this measurement, which does not have much variation within the TZs, and for which this effect becomes more marked when the data are aggregated at the cohort level giving a more homogenous land use mix in the pseudo panel dataset.

The correlation matrix of all variables in the pseudo panel dataset is shown in Table 6.2. Similar to the GWR dataset in Table 3.7, high correlations occur among bus frequency, population density, pseudo nodes, and distance to the nearest public transport stop. The correlation coefficients appear to be generally higher than they are in the GWR dataset in Table 3.7, as a result of fewer observations (256 cohorts) in the pseudo panel data model as opposed to 1,824 observations (TZs) in the GWR model. Given the potential for multi-collinearity, the model estimations presented in the following sections also test the magnitude of collinearity based on the Variance Inflation Factor (VIF) of parameters.

Table 6.2 Correlation Matrix of Variables in the Pseudo Panel Dataset

	PTTRIP	PRICE	INCOME	AGE	BUS FREQ ¹	DENSITY ²	LAND MIX	PSEUDO ³	DISTANCE ⁴	PT STOPS
PTTRIP	1									
PRICE	-0.26	1								
INCOME	-0.35	0.53	1							
AGE	-0.41	-0.60	-0.01	1						
BUS FREQ ¹	0.65	-0.26	0.06	-0.10	1					
DENSITY ²	0.66	-0.30	0.12	-0.08	0.82	1				
LAND MIX	-0.29	0.23	0.09	0.02	-0.21	-0.36	1			
PSEUDO ³	-0.56	0.40	-0.07	-0.06	-0.59	-0.78	0.36	1		
DISTANCE ⁴	-0.38	0.28	-0.10	-0.06	-0.47	-0.61	0.12	0.64	1	
PT STOPS	0.60	-0.23	0.08	-0.11	0.68	0.83	-0.34	-0.81	-0.64	1

¹BUS FREQUENCY

²POPULATION DENSITY

³PSEUDO NODES

⁴DISTANCE TO PT STOP

6.2.3 Estimation techniques

The estimation techniques for pseudo panel data models are reviewed in Section 2.4 and examined in Chapter 5 through Monte Carlo simulation experiments. The simulation results for static pseudo panel data models in Section 5.4.1 suggest that there is no one superior estimator among the pooled Ordinary Least Squares (OLS), Fixed Effect (FE), Random Effect (RE), and Panel-Corrected Standard Error (PCSE) estimators when explanatory variables are equally

distributed across groups and across time, that is, between-group standard deviations are the same as within-group standard deviations. However, when explanatory variables have substantially larger between-group standard deviations than within-group standard deviations, the FE estimator will become inefficient and cause inflation of the standard errors of parameters. Under this circumstance, the pooled OLS and PCSE estimators outperform to other estimators in terms of RMSE, with PCSE giving the lowest RMSE when heteroscedasticity is present in the model.

The static pseudo panel data model in the following section is first estimated by the pooled OLS estimator. The omitted variable bias and heteroscedasticity are also tested through RESET test and Breusch-Pagan Test respectively. If heteroscedasticity is present, the pooled OLS estimation will need to be corrected by using the PCSE estimator which allows for serial correlation or cross-sectional dependency in the error terms. If there is no evidence of non-spherical errors, the pooled OLS estimator and the PCSE estimator will give the exactly same estimated coefficients, with slight differences in standard errors. A comparison of estimation results using the FE, RE, and PCSE estimators are demonstrated in Section 6.3.4 to validate the choice of estimator corresponding to findings from the simulation results of Section 5.4.1.

6.3 Estimation results

6.3.1 Base model

The estimation results of the static linear public transport demand model estimated by the pooled OLS estimator are shown in Table 6.3. This model is estimated as a base model to be compared to other functional forms. In terms of general estimation performance, the model goodness-of-fit is fairly good given the adjusted R-squared of 0.782 which suggests that 78.2 percent of variation in public transport demand can be explained by the explanatory variables. The F-test suggests that the joint relationship between dependent variable and predictors is significant (P-value=0).

Table 6.3 Pooled OLS Estimation Results of Static Linear Model (Base Model)

Dependent Variable: PTTRIP	Coef.	Std. Err.	T	P-value	[95% C.I.]		VIF
PRICE	-0.065	0.033	-1.98	0.049	-0.130	0.000	5.68
INCOME	-0.007	0.001	-6.82	0.000	-0.009	-0.005	2.84
AGE	-0.007	0.001	-9.18	0.000	-0.009	-0.006	3.01
BUS FREQUENCY	0.507	0.099	5.13	0.000	0.313	0.702	3.34
POPULATION DENSITY	0.012	0.004	3.11	0.002	0.004	0.019	6.97
LAND MIX	0.024	0.671	0.04	0.972	-1.298	1.346	1.29
PSEUDO NODES	-0.101	0.026	-3.82	0.000	-0.153	-0.049	4.13
DISTANCE TO PT STOP	0.009	0.151	0.06	0.953	-0.289	0.307	1.97
PT STOPS	-0.001	0.002	-0.46	0.645	-0.006	0.003	4.63
CONSTANT	0.888	0.176	5.04	0.000	0.541	1.236	
Observations	256						
F(9, 246)	102.08						
P-value	0						
R-squared	0.789						
Adj. R-squared	0.782						
Root MSE	0.129						
Ramsey RESET Test (Ho: Model has no omitted variables)							
F(3, 243)	16.46						
P-value	0.000						
Breusch-Pagan Test for heteroscedasticity (Ho: Constant variance)							
chi2(1)	55.22						
Prob >Chi ²	0.000						
Wooldridge test for autocorrelation (Ho: No first order autocorrelation)							
F(1, 19)	0.581						
Prob > F	0.455						

In terms of the significance of individual explanatory variables, most variables are significant at 95 percent statistical confidence level with the expected sign. Price has a significantly negative impact on public transport demand which conforms to the economic theory of a negative relationship between demand and price. The socio-economic measures of personal income and age, are negatively significant to public transport demand changes, indicating that public transport demand is expected to decrease with higher income and age. The public transport supply measure, bus frequency, has a positive sign as expected, which confirms that higher bus frequency is expected to increase public transport demand.

The relationships between public transport demand and land use variables mostly confirm the hypotheses of this study, apart from land use mix and accessibility measures are inconclusive with their coefficients insignificantly

different from zero. The estimation results of this static linear model support the findings from the linear GWR global model as shown in Table 3.8 in terms of the relationships between public transport demand and its explanatory variables. As discussed in Section 3.4.2, the insignificance of land use mix and accessibility measures are likely to be due to the aggregation level of land use categories and the inclusion of the public transport supply measure which explains the variation in public transport demand better than distance to the nearest public transport stop and number of bus stops.

In general, the static pseudo panel data model performs better than the GWR global model in terms of model goodness-of-fit and the prediction power of the explanatory variables, even when the number of observations (256 cohorts) is considerably lower than the observations (1,824 TZs) in the GWR global model. This suggests that the pseudo panel dataset, constructed in a way to increase inter-group heterogeneity, generates more variation in public transport demand with respect to the explanatory variables and offers a better way of explaining public transport demand.

Although the general performance of the static linear model appears to be quite good, it is essential to diagnose potential multi-collinearity and omitted variable bias, as well as heteroscedasticity in error terms which may lead to problematic estimation results. The multi-collinearity is tested by the (VIF) of each variable. As a rule of thumb, a VIF larger than ten indicates strong a multi-collinearity which gives unacceptable inflation of standard errors, and a VIF between five and ten suggests a certain degree of multi-collinearity which may inflate the standard errors slightly but it will not alter the coefficient estimates. The VIF shown in Table 6.3 suggest that most variables have a VIF lower than five, except for the price and population density which are moderately impacted by multi-collinearity at VIF of 5.68 and 6.97 respectively. Despite these higher VIF values, price and population density are both significant at 95 percent confidence level, so the standard errors are not so over-inflated to change their significance. The insignificance of land use mix, distance to public transport stops, and number of public transport stops, are not likely to be a result of multi-collinearity because their VIF indicators are lower than five.

Potential omitted variable bias is tested by the Ramsey's RESET test, with results showing a null hypothesis of no omitted variables is rejected. This implies that although the relationship between the dependent variable and explanatory variables is significant, there may be potential omitted variable bias so the coefficients cannot be interpreted as the true magnitudes of explanatory variables' impacts on dependent variables. Therefore, other model forms should be further investigated to justify a preferred model where no omitted variable bias is identified.

The Breusch-Pagan Test for heteroscedasticity in error terms is also reported in Table 6.3, with results showing that the assumption of homogenous error terms is rejected, although serial correlation is not significant according to the Wooldridge test. The presence of heteroscedasticity is likely to alter the standard errors of estimated parameters. Therefore, not only the model functional form but also the estimation techniques should be further explored for the static pseudo panel data model estimation.

6.3.2 Test of functional forms

The static models with various functional forms (as defined in Section 6.2.1) are evaluated according to their model fits and RESET test results as presented in Table 6.4. The model fit, in terms of adjusted R-squared, can only be compared for models with the same dependent variable. Hence, the linear model and the linear-log model, as well as the double-log model and the log-linear model are compared as two pairs where each pair has the same dependent variable. Although the differences are not substantial, the linear-log model and the double-log model are the two preferred models given their higher adjusted R-squared values than the linear model and the log-linear model respectively. It is also important to note that the significance of the explanatory variables across all the models is very similar, apart from the price and number of public transport stops which show different results in the linear-log model.

Table 6.4 Comparison of Static Model Functional Forms

	LINEAR	LINEAR LOG	DOUBLE LOG	LOG LINEAR
PRICE	-0.065** (0.033)	-0.005 (0.038)	-0.269*** (0.0709)	-0.243*** (0.0639)
INCOME	-0.007*** (0.001)	-0.210*** (0.031)	-0.221*** (0.0572)	-0.012*** (0.002)
AGE	-0.007*** (0.001)	-0.260*** (0.035)	-0.750*** (0.0656)	-0.017*** (0.002)
BUS FREQUENCY	0.507*** (0.099)	0.139*** (0.025)	0.206*** (0.0462)	0.438** (0.192)
POPULATION DENSITY	0.012*** (0.004)	0.286*** (0.077)	0.753*** (0.142)	0.042*** (0.007)
LAND MIX	0.024 (0.671)	0.071 (0.064)	0.000963 (0.120)	-0.589 (1.302)
PSEUDO NODES	-0.101*** (0.026)	-0.189*** (0.053)	-0.570*** (0.0986)	-0.276*** (0.051)
DISTANCE TO PT STOP	0.009 (0.151)	0.018 (0.034)	0.0610 (0.0634)	0.386 (0.293)
PT STOPS	-0.001 (0.002)	-0.151* (0.079)	-0.200 (0.146)	0.004 (0.004)
CONSTANT	0.888*** (0.176)	1.943** (0.858)	0.0227 (1.595)	-0.314 (0.342)
Observations	256	256	256	256
R-squared	0.789	0.801	0.872	0.851
Adjusted R-squared	0.781	0.794	0.867	0.846
Ramsey RESET Test				
Prob > F	0.000	0.000	0.007	0.000
Breusch-Pagan Test				
Prob >Chi ²	0.000	0.000	0.044	0.307

Note: Standard errors in parentheses; * P<0.10, ** P<0.05, *** P<0.01;
Models are estimated by OLS.

The comparison between the linear-log model and the double-log model is evaluated by the Ramsey's RESET test. The result suggests that both the linear-log model and the double-log model reject the null hypothesis of no omitted variable bias. However, the double-log model has a slightly lower F-statistic and thus the probability of rejecting the null hypothesis is slightly higher than all other models, so the double-log model is considered as the preferred functional form for the static public transport demand model although the omitted variable bias is present in each model. The omitted variable bias appears to exist in the static models possibly because the static models do not take account of the

temporal effects of demand adjustment, that is, the lagged dependent variable which captures travellers' lagged adjustments of travel behaviour (as reviewed in Section 2.1.2). As a result, although the goodness-of-fit of the double-log model is considered good given an adjusted R-squared of 0.867, the static model still has potential omitted variable bias that requires further investigation, and the parameter estimates should not be interpreted for economic and policy implications. These issues are further explored in Chapter 7.

6.3.3 Model Diagnostics

In a multiple regression model analysis, it is essential to test the normality of error term distribution. The Breusch-Pagan Test for heteroscedasticity has been conducted for the double-log model with the result suggesting that it rejects the null hypothesis of constant variance in error terms at 95 percent confidence level, but the p-value of 0.044 implies that the degree of heteroscedasticity may not be substantial.

The heteroscedasticity can be further investigated by residual plots of the regression model as shown in Figure 6.1 to Figure 6.7. The scatter plot of residuals and fitted values (i.e., predicted values of the dependent variable) in Figure 6.1 can be used to detect linearity, heteroscedasticity, and outliers from the regression model. The distribution of residuals in Figure 6.1 does not show noticeable patterns and it appears to be fairly random. There are several residuals slightly greater than 0.5 or smaller than -0.5. Those data points are mostly cohorts constituted of less than 50 members. Although this may suggest some degree of measurement errors for those cohorts, they do not strongly distort the distribution. Therefore, the regression analysis does not drop these observations in order to improve the estimation efficiency.

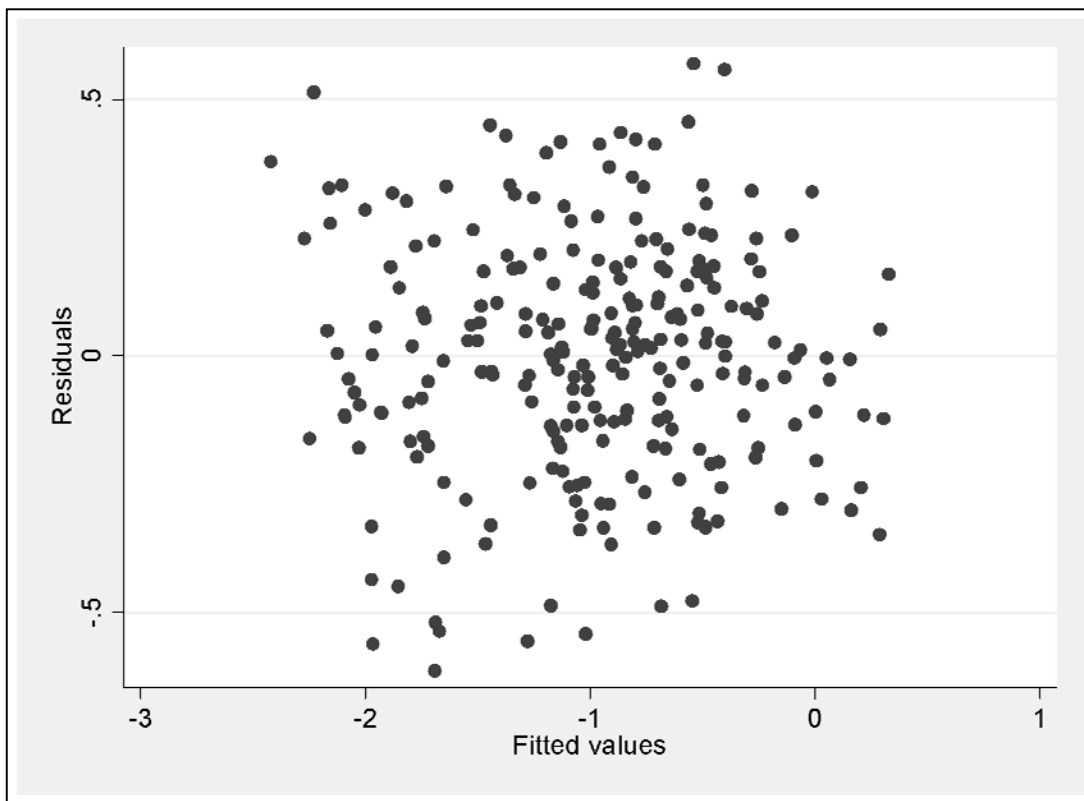


Figure 6.1 Scatter plot of Residuals and Fitted Values

The residuals can also be plotted against each of the explanatory variables to evaluate the performance of a predictor on the dependent variable. As with the residual plots versus fitted values, the distribution of the data points is expected to be random with no distinctive patterns. In Figure 6.2 which shows the scatter plot of residuals and the price variable, it is noticeable that there are two moderately different populations of the data points. The data points located on the left hand side of the plot have relatively smaller trip price in natural logarithms, and there is a gap between the two populations of data points. This is because the data points on the left hand side are from cohorts with an average age older than 65 years, and people at this age are mostly eligible to a pensioner excursion tickets which allows for unlimited travel using a ticket with a face value of 2.5 dollars and results in a smaller average price per trip. This effect is captured by the age variable included in the pseudo panel data model which reduces its influence on the overall estimation result.

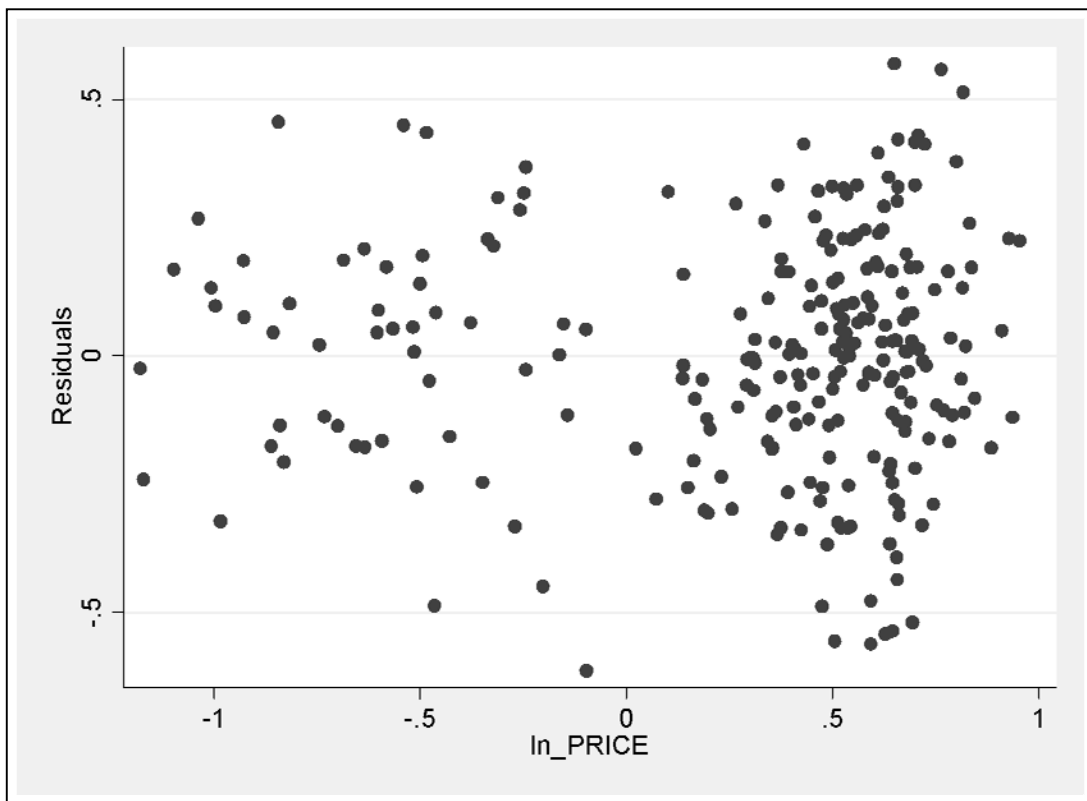


Figure 6.2 Scatter Plot of Residuals and Price

The scatter plot of residuals against income in Figure 6.3 shows that most of the data points are randomly distributed, apart from a potential outlier with the lowest income in the natural log term. This data point is the cohort at an average age of 18 years so the low average income is expected. This effect is also noticeable to some data points with lower income on the left hand side of the graph, and they are all identified as cohorts of younger generations.

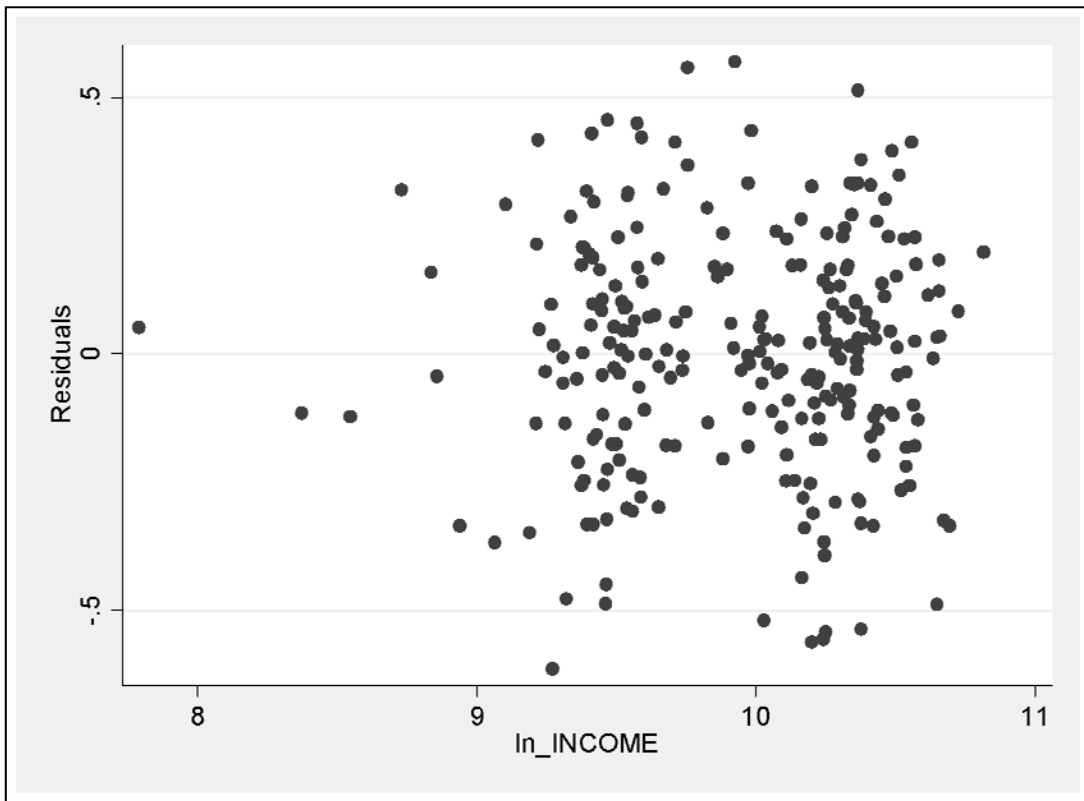


Figure 6.3 Scatter Plot of Residuals and Income

The residual plots of age and bus frequency in Figure 6.4 and Figure 6.5 do not show distinctive patterns. The only noticeable pattern is in Figure 6.4 in which there is a small gap between age 4 and 4.2 in natural logarithms. Their corresponding ages in real terms are 55 and 65 approximately. As discussed in Figure 4.4, there is an age gap of public transport users in the SGMA between 55 and 65 years as a result of the pensioner ticket discount for people above 65 years old. Apart from this, the residuals of age and bus frequency do not show distinctive distributions.

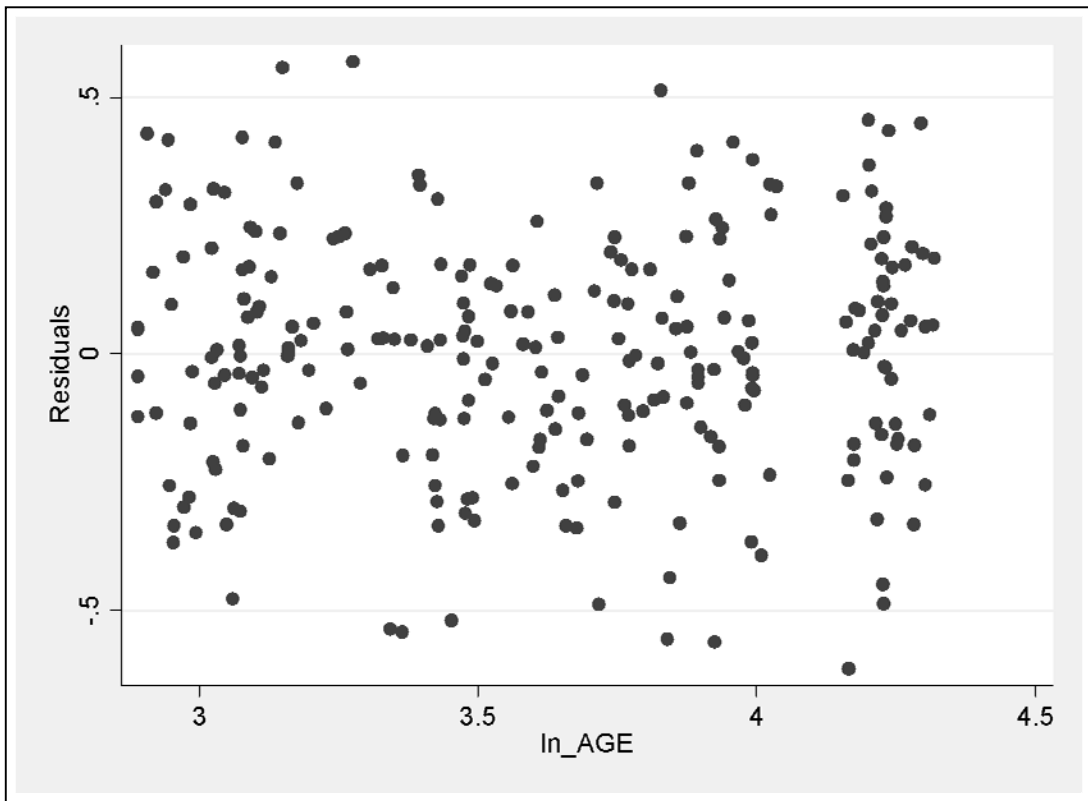


Figure 6.4 Scatter plot of Residuals and Age

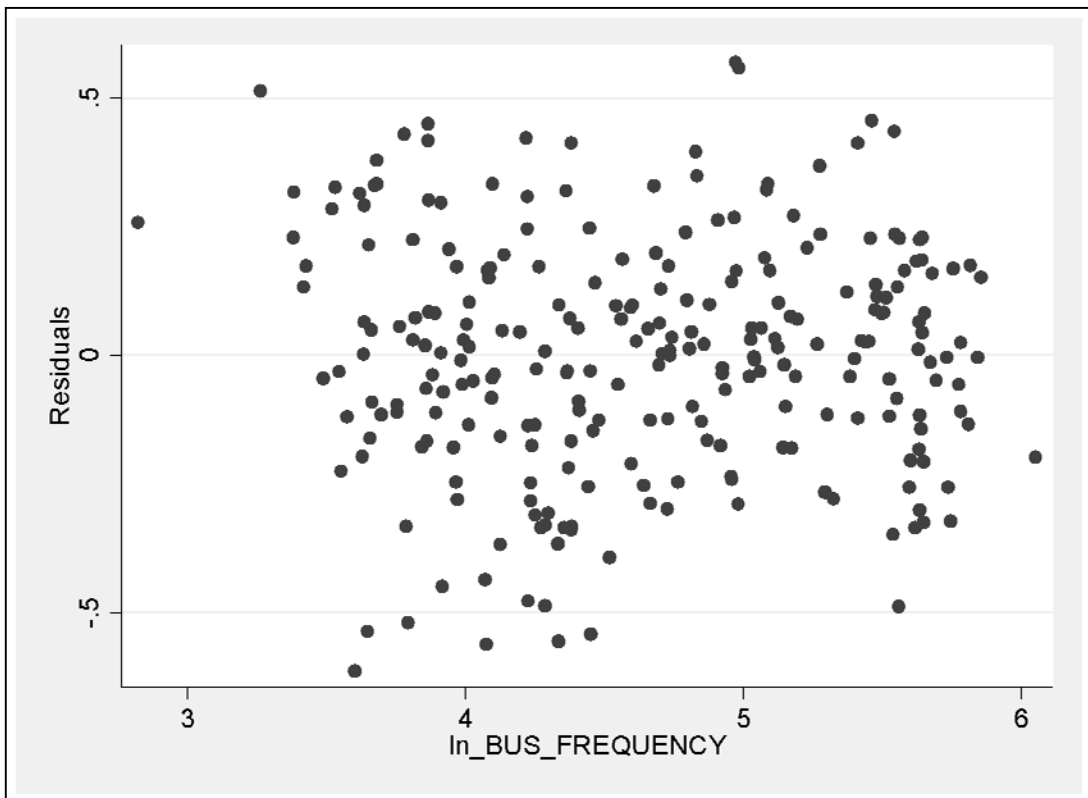


Figure 6.5 Scatter Plot of Residuals and Bus Frequency

Figure 6.6 and Figure 6.7 show the residual plots of population density and pseudo nodes respectively. Both plots show two noticeable distributions, observed from one distribution of lower population density and one distribution of higher pseudo nodes in each of the plots. Those data points with lower density and higher number of pseudo nodes are cohorts located in Zone 4 of the pseudo panel dataset, which is the furthestmost area from the Sydney CBD. Therefore, these areas have substantially lower population density and more pseudo nodes as compared to other areas closer to the CBD, and these two distributions demonstrate an inverse relationship between population density and pseudo nodes as a result of their high correlation with a negative sign (correlation coefficient: -0.78).

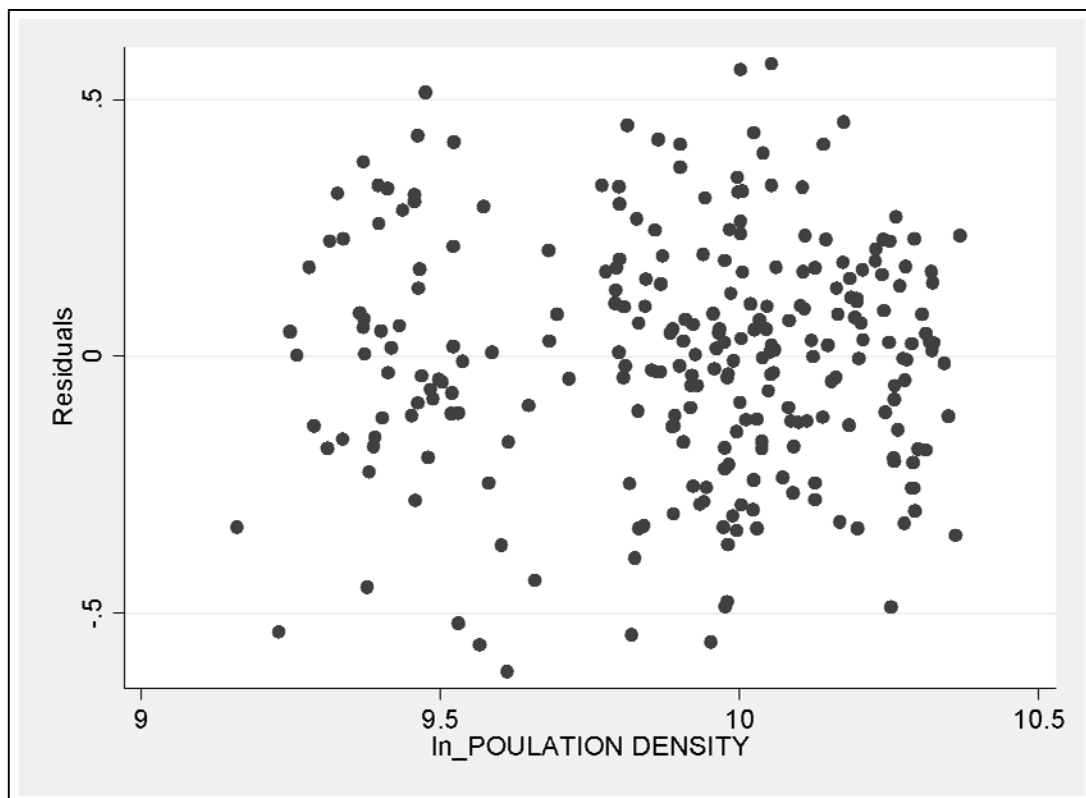


Figure 6.6 Scatter Plot of Residuals and Population Density

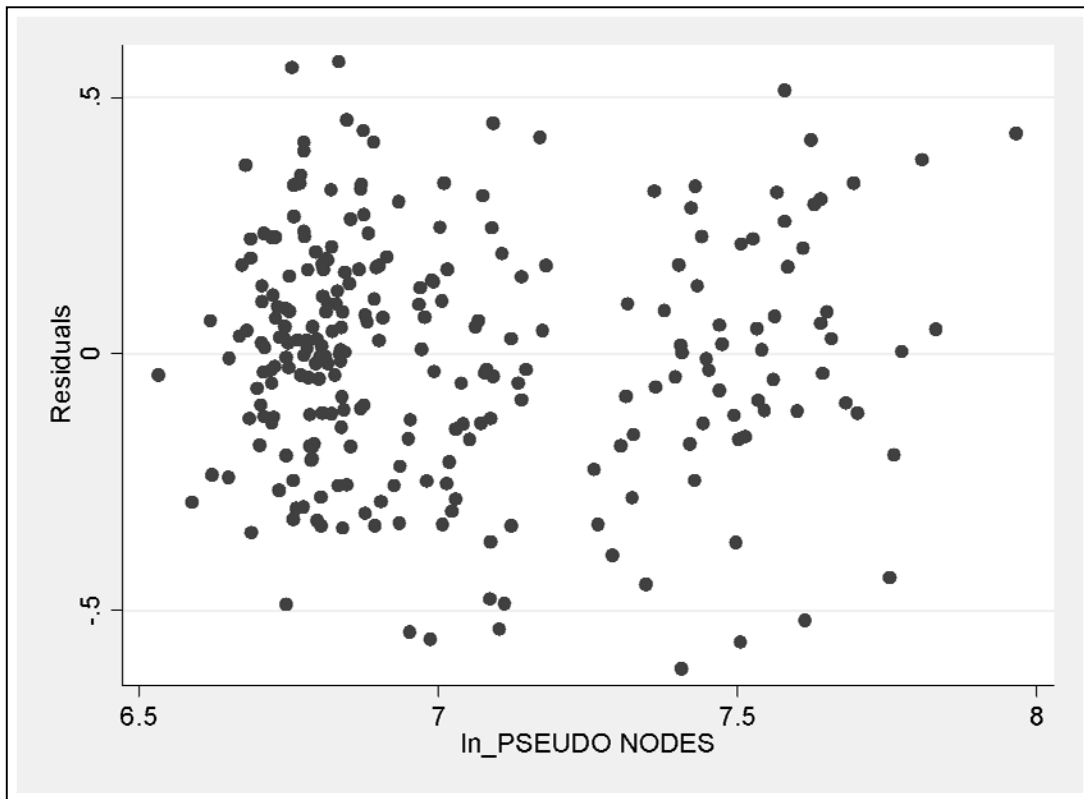


Figure 6.7 Scatter plot of Residuals and Pseudo Nodes

The scatter plots of residuals and each of the explanatory variables provide more insight into the nature of data and their potential impacts on model estimation. Although there are some distinctive patterns and outliers identified in the residual plots, the plot of residuals and fitted values in Figure 6.1 demonstrates that those potential impacts from each explanatory variable are mitigated through the model specification. Despite the presence of heteroscedasticity evident from the Breusch-Pagan Test, the analysis on the residuals does not suggest strong non-linearity or non-normality of the static pseudo panel data model.

6.3.4 Comparison of estimation techniques

The estimation results presented above are based on the pooled OLS estimation. The pooled OLS estimator is suggested by the Monte Carlo simulation experiment discussed in Chapter 5 as the preferred estimator for static pseudo panel data models. This section compares the estimation results of the static pseudo panel data model based on various estimators examined in the Monte Carlo experiment. This comparison is conducted not only to evaluate the

performance of these estimation techniques, but also to validate the simulation results presented in Chapter 5.

Table 6.5 summarises the estimation results of the static pseudo panel data model using the pooled OLS, FE, RE, and PCSE estimators. The FE estimator, which has been commonly used in previous pseudo panel data studies, essentially applies the OLS estimation to the static model whilst controlling the unobserved group effects and gives the same estimation results as using the Least Square Dummy Variable (LSDV) estimation. Its estimation results do not show a good model fit based on the low adjusted R-squared and the insignificances of explanatory variables. This is because most explanatory variables in this pseudo panel dataset have a much larger between-group variance than the within-group variance. As a result, the FE estimator, which only takes account of the within-group variance in order to eliminate unobserved group effects, is inefficient and thus inflates the standard errors of the parameters and shows a poor fit. These results correspond to the findings from the simulation experiment which suggests that the FE estimator has poor efficiency when the between-group variance is larger than the within-group variance.

The RE estimator, which controls for the unobserved group effects but assumes no correlation between explanatory variables and error terms, gives similar estimation results to the pooled OLS estimator. All the explanatory variables have the same signs as the pooled OLS estimation with very minor differences in the coefficient values. The RE estimation also identifies that the contribution of the variance from the unobserved group effects (ρ) is only around ten percent of total variance in the estimation. This indicates that the effect of the variance of unobserved group effects (σ_u) in the static model is fairly small and thus is not expected to cause strong bias in the pooled OLS estimation.

Table 6.5 A Comparison of Static Model Estimation Results with Various Estimators

	OLS	FE	RE	PCSE
PRICE	-0.269*** (0.071)	0.017 (0.097)	-0.262*** (0.079)	-0.269*** (0.068)
INCOME	-0.221*** (0.057)	-0.108 (0.068)	-0.184*** (0.063)	-0.221*** (0.056)
AGE	-0.750*** (0.066)	-0.967*** (0.165)	-0.751*** (0.079)	-0.750*** (0.065)
BUS FREQUENCY	0.206*** (0.046)	-0.004 (0.060)	0.177*** (0.051)	0.206*** (0.045)
POPULATION DENSITY	0.753*** (0.142)	0.025 (0.184)	0.741*** (0.151)	0.753*** (0.138)
LAND MIX	0.001 (0.120)	-0.021 (0.112)	-0.011 (0.119)	0.001 (0.119)
PSEUDO NODES	-0.570*** (0.099)	-0.003 (0.134)	-0.543*** (0.105)	-0.570*** (0.102)
DISTANCE TO PT STOP	0.061 (0.063)	0.047 (0.058)	0.0570 (0.062)	0.061 (0.063)
PT STOPS	-0.200 (0.146)	-0.011 (0.148)	-0.171 (0.149)	-0.200 (0.149)
CONSTANT	0.023 (1.595)	3.139 (1.910)	-0.391 (1.711)	0.023 (1.587)
Observations	256	256	256	256
R-squared	0.872	0.248	0.871	0.872
Adjusted R-squared	0.867	0.155		
σ_α			0.071	
σ_ε			0.205	
rho ¹			0.107	

Note: Standard errors in parentheses; * P<0.10, ** P<0.05, *** P<0.01;

Double log models are employed and estimated by OLS.

¹fraction of variance due to unobserved group effect

The PCSE estimator, which corrects for heteroscedasticity from the pooled OLS estimator, shows exactly the same coefficients as the pooled OLS estimator. This is expected because the PCSE estimator only corrects the standard errors from the pooled OLS estimator due to the presence of heteroscedasticity. It can also be observed that the standard errors of parameters in the PCSE estimator differ only slightly from the standard errors in the pooled OLS estimation. This confirms that the heteroscedasticity in the static model is not substantial. The comparison of the estimation techniques validates the use of the pooled OLS estimator and also confirms the findings from the simulation experiments in Chapter 5 in this empirical case study.

6.4 Summary

This chapter presents the analysis of public transport demand in the SGMA using a static pseudo panel data model with standard procedures of statistical analysis, including descriptive statistics, model estimations and evaluations, and model diagnostics. The step-by-step analysis investigates the nature of the data employed and tests the validity of model forms and estimation techniques in use.

The discussion of the descriptive statistics highlights the unique properties of the pseudo panel dataset, including the large between-group variation and the issues related to the aggregation level of the cohort data. These properties in turn impact on the estimation results. The large between-group variation in the explanatory variables leads to inflated standard errors of parameters for the FE estimation, the importance of the Monte Carlo simulation to underpin the choice of the estimation technique. The aggregation level of cohort data is shown to reduce measurement errors as evident by the smaller standard deviations of variables, as compared to the GWR global model, thus improving the model goodness-of-fit despite the smaller number of observations being estimated in the pseudo panel data model.

From the model diagnostics, the static model shows some degree of heteroscedasticity but this is not considered to have a strong impact on the estimation results, and the biased standard errors due to heteroscedasticity can also be corrected by the PCSE estimator. However, omitted variable bias is evident in the static models, regardless of which functional form is used. This is suspected to be a consequence of omitting the lagged dependent variable which captures the dynamics of travel behaviour changes. Thus, Chapter 7 accommodates this issue by employing a dynamic partial adjustment model to identify the lagged adjustments of public transport demand in the SGMA.

Although the estimation results from the static model are not further discussed in terms of their policy implications due to the potential bias of omitted variables, this chapter presents a rigorous statistical analysis which details each procedure in the econometric analysis. The contribution of this chapter to this study is the

investigation of the basic model structures and the preliminary findings of the pseudo panel data approach. The procedures of the analysis also contribute to the literature by highlighting the importance of these basic model assumption tests and the evaluation of model forms, which may lead to questionable research findings if they are ignored in an econometric analysis.

CHAPTER 7 DYNAMIC MODEL ESTIMATION

7.1 Introduction

As identified in Chapter 6, the static pseudo panel data model shows omitted variable bias and thus requires further investigation for a better model specification. This chapter presents the analysis of the dynamic pseudo panel data model taking account of the lagged demand adjustment which is not captured by the static model. The estimation results from this dynamic modelling are compared to the static models to identify the potential causes of the estimation bias identified in the static models⁵.

The theoretical background of dynamic models and their various functional forms and model specification are first introduced in Section 7.2. The estimation results and related model assumption diagnostics as well as model specification tests are presented and discussed in Section 7.3. The best dynamic model functional form, evaluated from Section 7.3, is then used to estimate short-run and long-run demand elasticities with respect to each of the explanatory variable, with a detailed discussion on their policy implications in Section 7.4. A final summary and discussion for future research directions extended from the dynamic model analysis is presented in Section 7.5.

7.2 Model specifications

7.2.1 *Dynamic models*

The purpose of dynamic modelling is reviewed in Section 2.1.2 and the dynamic models can be specified in various forms. The general form of a dynamic model specified by Equation (7.1) is known as a distributed-lag model in which the duration of the lagged adjustment is infinite. The lagged adjustment can be captured by a short-run multiplier given coefficient β_0 and the long-run multiplier given by $\sum_{i=1}^k \beta_i$ which represents the cumulated effects of independent variables X on the dependent variable Y .

⁵ Parts of the work presented in this chapter were presented in the 13th International Conference on Travel Behavior Research in July 15-20, 2012 (Tsai and Mulley, 2012). A revised paper has been accepted for publication in Journal of Transport and Economics Policy (Tsai and Mulley, 2013).

$$Y_t = \alpha + \beta_0 X_t + \beta_1 X_{t-1} + \beta_2 X_{t-2} + \dots + \beta_k X_{t-k} + u_t \quad \text{Equation (7.1)}$$

The issue in the estimation of this distributed-lag model is the determination of the number of lags to be included. Although this may be determined by using a general-to-specific approach (Campos et al., 2005) which applies a testing-down procedure to eliminate insignificant lagged variables, the successive lags remaining in the model tend to be highly correlated and thus lead to multicollinearity problems.

An alternative approach to estimating the distributed-lag model is to impose prior restrictions on the coefficients of the lagged variables. Koyck (1954) assumed that the coefficients of the lagged values of X decline geometrically as Equation (7.2).

$$\beta_k = \beta_0 \lambda^k, \quad 0 < \lambda < 1 \quad \text{Equation (7.2)}$$

Equation (7.2) shows that the effect of the lags on the dependent variable becomes progressively smaller and are captured by the coefficient λ . Thus, the distributed-lag model (Equation (7.1)) can be re-written as Equation (7.3).

$$Y_t = \alpha + \beta_0 X_t + \beta_1 \lambda X_{t-1} + \beta_2 \lambda^2 X_{t-2} + \dots + u_t \quad \text{Equation (7.3)}$$

To transform Equation (7.3) into a linear function and mitigate the multicollinearity problem, Koyck proposed to lag Equation (7.3) by one time period and multiply all variables by λ :

$$\lambda Y_{t-1} = \alpha \lambda + \beta_0 \lambda X_{t-1} + \beta_1 \lambda^2 X_{t-2} + \beta_2 \lambda^3 X_{t-3} + \dots + \lambda u_{t-1} \quad \text{Equation (7.4)}$$

Subtracting Equation (7.4) from Equation (7.1) gives:

$$Y_t = \alpha(1 - \lambda) + \beta_0 X_t + \lambda Y_{t-1} + v_t, \quad v_t = u_t - \lambda u_{t-1} \quad \text{Equation (7.5)}$$

Equation (7.5) is known as Koyck's model which eliminates the multiple lags of X , transforming the distributed lag model to an autoregressive model.

A different form of the geometric lag model is the Partial Adjustment Model (PAM) introduced by Nerlove (1958). The PAM assumes that the desired level of Y is a linear function of X as Equation (7.6),

$$Y_t^* = \alpha + \beta_0 X_t + u_t \quad \text{Equation (7.6)}$$

and an adjustment Equation (7.7).

$$Y_t - Y_{t-1} = (1 - \lambda)(Y_t^* - Y_{t-1}) \quad \text{Equation (7.7)}$$

Substituting Equation (7.6) into Equation (7.7) gives:

$$\begin{aligned} Y_t &= \alpha(1 - \lambda) + \beta_0(1 - \lambda)X_t + \lambda Y_{t-1} + (1 - \lambda)u_t \\ &= \alpha' + \beta_0' X_t + \lambda Y_{t-1} + u_t' \end{aligned} \quad \text{Equation (7.8)}$$

As a result, the PAM (Equation (7.8)) contains only the first lag of the dependent variable and eliminates the lags of independent variables. The PAM has been widely applied in modelling the dynamics of economic behaviour (as reviewed in Table 2.1) because of its practical advantages and the parsimonious functional form. It is an unrestricted linear function with a non-auto-correlated disturbance, so this model can be estimated using Ordinary Least Square (OLS). In contrast, the Koyck's model has an auto-correlated disturbance v_t which leads to problematic estimation results by the OLS estimator because of autocorrelation.

A dynamic model form which can be used to accommodate the auto-correlation problem thus addressing the issues raised by Koyck's model is the Error Correction Model (ECM). A general form of the ECM can be specified as Equation (7.9).

$$\Delta Y_t = \sum_{i=1}^k \lambda_i Y_{t-k} + \beta_1 \Delta X_t + \sum_{i=1}^k \beta_i X_{t-k} + \varepsilon_t \quad \text{Equation (7.9)}$$

The ECM applies a first-differencing approach to eliminate the nonstationarity in the model. It can also be used to estimate the short-run and long-run elasticities for time-series data. Jevons et al. (2005) compared the elasticities estimated from a PAM and a ECM and suggested that the results may vary depending on the model and time intervals in use. However, in a panel data analysis, the ECM is not able to incorporate the between-group variance because it only takes account of the over time changes in the dependent variable and exogenous variables. Given the substantial between-group variance in the pseudo panel dataset of this study, the ECM is not considered for this study and the partial PAM is employed to estimate the dynamic pseudo panel data models instead.

The dynamic public transport demand model for this analysis in the PAM form expanded from the general dynamic pseudo panel data model (Equation (4.5)) is defined by Equation (7.10).

$$\begin{aligned} \bar{D}_{g,t} = & \beta_0 + \lambda_1 \bar{D}_{g,t-1} + \beta_1 \overline{PRICE}_{g,t} + \beta_2 \overline{INCOME}_{g,t} + \beta_3 \overline{AGE}_{g,t} + \beta_4 \overline{FREQ}_{g,t} \\ & + \beta_5 \overline{DENSITY}_{g,t} + \beta_6 \overline{LANDMIX}_{g,t} + \beta_7 \overline{PSEUDO}_{g,t} + \beta_8 \overline{DISTANCE}_{g,t} \quad \text{Equation (7.10)} \\ & + \beta_9 \overline{STOPS}_{g,t} + \bar{u}_{g,t}, \quad \bar{u}_{g,t} = \bar{\alpha}_{g,t} + \bar{\varepsilon}_{g,t} \end{aligned}$$

Compared to the static pseudo panel data model presented in Chapter 6 (Equation (6.13)), this dynamic pseudo panel data model adds a lagged dependent variable of public transport demand ($\bar{D}_{g,t-1}$) to estimate the impact of demand at time period $t - 1$ on the current demand at time period t . As discussed in Section 2.1.2, this lagged dependent variable captures the temporal effects of demand adjustment, which has been suggested as a result of travellers' habits or other factors such as household locations that are not able to change in the short term. Other lag structures of the dependent variable were initially tested but only the first lag of dependent variable was found significant, and hence further lags were removed from the PAM model.

Although the dynamic pseudo panel data model form is similar to the static model, the introduction of the lagged dependent variable leads to more issues related to model estimation that need to be accommodated. These issues are reviewed in Section 2.4.2 and briefly summarised in the next section.

7.2.2 Estimation techniques

Section 2.4.2 reviews the theory of the dynamic panel data model estimation and discusses the performance of estimators including the pooled Ordinary Least Squares (OLS), Fixed Effect (FE), Random Effect (RE), Panel-Corrected Standard Error (PCSE), and Instrumental Variable (IV) estimators when they are employed to estimate a dynamic panel data model. In short, the pooled OLS estimator is expected to be biased upwards due to the presence of unobserved individual effect in a genuine panel data model and unobserved group effects in a pseudo panel data model. The FE and RE estimators theoretically can be shown to be biased because of the endogeneity between the lagged dependent variable and the composite error term. The PCSE estimator that corrects for the non-spherical errors is also biased when the pooled OLS estimator is biased. The IV estimator, although able to control for the endogeneity by introducing an instrumental variable, is practically difficult to implement because the appropriateness of the instrumental variable is hard to justify and it has been suggested that it may be inefficient when the number of panels is small in a panel dataset as discussed in Section 2.4.2.

Given that most of the estimators are potentially problematic in estimating a dynamic panel data model, it is likely that a Best Linear Unbiased Estimator (BLUE) does not exist, especially in pseudo panel data analysis in which there are more restrictions in model estimation such as small sample size and time-varying group effects. However, a preferred estimator can still be determined by evaluating the relative bias and efficiency among estimators, and this is shown by the Monte Carlo simulation experiment presented in Chapter 5.

The Monte Carlo experiment in Chapter 5 shows that the FE estimator is the preferred estimator using the overall RMSE as a justification when the exogenous variable is identically distributed across groups and time, that is, where there is the same between-group variance and within-group variance. However, when the explanatory variable has a substantially larger between-group variance than the within-group variance, the FE estimator is extremely inefficient and thus the pooled OLS and the PCSE estimator perform better than the FE estimator. Although theoretically there is still a certain degree of bias existing in the pooled OLS and PCSE estimators, the bias can be mitigated by reducing the variance of unobserved group effects as shown in the simulation results in Chapter 5. .

Therefore, based on the findings from the simulation experiments, the dynamic pseudo panel data model analysed in this chapter is first estimated by the pooled OLS estimator with the results being compared to other estimators in next section.

7.3 Estimation results of dynamic pseudo panel data models

7.3.1 Base model

As with the static pseudo panel data analysis in Chapter 6, the dynamic pseudo panel data model is estimated by the pooled OLS estimator in the linear functional form as a base model. The pooled OLS estimation results of the base model are presented first in Table 7.1 and are then compared to the results estimated from the static base model in Table 7.2.

Table 7.1 shows that the linear dynamic model has a fairly high adjusted R-squared value of 0.820 with no strong multi-collinearity identified, but with significant omitted variable bias, heteroscedasticity and first-order autocorrelation as shown by the results of RESET test, Breusch-Pagan Test and Wooldridge test.

In terms of the parameter estimates, the lagged dependent variable is significant at 0.433 with a positive sign which suggests that the demand of the previous time

period has an intermediate effect on current period demand given the same units of the dependent variable and the lagged dependent variable. For other explanatory variables, personal income and age are negatively significant at a 95 percent confidence level, and population density and pseudo nodes are only significant at the 90 percent confidence level. Other variables do not show coefficients significantly differing from zero in this dynamic base model.

Table 7.1 Pooled OLS Estimation Results of the Linear Dynamic Model (Base Model)

Dependent Variable:							
PTTRIP	Coef.	Std. Err.	T	P-value	[95% C.I.]		VIF
LAG1	0.433	0.062	6.99	0.000	0.311	0.555	5.38
PRICE	-0.016	0.031	-0.52	0.603	-0.077	0.045	5.97
INCOME	-0.004	0.001	-3.91	0.000	-0.006	-0.002	3.18
AGE	-0.003	0.001	-3.67	0.000	-0.005	-0.002	4.56
BUS FREQUENCY	0.321	0.094	3.41	0.001	0.136	0.507	3.65
POPULATION DENSITY	0.006	0.004	1.80	0.073	-0.001	0.014	7.24
LAND MIX	-0.824	0.623	-1.32	0.188	-2.051	0.404	1.29
PSEUDO NODES	-0.043	0.025	-1.69	0.093	-0.093	0.007	4.43
DISTANCE TO PT STOP	-0.006	0.139	-0.04	0.966	-0.280	0.268	1.92
PT STOPS	0.00004	0.002	0.02	0.985	-0.004	0.004	4.66
CONSTANT	0.489	0.168	2.92	0.004	0.159	0.820	
Observations	236						
F(10, 225)	109.48						
Prob > F	0.000						
R-squared	0.830						
Adjusted R-squared	0.820						
Root MSE	0.114						
Ramsey RESET Test (Ho: Model has no omitted variables)							
F(3, 243)	5.07						
Prob > F	0.002						
Breusch-Pagan Test for heteroscedasticity (Ho: Constant variance)							
chi ² (1)	89.35						
Prob > Chi ²	0.000						
Wooldridge test for autocorrelation (Ho: No first order autocorrelation)							
F(1, 19)	8.628						
Prob > F	0.009						

Comparing the estimation results of the static and dynamic base models in Table 7.2, it can be observed that the dynamic model has a better model goodness-of-fit according to its higher adjusted R-squared value than the static model, even though 20 observations in the dynamic model are removed as a result of the

missing lagged values in the first time period of each of the 20 groups. As mentioned, both models show significant omitted variable bias and heteroscedasticity, but the dynamic model also shows significant autocorrelation which is not evident in the static model. This typically occurs when there is a lagged dependent variable which is very likely to be correlated with its own value of the previous time period which introduces autocorrelation. The Variance Inflation Factors (VIFs), which indicate the magnitude of multi-collinearity, are generally higher in the dynamic model than the static model. This is also considered as a result of the inclusion of the lagged dependent variable which is a predicted value of the explanatory variables at time period $t - 1$, and this in turn increases the degree of multi-collinearity among all explanatory variables. Nevertheless, the impact does not appear to be strong given that all the VIFs are lower than ten.

The parameter estimates show some differences between the static and dynamic models. The price variable becomes insignificant in the dynamic model, and population density as well as pseudo nodes are only significant at 90 percent confidence level, whereas they are all significant at 95 percent confidence level in the static model. A possible reason is that the lagged dependent variable has a stronger explanatory power than these explanatory variables, so the variation in public transport demand is explained by the lagged dependent variable more than price, population density, and pseudo nodes in the dynamic model. However, the presence of the omitted variable bias in both models may also confound the estimation results so this comparison can only be seen as exploratory analysis, and the justification of the best dynamic model form requires further analysis as presented in the next section (Section 7.3.2).

Table 7.2 A Comparison of Pooled OLS Estimation Results between the Static and Dynamic Base Models

Dependent Variable: PTTRIP	Static Model			Dynamic Model		
	Coef.	P-value	VIF	Coef.	P-value	VIF
LAG1	n/a	n/a	n/a	0.433	0.000	5.38
PRICE	-0.065	0.049	5.68	-0.016	0.603	5.97
INCOME	-0.007	0.000	2.84	-0.004	0.000	3.18
AGE	-0.007	0.000	3.01	-0.003	0.000	4.56
BUS FREQUENCY	0.507	0.000	3.34	0.321	0.001	3.65
POPULATION DENSITY	0.012	0.002	6.97	0.006	0.073	7.24
LAND MIX	0.024	0.972	1.29	-0.824	0.188	1.29
PSEUDO NODES	-0.101	0.000	4.13	-0.043	0.093	4.43
DISTANCE TO PT STOP	0.009	0.953	1.97	-0.006	0.966	1.92
PT STOPS	-0.001	0.645	4.63	0.00004	0.985	4.66
CONSTANT	0.888	0.000	n/a	0.489	0.004	n/a
Observations	256			236		
Adjusted R-squared	0.782			0.820		
Root MSE	0.129			0.114		
Ramsey's RESET TEST:						
F-statistics	16.46			5.07		
Prob > F	0.000			0.002		
Breusch-Pagan Test						
chi2(1)	55.00			89.35		
Prob >Chi ²	0.000			0.000		
Wooldridge test						
F-statistics	0.581			8.628		
Prob > F	0.455			0.009		

7.3.2 Test of functional forms

Table 7.3 summarises the estimation results of the dynamic pseudo panel data model with four different functional forms. Based on the adjusted R-squared and the Ramsey's RESET test, the double-log model outperforms other functional forms given the highest adjusted R-squared at 0.872, with no significant omitted variable in contrast to all other models where the omitted variable bias is significant. The double-log model has no heteroscedasticity or autocorrelation present, evident in both the linear model and the linear-log models. The double-log model also demonstrates better explanatory power as demonstrated by the significance of explanatory variables. Variables that are not significant at 95 percent confidence level in the dynamic linear model including price, population density, and pseudo nodes, are significant in this double-log model. The evidence

discussed above collectively suggests that the double-log is the best functional form for the dynamic pseudo panel data model in this study.

Table 7.3 Evaluation of Dynamic Model Functional forms

	LINEAR	LINEAR- LOG	DOUBLE- LOG	LOG- LINEAR
LAG1	0.433*** (0.062)	0.425*** (0.061)	0.245*** (0.067)	0.360*** (0.066)
PRICE	-0.016 (-0.031)	-0.003 -0.037	-0.219*** (-0.076)	-0.146** -0.066
INCOME	-0.004*** -0.001	-0.113*** (-0.032)	-0.160** (-0.062)	-0.007*** -0.002
AGE	-0.003*** -0.001	-0.141*** (-0.039)	-0.573*** (-0.086)	-0.01*** -0.002
BUS FREQUENCY	0.321*** -0.094	0.074*** (0.026)	0.148*** (0.051)	0.296 -0.190
POPULATION DENSITY	0.006* 0.004	0.171** (0.0741)	0.596*** (0.152)	0.027*** 0.008
LAND MIX	-0.824 -0.623	-0.001 (-0.061)	-0.028 (-0.121)	-1.240 -1.287
PSEUDO NODES	-0.043* -0.025	-0.103** (-0.052)	-0.458*** (-0.109)	-0.166*** -0.055
DISTANCE TO PT STOP	-0.006 -0.139	0.019 (0.033)	0.0679 (0.066)	0.207 -0.293
PT STOPS	0.040 -0.002	-0.079 (-0.075)	-0.174 (-0.151)	0.003 -0.004
CONSTANT	0.489*** (0.168)	0.767 (0.828)	-0.164 (-1.645)	-0.184 -0.342
Observations	236	236	236	236
R-squared	0.830	0.828	0.877	0.866
Adjusted R-squared	0.822	0.820	0.872	0.860
Ramsey RESET Test				
Prob > F	0.002	0.000	0.072	0.027
Breusch-Pagan Test				
Prob > Chi ²	0.000	0.000	0.074	0.097
Wooldridge test				
Prob > F	0.009	0.020	0.423	0.227

Note: Standard errors in parentheses; * P<0.10, ** P<0.05, *** P<0.01;
Models are estimated by OLS.

The implications of these estimation results are next discussed and compared to the static double-log model. Table 7.4 shows that the dynamic model has a slightly better adjusted R-squared than the static model, and both the omitted variable bias and heteroscedasticity identified in the static model are not significant in the dynamic model. This implies that the omitted variable bias in

the static model may be a consequence of omitting the lagged dependent variable which is therefore identified as significant in public transport demand. Moreover, the coefficients are relatively smaller in the dynamic model than in the static model arising from the way in which the lagged dependent variable has a certain degree of impacts on the variation in the dependent variable, which in turn reduces the impacts from other explanatory variables. This indicates that the static model over-estimates the influence of the explanatory variables on public transport demand as a result of omitting the lagged adjustments of demand changes.

Table 7.4 A Comparison of the Best Static and Dynamic Functional Forms

Dependent Variable: PTTRIP	Static Model (DOUBLE-LOG)	Dynamic Model (DOUBLE-LOG)
LAG1	n/a	0.245***
PRICE	-0.269***	-0.219***
INCOME	-0.221***	-0.160**
AGE	-0.750***	-0.573***
BUS FREQUENCY	0.206***	0.148***
POPULATION DENSITY	0.753***	0.596***
LAND MIX	0.001	-0.028
PSEUDO NODES	-0.570***	-0.458***
DISTANCE TO PT STOP	0.0610	0.0679
PT STOPS	-0.200	-0.174
CONSTANT	0.0227	-0.164
Observations	256	236
R-squared	0.872	0.877
Adjusted R-squared	0.867	0.872
Ramsey RESET Test		
Prob > F	0.007	0.072
Breusch-Pagan Test		
Prob >Chi ²	0.044	0.074

Note: Standard errors in parentheses; * P<0.10, ** P<0.05, *** P<0.01;
Models are estimated by OLS.

Comparing the parameter estimates between the static and dynamic models, all the explanatory variables have the same significance at 95 percent confidence level with the same expected signs. Price, income, age, and pseudo nodes have negative impacts on public transport demand in the Sydney Greater Metropolitan Area (SGMA), whereas bus frequency and population density have

positive impacts on public transport demand. The lagged dependent variable does not dominate the model prediction power or change the significance or signs of other exogenous variables suggesting that the selected explanatory variables properly explain variations in public transport demand. The coefficient of the lagged dependent variable at 0.245 suggests that if public transport demand in the previous period was to increase by one hundred percent, then current public transport demand would increase by 24.5 percent change in current demand. This parameter can also be used to distinguish between short-run and long-run demand elasticities as presented in the next section (Section 7.4).

Although the analysis above shows that the double-log model appears to be the best functional form of the dynamic pseudo panel model, there are three insignificant variables that require further investigation for the best model specification. These are land use mix, distance to the nearest bus stop, and number of bus stops, which are selected as explanatory variables in the public transport demand model because they have been suggested to be influential on travel behaviour in the literature (as discussed in Section 2.2.1). The possible reasons for the insignificance of these three variables in this public transport demand model are discussed in Section 6.3, with no evidence suggesting that the insignificance is resulted from multi-collinearity. Nevertheless, there is always a trade-off between a parsimonious regression model with significant variables only and an unrestricted regression model that consists of variables suggested from the theory. Hence, a test of model specifications is conducted by removing each of these three insignificant variables from the double-log model to investigate the impact on the estimation results. This is summarised in Table 7.5.

Table 7.5 Test of Dynamic Model Specifications

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8
LAG1	0.245*** (0.067)	0.248*** (0.067)	0.252*** (0.067)	0.246*** (0.067)	0.257*** (0.067)	0.253*** (0.067)	0.248*** (0.067)	0.257*** (0.067)
PRICE	-0.219*** (0.076)	-0.213*** (0.076)	-0.232*** (0.075)	-0.218*** (0.075)	-0.227*** (0.075)	-0.231*** (0.075)	-0.212*** (0.075)	-0.226*** (0.075)
INCOME	-0.160** (0.062)	-0.163*** (0.062)	-0.149** (0.062)	-0.161*** (0.062)	-0.150** (0.062)	-0.150** (0.061)	-0.164*** (0.062)	-0.151** (0.061)
AGE	-0.573*** (0.086)	-0.569*** (0.086)	-0.575*** (0.086)	-0.572*** (0.086)	-0.572*** (0.086)	-0.574*** (0.086)	-0.569*** (0.086)	-0.571*** (0.086)
BUS FREQUENCY	0.148*** (0.051)	0.146*** (0.051)	0.148*** (0.051)	0.146*** (0.050)	0.145*** (0.051)	0.144*** (0.050)	0.143*** (0.050)	0.141*** (0.050)
POPULATION DENSITY	0.596*** (0.152)	0.603*** (0.152)	0.500*** (0.127)	0.603*** (0.149)	0.491*** (0.127)	0.507*** (0.125)	0.610*** (0.149)	0.498*** (0.125)
PSEUDO NODES	-0.458*** (0.109)	-0.449*** (0.108)	-0.404*** (0.098)	-0.463*** (0.107)	-0.383*** (0.097)	-0.409*** (0.097)	-0.454*** (0.107)	-0.388*** (0.096)
LAND MIX	-0.028 (0.121)	-0.028 (0.121)	-0.038 (0.121)	n/a	-0.040 (0.121)	n/a	n/a	n/a
DISTANCE TO PT STOP	0.068 (0.066)	n/a	0.081 (0.065)	0.0679 (0.066)	n/a	0.081 (0.065)	n/a	n/a
PT STOPS	-0.174 (0.151)	-0.200 (0.148)	n/a	-0.176 (0.150)	n/a	n/a	-0.202 (0.148)	n/a
CONSTANT	-0.164 (1.645)	0.165 (1.614)	-0.391 (1.635)	-0.104 (1.622)	-0.031 (1.611)	-0.315 (1.614)	0.224 (1.591)	0.049 1.589
R-squared	0.877	0.877	0.877	0.877	0.876	0.877	0.877	0.876
Adj. R-squared	0.872	0.872	0.872	0.872	0.871	0.872	0.872	0.872
RESET TEST	0.072	0.072	0.040	0.072	0.038	0.039	0.072	0.038
HET TEST	0.074	0.066	0.058	0.080	0.048	0.063	0.070	0.052

Note: Standard errors in parentheses; * P<0.10, ** P<0.05, *** P<0.01; double log models are employed and estimated by OLS.

In Table 7.5, Model 1 is the unrestricted model and is used as the benchmark to compare to other restricted models. The three insignificant variables are placed at the bottom of the table for easier comparison. The models have minimal differences in their adjusted R-squares. However, models that include number of public transport stops (Model 1/2/4/7) appear to have a slightly better model specification than others, given their RESET test results suggesting no omitted variable bias (highlighted in bold text). This indicates that the number of public transport stops has a certain degree of explanatory power in public transport demand and even though it is not significant at the 90 percent confidence level or better.

Comparing the parameter estimates of these models, it can be seen that all the parameters have the same level of significance and the same signs with the exception of income which shows some slight difference in its level of significance. In terms of estimated coefficients, these are similar over the eight models. The lagged dependent variable, price, income, age, and bus frequency vary less than five percent between their highest and lowest values across the eight models. However, population density and the number of pseudo nodes show more sensitivity to the inclusion of insignificant variables, with population density varying from 0.491 in Model 5 to 0.610 in Model 7 and pseudo nodes ranging between -0.383 in Model 5 and -0.458 in Model 1. Moreover, models including number of bus stops (Model 1/2/4/7) and models not including number of bus stops (Model 3/5/6/8) appear to be two distinctive clusters according to the coefficients of population density and pseudo nodes. In general, Model 1/2/4/7 have higher values for the estimated coefficients of population density and pseudo nodes than Model 3/5/6/8 in absolute terms. This confirms the way in which the number of public transport stops has a certain degree of influence on explaining public transport demand in the model. Including number of public transport stops increases the coefficient of population density and decreases the coefficient of pseudo nodes, as a result of its inverse correlation to population density and pseudo nodes, but these changes are not substantial and do not distort the model prediction results.

As the difference of the coefficients between the unrestricted model and other restricted models is minimal, the three insignificant variables are kept in the model for the following analysis because collectively they still have some minor explanatory power on public transport demand although not significant at 90 percent statistical confidence level.

7.3.3 Model diagnostics

The model tests summarised in Table 7.3 have shown that the best dynamic model, the double-log model, has no significant omitted variable, heteroscedasticity, or autocorrelation. This section further investigates the distributions of residuals with respect to the fitted values and the lagged dependent variable as shown in Figure 7.1 and Figure 7.2 respectively. The two scatter plots show that the residuals are distributed randomly with no distinctive patterns identified. Some observations have relatively small residuals because of the small cohort size of those observations. Similar to the static model residual plots, those observations are not expected to distort the estimation results and are kept in the model estimation to improve estimation efficiency.

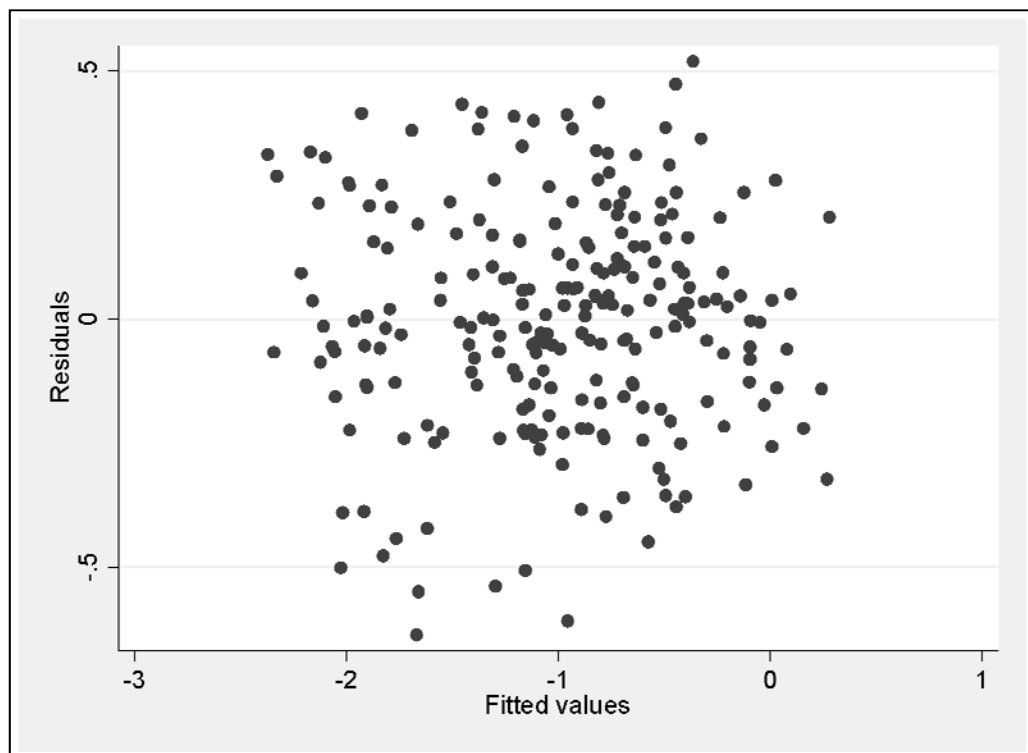


Figure 7.1 Scatter Plot of Residuals and Fitted Values from the Dynamic Model

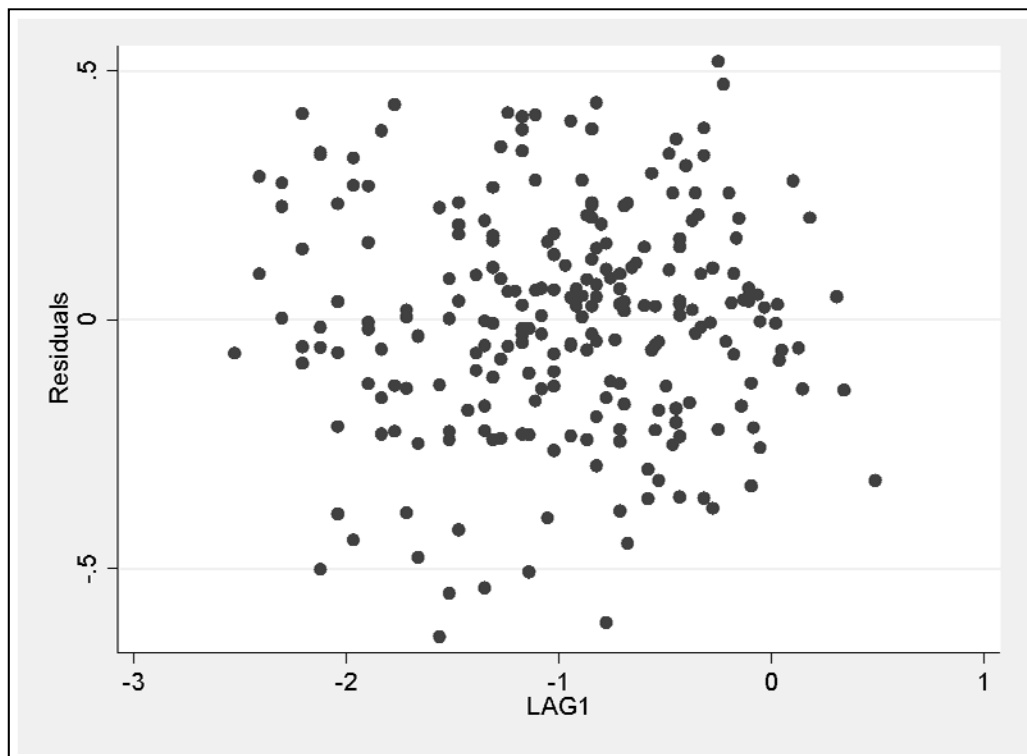


Figure 7.2 Scatter Plot of Residuals and the Lagged Variable from the Dynamic Model

In short, the double-log form of the dynamic pseudo panel data model demonstrates the best model goodness-of-fit, with no omitted variables or any significant non-spherical errors identified. There is a small degree of multicollinearity among some of the explanatory variables but they are not influential to the prediction power of this model. This choice of the best model functional form is thus made on the basis of the model specification tests and diagnostics presented above.

7.3.4 Comparison of estimation techniques

The dynamic pseudo panel data models presented above are estimated by the pooled OLS estimator. The use of the pooled OLS estimator for the characteristics displayed by the dataset for this study is recommended by the Monte Carlo experiment results concluded in Chapter 5. This section further investigates the estimation results from other estimators using the findings in the simulation experiments for completeness.

Table 7.6 summarises the estimation results of the dynamic model in the double-log functional form using the pooled OLS, FE, RE, and GMM estimators. The PCSE estimator which corrects the non-spherical errors from OLS estimation is not included here because the double-log model does not show any significant non-spherical errors as discussed in the previous section. In Table 7.6, it can be clearly seen that the FE estimator has the lowest model goodness-of-fit with a weak model predictive power, given its low adjusted R-squared and the insignificance of explanatory variables. This is the result of the large ratio of between-group variance to within-group variance in the variable, and also identified in the static model estimation (see Section 6.3.4).

Table 7.6 Dynamic Model Estimation Results using Various Estimators

	OLS	FE	RE	GMM
LAG1	0.245*** (0.067)	-0.048 (-0.072)	0.245*** (0.067)	-0.678 (-0.674)
PRICE	-0.219*** (-0.076)	-0.003 (-0.100)	-0.219*** (-0.076)	-0.489** (-0.225)
INCOME	-0.160** (-0.062)	-0.087 (-0.080)	-0.160** (-0.062)	-0.297* (-0.160)
AGE	-0.573*** (-0.086)	-1.024*** (-0.201)	-0.573*** (-0.086)	-1.318** (-0.550)
BUS FREQUENCY	0.148*** (0.051)	0.020 (0.064)	0.148*** (0.051)	0.368** (0.179)
POPULATION DENSITY	0.596*** (0.152)	-0.048 (-0.201)	0.596*** (0.152)	1.147** (0.500)
LAND MIX	-0.028 (-0.121)	-0.049 (-0.117)	-0.028 (-0.121)	-0.200 (0.199)
PSEUDO NODES	-0.458*** (-0.109)	0.007 (0.143)	-0.458*** (-0.109)	-0.912** (-0.406)
DISTANCE TO PT STOP	0.068 (0.0656)	0.080 (0.062)	0.068 (0.066)	0.054 (0.117)
PT STOPS	-0.174 (-0.151)	-0.0001 (-0.156)	-0.174 (-0.151)	-0.301 (0.298)
CONSTANT	-0.164 (1.645)	3.375 (2.128)	-0.164 (-1.645)	
Observations	236	236	236	236
R-squared	0.877	0.200	0.877	
Adjusted R-squared	0.872	0.087		
σ_α			0	
σ_ε			0.208	
rho			0	

Note: Standard errors in parentheses; * P<0.10, ** P<0.05, *** P<0.01;
Double log models are employed

The estimation results from the GMM estimator show larger standard errors for every parameter than other estimators. As reviewed in Section 2.4.2, the GMM estimator tends to be inefficient and thus inflates the standard errors when the number of groups (or number of panels in a genuine panel dataset) is small (Bruno, 2005a) and this is the case given the number of groups in the pseudo panel dataset is 20 which is considered to be rather small. Although most explanatory variables are still significant in the GMM estimation, albeit with a lower confidence level, the lagged dependent variable is insignificant. This finding might be expected from the simulation experiment results for dynamic models presented in Table 5.4, which identify that the GMM estimator substantially inflates the standard errors of the lagged dependent variable much more than the standard errors of the explanatory variable. Thus, the lagged dependent variable of the pseudo panel data model becomes insignificant because of estimation inefficiency, whilst other exogenous variables have only marginal increases in standard errors but remain significant.

The RE model degenerates to the OLS model as evident from exactly same estimation results between the two estimators in Table 7.6. This effect has been identified by Baltagi (2008, p. 20) and the reason is that the variance of the unobserved individual effect (σ_{α}^2) may be negative from the RE estimation process⁶ and will be replaced by zero when it is negative. This happens when the unobserved individual effects (or group effects in a pseudo panel dataset) is minimal and the variance of the i.i.d error terms is substantially larger than the variance of the unobserved individual effects. This effect is confirmed in this dynamic model by the zero value of σ_{α} in Table 7.6 and the *rho* (representing the fraction of variance due to unobserved group effects) as zero. This effect is not evident in the static model but in the dynamic model where the model specification is improved by the inclusion of the lagged dependent variable which reduces the variance of the unobserved group effects. This suggests that the bias of the pooled OLS estimator or the RE estimator may not be substantial since the unobserved heterogeneity which causes bias in the pooled OLS model is substantially reduced after including the lagged dependent variable in the model.

⁶ The variance of unobserved individual effect (Baltagi, 2008, p.20): $\hat{\sigma}_{\alpha}^2 = [(T \sum_{G=1}^N \bar{\alpha}_g^2 / G) - \hat{\sigma}_{\bar{\epsilon}}^2] / T$

To conclude, although a pseudo panel data model in theory may require cohort dummies or may be better estimated using the FE estimator to control the unobserved group effects, this study finds that the unobserved group effects are not substantial in this public transport demand model because of the proper model specification. Moreover, the FE estimator has been shown to be inefficient for the pseudo panel data model because its inability of capturing the between-group variance as demonstrated in Section 5.4.2. Hence, the use of the pooled OLS estimator for the dynamic pseudo panel data model of this study is justified, and it is considered to be the most appropriate estimator with the lowest RMSE among all the estimators presented in this section.

7.4 Estimation of demand elasticities

Public transport demand elasticity is a measure of how travellers' demand changes in response to the changes in its determinants. The demand elasticity is not necessarily constant over time. Instead, it may vary over time in accordance with the speed of demand adjustment and thus there is a distinction between short-run and long-run demand elasticities as reviewed in Section 2.1.2. The short-run elasticity (\bar{e}_k^{SR}) estimated from the double-log regression model is the coefficient of the variable concerned, whereas the long-run elasticity (\bar{e}_k^{LR}) is a function of the short-run elasticity and the coefficient of the lagged dependent variable as specified in Equation (7.11), with a speed of adjustment derived from Equation (7.12). The speed of adjustment and demand elasticities presented in this section are estimated from the double-log form of the dynamic pseudo panel data model which is justified as the preferred model, based on the pooled OLS estimator as the preferred estimator for the dynamic public transport demand model discussed in Section 5.4.2.

$$e_k^{LR} = \beta_k / 1 - \lambda \quad \text{Equation (7.11)}$$

$$T = \ln(1 - A) / \ln(\lambda) \quad \text{Equation (7.12)}$$

where

T : number of years for A percent of demand to adjust

A : proportion of demand adjustment

λ : coefficient of the lagged dependent variable

The timeframes for the short-run demand and long-run demand are referred to by Jevons (2005), who defines the long-run time period as the number of years for 95 percent of long-run demand to work through. Although some studies have suggested using one hundred percent of long-run demand adjustment to determine the long-run time period (Dargay, 2002), the speed of demand adjustment derived from Equation (7.12) (where $\lambda = 0.245$) shown in Figure 7.3 shows that the difference in the number of years required for 95 percent and one hundred percent demand adjustment is very minimal. This minimal difference is not expected to make a significant difference in the long-run demand estimation. From Figure 7.3, it can be seen that the speed of adjustment decreases over time, with around 75 percent of demand adjusted within the first year which is defined as the timeframe for short-run demand change. It then takes around 2.13 years for 95 percent of demand to be fully adjusted and this is the timeframe required to reach the long-run demand equilibrium. The speed of demand adjustment implies that the responsiveness of public transport users to system changes, such as fare or population density, may take up to 2.13 years to be fully observed. This gives a strong message to policy makers that long-run demand changes are important to take account of when planning public transport systems.

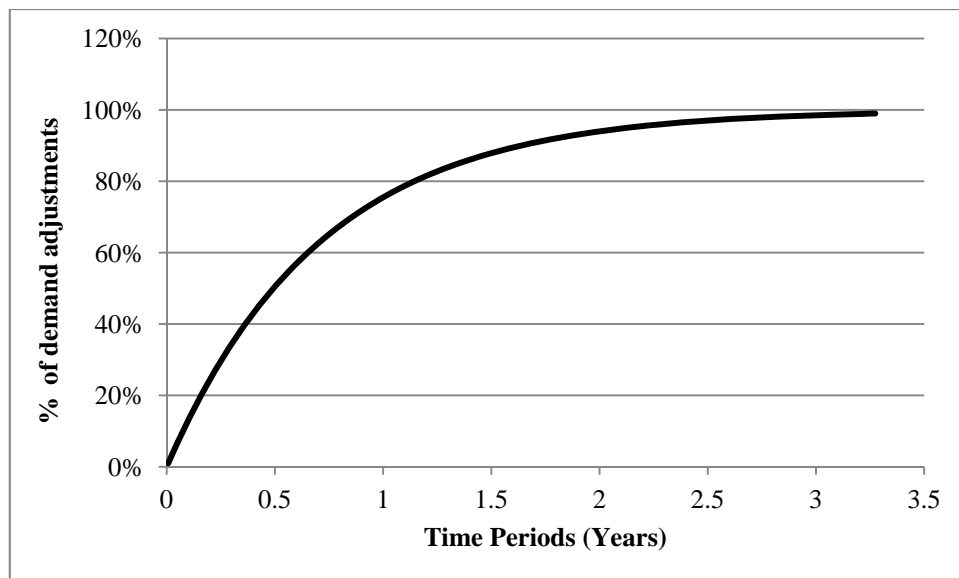


Figure 7.3 The Speed of Public Transport Demand Adjustments

The short-run and long-run demand elasticities with respect to the significant explanatory variables in the dynamic double-log model are summarised in Table 7.7.

Table 7.7 Demand Elasticities Derived from the Best Dynamic Public Transport Demand Models

	Dynamic Model	
	Short-Run	Long-Run
PRICE	-0.22	-0.29
INCOME	-0.16	-0.21
AGE	-0.57	-0.76
BUS FREQUENCY	0.15	0.20
POPULATION DENSITY	0.60	0.79
PSEUDO NODES	-0.46	-0.61

The short-run and long-run price elasticities estimated from the dynamic model are -0.22 and -0.29 respectively, suggesting that a ten percent increase in price is expected to reduce public transport demand by 2.2 percent in the short run (i.e., one year), but it will reduce public transport demand by 2.9 percent in the long run (i.e., 2.13 years). The estimated price elasticities from the dynamic pseudo panel data model are of the same order of magnitude as the public transport price elasticities estimated by Hensher (1998) using mixed Reveal Preference and Stated Preference data in Sydney, in which the price elasticity of train travel was found to be between -0.093 and -0.218 and the price elasticity of bus travel ranged between -0.098 and -0.357 varying with the ticket types in use, where the elasticities were estimated at a point of time so the distinction between short-run and long-run elasticities was not examined. The price elasticity found in this study and Hensher's finding, which both use disaggregate data, are generally smaller than the international evidence based on aggregate data, and the difference between the short-run and long-run elasticity in this study (32 percent) is also smaller than international evidence which has generally found that the long-run price elasticity is two to three times larger than short-run elasticities (Voith, 1991; Dargay and Hanly, 2002; Bresson et al., 2003; Graham et al., 2009; Dargay et al., 2010), regardless the type of functional form in use. This is possibly because the price variation in a pseudo panel dataset for a specific study area is smaller than would be exhibited using a panel data analysis of aggregate

data from multiple transport systems. This difference shows the demand elasticity of price may be sensitive to the methodology in use, and the elasticity estimated from the pseudo panel approach is representative of the focussed study area which is more relevant for local transport planning.

The age elasticity is -0.57 in the short run and -0.76 in the long run. The age elasticities appear to be high since these suggest a one-hundred percent increase in age gives a significant change over a life cycle. For example, students aged around 20 years old with high public transport demand will become middle-age people in the workforce after a one-hundred percent increase in age, who are expected to have a lower usage of public transport in the context of Sydney. Thus, public transport demand in Sydney is very sensitive to age in terms of percentage changes.

The two land use variables, population density and pseudo nodes, also have moderately high elasticities because a one-hundred percent change in population density and pseudo nodes indicates a dramatic changes in land use, so population density and number of pseudo nodes have strong impacts on public transport demand in terms of percentage changes, and the magnitudes of the impacts are greater than price, income, and bus frequency.

Comparing the elasticities estimated from the dynamic model with the static model shows that the elasticities of the static model are higher than those short-run elasticities of the dynamic model. Some, such as income and bus frequency, have even greater elasticities in the static model than those long-run elasticities estimated from the dynamic model. This suggests that the static model, which does not take the lagged demand adjustment into account, is likely to over-estimate the demand elasticities for some explanatory variables.

The differences between short-run and long-run demand elasticities discussed above confirm that public transport users do take time to change their travel behaviour in response to changes in the explanatory variables. This implies that failing to recognise the long-run travel behaviour may mislead policy formulation

and implementation by mistakenly under-estimating the influence of system changes on long-run demand. For example, if a public transport fare is increased by one hundred percent on average, the public transport demand is expected to decrease by 22 percent in the first year according to the short-run price elasticity, but there is another seven percent reduction in public transport demand which is expected to occur in following 1.13 years. If this long-term effect is neglected by the policy maker, the influence of fare increases on changes in passenger volume or the adjustment of service level would be under-estimated.

Income of travellers is an important socio-economic factor which influences the impact of public transport system changes. Dargay (2001) has shown that demand elasticity would be different across people with different levels of income. By classifying all the observations (i.e., cohorts) in the pseudo panel dataset into two income levels based on the median annual personal income (AU\$28,825) of all observations, the elasticities can be estimated separately for these two clusters. Table 7.8 summarises the descriptive statistics of cohorts with lower income and higher income. It can be observed that cohorts with lower income have higher public transport use at 0.54 public transport trips per person as opposed to 0.37 trips per person for higher income cohorts. The average public transport trip price is also lower for lower-income cohorts at 1.40 dollars as opposed to 2.06 dollars for higher income cohorts because concession tickets and school buses are more commonly available for people with lower income such as students and retired people. Other land use variables and bus frequency do not show substantial differences between the two clusters which suggests that the effects of land use characteristics and bus frequency is not distinguishable between low-income and high-income residential areas in Sydney.

Table 7.9 displays the estimation results of the dynamic pseudo panel data model in the double-log functional form for lower income cohorts and higher income cohorts separately.

Table 7.8 Descriptive Statistics of Low Income and High Income Cohorts Divided by the Median Income

Variable (Units)	Low Income			High Income		
	Obs	Mean	S.D.	Obs	Mean	S.D.
PTTRIP (Trips/person)	128	0.54	0.31	128	0.37	0.20
PRICE (AU dollars)	128	1.40	0.60	128	2.06	0.33
INCOME (AU dollars p.a.)	128	17494.55	6170.47	128	39790.32	7053.93
AGE (years)	128	42.72	23.19	128	39.92	9.13
BUS FREQUENCY	128	193.84	159.33	128	193.63	140.20
DENSITY (population/800 m ²)	128	21623.20	5475.54	128	22531.67	5690.77
PSEUDO NODES (Nodes)	128	1387.45	637.83	128	1336.83	609.67
LAND MIX (Entropy)	128	0.12	0.01	128	0.13	0.01
DISTANCE TO PT STOP (meter)	128	246.87	84.63	128	235.40	64.44
PT STOPS (no. of stops)	128	40.77	7.80	128	42.13	7.32

Table 7.9 Estimation Results classified by Personal Income

	Lower Income	Higher Income	Two-sample t-test presented by t-value
LAG1	0.193* (0.101)	0.283*** (0.0914)	-3.39
PRICE	-0.234** (0.114)	0.0621 (0.183)	-25.8
INCOME	-0.0677 (0.105)	-0.109 (0.141)	3.09
AGE	-0.594*** (0.137)	-0.483*** (0.119)	-2.50
BUS FREQUENCY	0.175** (0.0726)	0.103 (0.0763)	6.46
POPULATION DENSITY	0.457** (0.228)	0.828*** (0.211)	-4.89
PSEUDO NODES	-0.350** (0.167)	-0.569*** (0.150)	4.18
LAND MIX	-0.0836 (0.173)	0.144 (0.189)	-11.44
DISTANCE TO PT STOP	0.146 (0.104)	0.0153 (0.0900)	14.07
PT STOPS	0.155 (0.223)	-0.404* (0.214)	13.75
CONSTANT	-2.189 (2.687)	-1.003 (2.679)	
R-squared	0.864	0.884	
Adjusted R-squared	0.851	0.873	

Note: Standard errors in parentheses; * P<0.10, ** P<0.05, *** P<0.01;

Double log models are employed and estimated by the OLS estimator

Both estimations show good model goodness-of-fit although the number of observations is reduced for each of them as compared to the estimation for the full dataset (see Table 7.6). Two-sample student's t-tests are undertaken to compare the parameters of the two models, with result showing the differences between the parameters are significant. The most noticeable difference between the two sets of parameter estimates is the coefficients for the lagged dependent variable and the price variable. The lower income cohorts are more sensitive, than higher income cohorts, to price variations but less elastic in their behaviour for the previous time period. The price variable is insignificant for the higher income cohorts suggesting that people with higher income are not sensitive to public transport price changes. Instead, their travel behaviour in terms of public transport use is more dependent on their previous behaviour as captured by the lagged dependent variable.

The influence of bus frequency on public transport demand is different between the two income clusters. Bus frequency is significant for the lower income cohorts but insignificant for higher income cohorts. This is possibly because people with lower income have higher dependency on public transport than people with higher income, given that the average number of public transport trips per person is higher in the lower income cohorts than higher income cohorts as shown in Table 7.8. Other variables do not show substantial difference in terms of the parameter significance.

This analysis shows the behavioural differences in people with difference income. People with lower income choose to use public transport because of its lower monetary costs and better level of service in terms of frequency, whereas the reason for people with higher income choosing to use public transport is more related to their previous behaviour, maybe as a result of habit or other factors that require them to use public transport such as inconvenient or inability of parking at their destination. This provides important information for public transport policy in terms of the impact of fare increase on travellers with difference income status. For example, if the operator or policy maker needs to predict the reduction of passenger volume due to a fare increase, the prediction is

expected to be more accurate if customers are segmented by income with specific care being taken to separate lower income from the overall population since they are more sensitive to fare changes.

7.5 Summary and discussion

This chapter presents a comprehensive analysis for the dynamic pseudo panel data model. Given the various forms of dynamic models, the best dynamic public transport demand is justified using statistical tests from models with different model specifications and estimation techniques. These suggest that the double-log model using the pooled OLS estimator determines the most plausible estimation results from the dynamic pseudo panel data model for this study.

Public transport demand elasticities are computed using the preferred model. The price elasticity is -0.22 in the short run and -0.29 in the long run which are similar to previous studies conducted in Sydney. Other explanatory variables including age, income, bus frequency, and land use characteristics in terms of population density and pseudo nodes are also demonstrated to be significantly influential to public transport demand. The distinction between short-run and long-run provides important policy implications by highlighting the necessity of realising the long-term effects in response to system changes which are sometimes neglected by policy makers.

In distinguishing the short-run and long-run effects, this study employs the dynamic PAM which only takes account of the first lag of dependent variable in the model prediction process. The inclusion of the second lag in the dynamic model was also estimated but not presented in this chapter because the second lag parameter is not significantly different from zero. The results from the PAM imply that the differences between the short-run and long-run elasticities are the same across all the explanatory variables as a result of being determined by the coefficient of the lagged dependent variable.

The use of pseudo panel data approach has closed some research gaps in this field of study as discussed in Section 2.6. First, the dynamic pseudo panel data

model take account of the temporal effect of behaviour changes in the public transport demand model. Second, this public transport demand model incorporates a comprehensive set of land use variables covering the land use 3D and accessibility measures. Third, the use of pseudo panel data allows a longitudinal analysis on a specific study area, with a certain level of individual information being incorporated. This allows the identification of demand elasticities with respect to people with different socio-economic status such as income. The research findings discussed above collectively contribute to the literature of transport demand analysis and land use studies, and also provide practical policy suggestions for long-term transport and urban planning.

The model specification and model form of the dynamic public transport demand model is determined in this chapter. The predictive power of this model can be examined by conducting demand forecasting, which also demonstrates the usefulness and applicability of this pseudo panel data model in empirical transport planning. Demand forecasting for the SGMA using the dynamic pseudo panel data model is introduced in the next chapter.

CHAPTER 8 DEMAND FORECAST

8.1 Introduction

To demonstrate the applicability of the dynamic public transport demand model presented in Chapter 7, this chapter employs the demand model to forecast public transport demand for the Sydney Greater Metropolitan Area (SGMA). This process validates the demand model by comparing the forecast demand and the observed public transport demand in the HTS report in the past years.

In Section 8.2, the public transport demand model is validated by estimating data between 1997 and 2007 to compare the predicted demand and the observed demand in 2008 and 2009. Section 8.3 projects the public transport demand predictors for future years using various data sources in the preparation for the demand forecasting. Section 8.4 forecasts future public transport demand for the SGMA using the best dynamic pseudo panel data model selected in Chapter 7. Section 8.5 presents sensitivity analysis to investigate the potential public transport demand growth based on various policy scenarios. Section 8.6 concludes this chapter.

8.2 Model validation

One of the difficulties in assessing the accuracy of a demand forecast is the lack of information about actual demand in the future to be compared to the forecasted demand. In principle, the accuracy of a travel demand forecast can be evaluated by comparing differences between the forecast demand estimated from a demand model and the actual demand observed from real data. However, the demand model constructed in this study is based on the Sydney Household Travel Survey (SHTS) data from 1997 to 2009, and there has only been one more year of data in 2010 released by Bureau of Transport Statistics that could be used to compare with the demand forecast. Hence, before forecasting demand for future years, the demand model is first validated through taking a holdout sample approach using data between 1997 and 2007 in the pseudo panel dataset to estimate the public transport demand in 2008 and 2009. These results are compared to the demand observed in 2008 and 2009.

The estimation results of the model based on 1997 and 2007 data and the dynamic pseudo panel model presented in Chapter 7 are compared in Table 8.1. Both models have exactly the same functional form and explanatory variables in natural logarithms. The only difference is that one model only uses data between 1997 and 2007 out of the whole dataset for estimation (the holdout model). The estimation results show that there is little difference between the two models in terms of the model goodness-of-fit and the significance of the parameters. The two-sample t-test also confirms that the differences of the parameters between the two models are insignificant. The coefficients of the parameters of the holdout model are slightly different to the original model but their significances and signs remain the same. This demonstrates that the dynamic pseudo panel data model is well specified as the estimation results are not sensitive to the data from different time periods.

The public transport demand in 2008 and 2009 is predicted and compared to the demand observed in the original dataset. Table 8.2 summarises the mean values of the explanatory variables in real terms in 2008 and 2009, with the predicted public transport demand (i.e., *PTTRIP*, number of trips per person) in comparison to the observed demand in 2008 and 2009. In 2008, the predicted public transport demand is 0.32 trips per person as compared to 0.34 for the observed demand which is around six percent lower. In 2009, the difference is larger at around 11 percent of demand under-estimated. This larger prediction difference is considered to be as a result of unexpected changes in explanatory variables between 2008 and 2009 rather than prediction errors. Looking at the mean values of explanatory variables in 2009, some variables have lower values than might be expected from historical averages. For example, price, income, and population density have decreased in 2009 as compared to 2008 although their average trends are increasing between 1997 and 2009 (as shown in Section 8.3 below). As a result, this variation from the historical trend is likely to lead to larger prediction errors. These prediction differences, as a result of using the observed values of explanatory variables, is expected to be mitigated when using projected data for demand forecasting as demonstrated in the next section.

Table 8.1 Model Estimation Results Using Data from 1997-2009 and 1997-2007

Variable	Model 1997-2009	Model 1997-2007	Two-sample t-test presented by t-value
LAG1	0.245*** (0.067)	0.225*** (0.076)	0.43
PRICE	-0.219*** (0.076)	-0.251*** (0.086)	0.43
INCOME	-0.160** (0.062)	-0.142** (0.070)	0.57
AGE	-0.573*** (0.086)	-0.604*** (0.096)	0.18
BUS FREQUENCY	0.148*** (0.051)	0.208*** (0.062)	0.57
POPULATION DENSITY	0.596*** (0.152)	0.528*** (0.176)	0.49
LAND MIX	-0.028 (0.121)	-0.123 (0.138)	1.04
PSEUDO NODES	-0.458*** (0.109)	-0.411*** (0.127)	0.47
DISTANCE TO PT STOP	0.0679 (0.066)	0.075 (0.075)	0.96
PT STOPS	-0.174 -0.151	-0.191 -0.176	1.09
Constant	-0.164 -1.645	-0.352 -1.901	1.50 0.43
Observations	236	196	
R-squared	0.877	0.876	
Adjusted R-squared	0.872	0.869	

Note: Standard errors in parentheses; * P<0.10, ** P<0.05, *** P<0.01; double log modes are estimated by the OLS estimator.

Table 8.2 Predicted Public Transport Demand for 2008 and 2009

Variable	Unit	2008 Mean	2009 Mean
LAG1	trips/person	0.32	0.34
PRICE	AU dollars	1.46	1.43
INCOME	AU dollars	23,966.21	21,687.71
AGE	years	41.59	42.21
BUS FREQUENCY	services	100.59	102.82
POPULATION DENSITY	populations	20,119.12	19,271.70
PSEUDO NODES	nodes	1,112.15	1,045.48
LAND MIX	entropy	0.12	0.13
DISTANCE TO PT STOP	meter	146.29	158.89
PT STOPS	stops	38.35	37.65
Predicted PTTRIP (\hat{y})		0.32	0.33
Observed PTTRIP		0.34	0.37
Difference (%)		-6%	-11%

8.3 Projection of predictors

The first step of demand forecasting is projecting the predictors forward. As the dynamic pseudo panel data model is estimated using data from 1997 to 2009, 2009 is selected as the base year for forecasting future demand. The predictors, which are the explanatory variables in the dynamic model, are projected for 2010, and 2011 and then to 2026 in five year intervals using various data sources as summarised in Table 8.3. Public transport demand is forecast for 2011 to 2026 in order to be compared with population forecast years and Australian Census years with the same timeframe. Public transport demand in 2010 is also forecast using the demand model given that the number of public transport trips has been observed and published, so a comparison between the observed demand and forecast demand in 2009 and 2010 can be used as another approach to assess the accuracy of the forecast model.

Table 8.3 Projections of Predictors for Demand Forecasting

Variable	Annual % change	2009-2026 total change	Data Source
PRICE	1.03%	19%	ABS (2012b)
INCOME	1.90%	38%	ABS (2012a)
AGE	0.50%	9%	ABS (2008)
BUS FREQUENCY	0.90%	16%	BTS (2012c)
POPULATION DENSITY	1.40%	24%	BTS (2012d)
PSEUDO NODES	0%	0%	Assumed to be time-invariant
LAND MIX	0%	0%	Assumed to be time-invariant
DISTANCE TO PT STOP	0%	0%	Assumed to be time-invariant
PT STOPS	0%	0%	Assumed to be time-invariant

The projection of public transport price uses the Urban Transport Fare Index of New South Wales, published by Australian Bureau of Statistics (ABS) (2012b). This index is a subgroup of the Consumer Price Index (CPI) for which historical data are also available. As there is no specific methodology for predicting future public transport price, the historical average percentage increase in Urban Transport Fare Index from 1997 to 2009 is used as the average annual price change (1.03 percent per year) for the forecast years, with all indices being adjusted to real terms based on 1997 CPI.

Annual person income is projected forward using the historical weekly income released by ABS (2012a). This weekly income is equally weighted to the annual incomes for each year between 1997 and 2008. This historical trend shows that, on average, annual income has increased in NSW by 3.86 percent in money terms, which is slightly higher than the average increase of the Australian CPI at 2.64 percent. This is converted to an average of 1.9 percent in real terms.

The growth of the age variable is not as substantial as that of the income and price variables. According to the Australian Historical Population Statistics published by ABS (2008), the median age in NSW has been increasing by around 0.5 percent per year since 1998, from 35.2 years in 1998 to 36.9 in 2007. In this study this is used as the future age increase for demand forecasting.

The three variables discussed above are projected on the basis of historical statistics collected by ABS. These ABS statistics are based on the geography of NSW state. As no further level of aggregation is publicly available, the projections for the SGMA are assumed to be the same as for NSW.

On the other hand, bus frequency and population density are projected forward using the SGMA data forecast by the Bureau of Transport Statistics (BTS). BTS forecasts travel demand by trip mode for the SGMA using the Strategic Travel Model. Bus frequency and population density are two inputs for this Strategic Travel Model and are used in this study. The projection of bus frequency can be retrieved from Transport Supply and Demand Forecasts for the Greater Metropolitan Area published by BTS (2012c), in which the bus frequency is assumed to increase by around 0.9 percent per year between 2006 and 2036. The increase of population density is assumed to be proportional to total population growth in the SGMA, since land area is fixed over time. The population growth in the SGMA is forecast by BTS (2012d) based on 2006 Census data for each five-year interval between 2006 and 2036. This population forecast is non-linear and estimated by taking account of various factors such as the supply of dwellings, birth and deaths rates, and migration flows. Populations between each two forecast years are linearly weighted. For this study, the average population growth is estimated at 1.4 percent per year between 2009 and 2026.

Other variables including pseudo nodes, land use mix, walk distance to the nearest public transport stop, and number of bus stops are assumed to be time-invariant in the forecast model. This is because they are time-invariant variables in the pseudo panel dataset and there is no historical data or forecast data available to estimate a reasonable increase rate. However, these are subject to sensitivity tests in Section 8.5 to investigate how public transport demand might change in response to the changes in these time-invariant variables as this relevant for policy analysis.

8.4 Public transport demand forecast

Based on the projected variables introduced above, public transport demand in the SGMA for future years is forecast using the double-log dynamic pseudo panel data models. The unrestricted model and the restricted model which excludes the insignificant variables are both employed. Public transport demand is first predicted for 2009 as the base year demand, followed by 2010, 2011, and then every five years until 2026.

As the dynamic demand model of Chapter 7 defines the dependent variable as number of public transport trips per person per day, this is aggregated to total public transport demand on an average day in the SGMA in 2009 by multiplying the predicted number of trips per person (0.336 trips in 2009) by the total population of the SGMA (5,317,330 persons in 2009). The daily public transport demand is then multiplied by 365 to estimate the annual public transport trips in the SGMA in order to compare the annual number of public transport trips published in the HTS report (BTS, 2012e). Public transport demand for future years is forecast using the dynamic models based on the projected data in future years.

Table 8.4 summarises these forecasted results, based on unrestricted and restricted dynamic model separately, and compares these to the reported demand published by BTS (2012e). The reported demand published by BTS is estimated from the SHTS data by expanding the sample observations to the total population of the SGMA through a weighting scheme (2011c). This statistic is published annually and the most recent data available is in 2010.

Table 8.4 Results of Annual Public Transport Demand Forecast

Year	Reported Demand (HTS report)		Forecast Demand by unrestricted model			Forecast Demand by restricted model		
	PT Trips (million)	Growth ¹	PT Trips (million)	Growth ¹	Difference ²	PT Trips (million)	Growth ¹	Difference ²
2009	606.7	N/A	651.5	N/A	7.32%	651.8	N/A	7.43%
2010	624.4	2.90%	659.6	1.29%	5.64%	659.1	1.12%	5.57%
2011	N/A	N/A	675.8	2.47%	N/A	673.3	2.14%	N/A
2016	N/A	N/A	714.3	5.69%	N/A	708.3	5.20%	N/A
2021	N/A	N/A	756.1	5.85%	N/A	745.4	5.24%	N/A
2026	N/A	N/A	797.6	5.49%	N/A	781.9	4.90%	N/A

¹As compared to the demand forecast for the previous time period (2009-2010-2011-2016-2021-2026)

²As compared to the reported demand based on the SHTS report published by BTS (2012e).

Comparing the reported demand and forecast demand in the base year 2009, the forecast demand is higher than the reported demand by 7.32 percent for the unrestricted model and 7.43 percent for the restricted model. The difference between the unrestricted and restricted models is small (651.5 million for the unrestricted model and 651.8 for the restricted model). The unrestricted model is chosen as the preferred model for demand forecasting, given that its forecast results are closer to the HTS report in terms of the annual growth and the difference between the forecast demand and the reported demand.

This difference between the reported demand and the forecast demand by the unrestricted model is around 7.32 percent in 2009 and 5.64 percent in 2010. This difference could result from the weighting scheme used by BTS which is different to the aggregation process of this analysis. In addition, the forecast demand is based on the selected sample of this study, which is constituted of public transport users only. It is possible that people may stop using public transport or there are new public transport users in the future, and this effect is unable to be captured by this forecast model.

The validity of the demand forecast can also be evaluated by comparing the demand changes over time between the reported demand and forecast demand shown in Table 8.4. The growth rate of the reported demand between 2009 and 2010 is 2.90 percent, which is close to the growth rate of the forecast demand using the unrestricted model of 1.29 percent. The estimated demand changes

over time correspond to the reported demand changes which suggest that the demand model has sound forecasting power. The growth rate of the forecast demand between 2011 and 2026 is around five percent to six percent at a five-year basis and, although the reported demand is unknown for future years, this forecast demand is similar to the forecast demand estimated by the Strategic Travel Model conducted by BTS (2012c), which also predicts a growth rate of around six percent every five years for the same timeframe. This evidence validates the forecasting power of the dynamic pseudo panel data model.

8.5 Sensitivity analysis

As there are uncertainties of future changes in explanatory variables as well as there being some time-invariant variables in the model, sensitivity analysis is conducted to forecast public transport demand based on various scenarios. These scenarios are designed by adjusting the projections of explanatory variables based on the base scenario as presented in Table 8.3. Only variables that are considered to be adjustable by policy makers are selected for this sensitivity analysis, which include the public transport price, bus frequency, population density, and pseudo nodes. Income and age are not included because it is less likely to adjust future changes of personal income and age. Land use mix, the distance to public transport stop, and the number of public transport stops are also excluded since they are insignificant in the public transport demand model.

The public transport demand is forecast from 2009 to 2026 using the unrestricted dynamic model. The results of the sensitivity analysis are summarised in Table 8.5. The base scenario, which assumes that all explanatory variables will change as projected in Table 8.3 in the future, shows that public transport demand is forecast to increase by around 26.53 percent between 2009 and 2026. This growth rate is used as a baseline to be compared to the following scenarios.

Table 8.5 Sensitivity Analysis of Public Transport Demand Forecasting

(Unit: million trips)

Year	Base Scenario	Price (+0%) +1.03% ¹	Bus Frequency (+1.5%) +0.9% ¹	Density (+2%) +1.4% ¹	Pseudo Nodes (-0.5%) +0% ¹	Combined Effect ²
2009	672.0	672.0	672.0	672.0	672.0	672.0
2010	684.0	685.5	684.6	685.4	685.6	689.1
2011	704.3	707.8	705.7	700.6	707.9	709.2
2016	751.9	766.9	757.7	773.8	767.3	811.8
2021	801.0	829.3	812.0	850.2	830.0	924.5
2026	850.3	893.4	867.0	931.8	894.5	1,050.3
2009-2026 Total Growth	26.53%	32.96%	29.02%	38.67%	33.12%	56.30%
Increased Demand ³	-	6.42%	2.48%	12.14%	6.59%	29.77%

¹Percent change in the Base Scenario²The scenario that combines all the scenarios on the left³The total increased public transport demand as compared to the base scenario in 2026

The first sensitivity analysis assumes a constant public transport price over time, in comparison to the assumption of a 1.03 percent annual increase in the base scenario, whilst keeping the changes of other variables the same as the base scenario. This scenario is tested to examine how public transport demand changes in response to adjustments of public transport price. The result shows that public transport demand is expected to increase by 32.96 percent from 2009 to 2026 which is higher than the demand growth of the base scenario by 6.43 percent, indicating that the public transport demand could be increased by 6.43 percent if the price of public transport was to remain constant in the future as compared to the current level of average price increase.

The next scenario investigates to which extent public transport demand can be increased by providing more frequent bus services. This scenario assumes a higher increase in bus frequency at 1.5 percent per year as opposed to the 0.9 percent of the base scenario. The chosen scenarios are based on likely outcomes, given known policy and, although other changes in bus frequency could be tested, they are not because this is not the focus of this study. The forecasted result shows that public transport demand will increase by 29.02 percent which is slightly higher than the base scenario as expected because the long run elasticity with respect to bus frequency is positive at 0.20, and the demand increase

between 2009 and 2026 is slightly lower than the first scenario as a result of the smaller elasticity as compared to the price elasticity at -0.29 in absolute term.

The third scenario assumes an increase in population density of two percent per annum which is higher than the base scenario of 1.4 percent. This assumption is to examine the impact of increasing population density on public transport demand. A two percent annual increase in population density means that there will be a 34 percent increase from 2009 to 2026. Figure 8.1 and Figure 8.2 visualise two built environments with around 34 percent difference in population density in Sydney in the current year. Figure 8.2 represents the image of the built environment after the population density is increased by 34 percent, and shows that there are more high-rise buildings as compared to Figure 8.1. As shown in Table 8.5, when population density increases at two percent annually, public transport demand is forecast to increase by 38.67 percent from 2009 to 2006 as compared to 26.53 percent increase in the base scenario. This result shows that public transport demand is more elastic to population density than with regard to public transport price or bus frequency.



Figure 8.1 A Built Environment with Lower Population Density as a Baseline
(Population: 31,958 persons/800m²; Source: Google Earth)



Figure 8.2 A Built Environment with Higher Population Density for Sensitivity Analysis
(Population: 42,489 persons/800m²; Source: Google Earth)

Similar results can be found in the next scenario which assumes a 0.5 percent reduction in pseudo nodes. Although the number of pseudo nodes is assumed to be time-invariant in the base scenario, in the long-term it is possible to slightly change the road network by reducing cul-de-sacs and by designing grid networks in new communities to improve the walking environment and accessibility to local public transport stops. Around 8.5 percent of pseudo nodes would be reduced by 2026 as a result of an annual reduction rate of 0.5 percent. Figure 8.3 and Figure 8.4 show two built environments with eight percent difference in the number of pseudo nodes within an 800-metre buffer of a Travel Zone centroid (marked as a dot in Figure 8.3 and 8.4). The two figures do not show noticeable difference in terms of the structure of network, although Figure 8.4 has 8 percent fewer pseudo nodes. Nevertheless, this scenario results in public transport demand increasing by 33.12 percent between 2009 and 2026 which is around 6.59 percent higher than the base scenario. This suggests that public transport demand could be effectively increased by only a slight improvement in the walking environment of the built environment and without a dramatic reform of road network.



Figure 8.3 A Built Environment with more Pseudo Nodes as a Baseline
 (Number of Pseudo nodes: 3,240; Source: developed from Sydney GIS layers)



Figure 8.4 A Built Environment with Fewer Pseudo Nodes for Sensitivity Analysis
 (Number of Pseudo nodes: 2,942; Source: developed from Sydney GIS layers)

The last scenario, which combines all the scenarios introduced in Table 8.5, shows that the total public transport demand could be increased by 48.81 percent between 2009 and 2026 which is 29.77 percent higher than the base scenario. This sensitivity analysis demonstrates the way in which public transport demand can be forecast, based on various policy scenarios. A comparison of all the policy scenarios with their incremental demand growth between 1997 and 2026 is

presented in Figure 8.5. It shows that land use changes in terms of population density and number of pseudo nodes are expected to have a greater impact on public transport demand than changes in price or bus frequency, and the public transport demand could be increased by a total number of 56.3 percent if all the four policy scenarios in the sensitivity analysis are achieved.

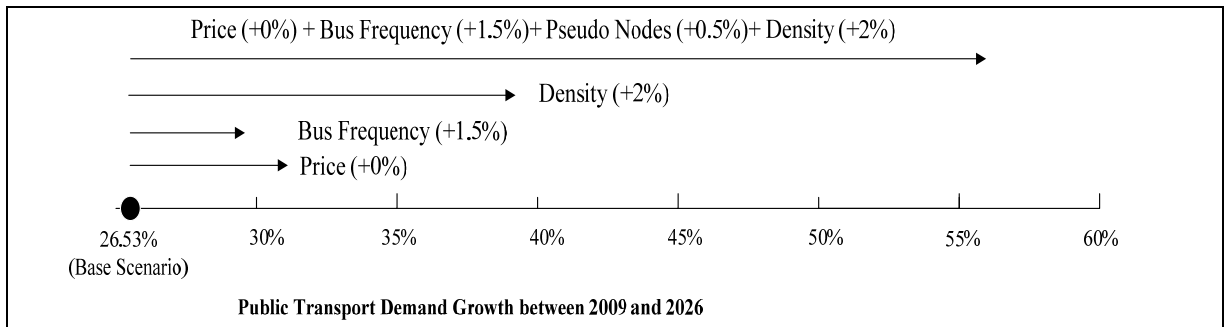


Figure 8.5 A comparison of Public Transport Demand Growth of all Scenarios

8.6 Summary

This chapter presents various demand forecasts using the dynamic pseudo panel data introduced in Chapter 7. The forecast model is first validated by investigating the difference between the observed demand and predicted demand using data between 1997 and 2007 only. The result shows that the demand model is robust to the reduction in time periods of the dataset, and it also demonstrates that the differences between the predicted demand and observed demand in 2008 and 2009 are not substantial.

The dynamic pseudo panel model based on the full dataset is then employed to forecast demand for the SGMA in future years. The forecast demand is close to the demand observed in 2009 and 2010 with similar growth rates being identified. This result confirms the validity of the demand elasticities estimated from the dynamic model which can reflect the demand changes over time in response to changes in predictors.

The sensitivity analysis of the demand forecasting demonstrates to which extent the public transport demand can be influenced by different policy scenarios. The findings suggest public transport demand is the more elastic to land use characteristics in terms of population density and number of pseudo nodes than

price changes or improvement of bus frequency. The outcomes of the demand forecasting have suggested the usefulness and potential empirical applications of the public transport demand model, given its capability of forecasting public transport demand for the SGMA.

CHAPTER 9 CONCLUSIONS

In this thesis, Chapter 1 and Chapter 2 introduce the research questions and research gaps addressed in this study. Chapter 3 through to Chapter 8 present the research work conducted to address the research questions and research gaps. This chapter summarises the research outcomes of the thesis in Section 9.1 and highlights the research contributions in Section 9.2. The limitations of this study and directions for future research are discussed in Section 9.3 and Section 9.4 respectively, followed by a concluding remark in Section 9.5.

9.1 Summary of research findings

The research findings of this study are concluded in turn in the context of research questions introduced in Chapter 1.

Question 1: What are the determinants of public transport demand and the demand elasticities with respect to each of the determinants in the SGMA?

The determinants of public transport demand in the Sydney Greater Metropolitan Area (SGMA) are investigated in Chapter 3 using the Geographically Weighted Regression (GWR) approach and Chapter 6 and Chapter 7 analysed using pseudo panel data models. The global model of the GWR analysis identifies that public transport trips price, travellers' personal income and age, bus frequency, population density, pseudo nodes, and the road distance to the CBD are significant determinants of public transport demand in the SGMA. On the other hand, the local model suggests that public transport demand consistently varies spatially with distance to the CBD and confirms the importance of spatial variability. However, the GWR analysis is here only an exploratory analysis which does not incorporate the time-series variations of public transport demand or explanatory variables which is addressed in the pseudo panel analysis.

The pseudo panel data models take account of the time-series variations in the demand model. The static and dynamic models demonstrate consistent findings in terms of the significance of the determinants. Both models confirm that public transport trip price, travellers' personal income and age, bus frequency, population density, and pseudo nodes are significant to public transport demand, whereas land use mix, number of public transport trips, and walk distance to the nearest bus stop are not significant at the 95 percent statistical confidence level. This result suggests that public transport service frequency is more important than accessibility to local public transport stops in the determination of public transport demand, and thus it needs to be considered in public transport demand modelling.

***Question 2:** Is the temporal effect of public transport demand significant in the SGMA? What are the short-run and long-run demand elasticities if the temporal effect is significant?*

The temporal effect of public transport demand is captured by the lagged dependent variable in the dynamic pseudo panel data model. The estimation results confirm the significance of the lagged dependent variable with a coefficient of 0.245, which suggests that the timeframe for public transport demand in the SGMA to reach the long-run equilibrium is around 2.13 years.

The short-run and long-run demand elasticities are summarised in Table 7.7. One of the key findings is that the price elasticity of demand is -0.22 in the short run and -0.29 in the long run, which corresponds to international evidence and a previous finding conducted in Sydney (Hensher, 1998). Elasticities of other variables suggest that public transport demand in the SGMA is expected to increase with decreasing income, lower age, higher bus frequency, higher population density, and a lower number of pseudo nodes, with evidence showing that public transport demand is the most sensitive to population density.

Question 3: *What are the magnitudes of the impact of land use density, diversity, design, and accessibility on public transport demand in the SGMA?*

The GWR and pseudo panel data analysis consistently indicate that population density and number of pseudo nodes as a measure of land use design are significant determinants of public transport demand in the SGMA, whereas other measures including land use mix and accessibility to local public transport services, as measured by the number of public transport trips and walk distance to the nearest public transport stop, are insignificant.

The magnitudes of the impact of the significant land use variables on public transport demand can be identified by the estimated elasticities. The demand elasticity with respect to population density is 0.60 in the short run and 0.79 in the long run, which implies that a ten percent increase in population density is expected to raise public transport by six percent in the first year (i.e., short run) and 7.9 percent within 2.13 years (i.e., long run). The impact of number of pseudo nodes on public transport demand is inversely related with a negative short-run elasticity of -0.46 and long-run elasticity of -0.61. The elasticities of population density and pseudo nodes are both greater than the elasticities of price, income, and bus frequency, which indicate that the influence of land use characteristics on public transport demand is fairly large as compared to the other determinants discussed under Question 1.

Land use mix, number of public transport stops, and walk distance to the nearest bus stop show insignificant results although they have been identified as influential on travel behaviour in some studies. As discussed in Section 7.3.2, land use mix is perhaps insignificant because of the low level of aggregation which does not allow for sufficient variation across space, whereas the impacts of number of public transport stops and walk distance to the nearest bus stop have been partly captured by the bus frequency measure, so these variables are found to be insignificant in the context of this study.

9.2 Research contributions

The research findings of this study do not only address the research questions, but also contribute to the literature and provide important policy implications for urban and transport planning. These research findings are summarised in this section.

9.2.1 Contributions to the literature and research methodology

As reviewed in Chapter 2, the linkage between public transport demand and land use characteristics as well as the lagged adjustments of travel demand has been identified in the literature, but not yet fully incorporated in conventional public transport demand models. The application of the pseudo panel approach applied in this study demonstrates its capability in addressing this research gap. The pseudo panel approach is comprehensively introduced in Chapter 4 and takes account of the research issues identified in previous applied pseudo panel research. One of the key contributions in this regard is applying the pseudo panel approach to a limited number of sample observations. Although the pseudo panel approach typically requires a large number of sample observations from repeated cross-sectional surveys, this study addresses the challenge of limited public transport observations in the SGMA by the way in which the individual records are equally assigned into groups and by the use of appropriate estimation technique to improve the estimation efficiency. This exercise is expected to extend the applicability of pseudo panel approach in future research.

One of the key findings with regard to the pseudo panel approach is the analysis of estimation techniques. As the debates on the appropriate use of the estimator for pseudo panel data models have been unhelpful in the literature, the Monte Carlo experiment conducted in Chapter 5 evaluates the performance of conventional panel data estimators for pseudo panel data models, under various scenarios observed from the empirical pseudo panel dataset of this study. The review and examination of the estimation techniques have been widely discussed in the discipline of econometrics in relation to panel data analysis but are less commonly evident in transport research. This simulation experiment can be used as a reference for future econometrics analysis in transport with its application

being not limited to pseudo panel data research but also to other econometric analysis with similar data properties. For example, the time-invariant variables or rarely-changing variables which have substantially larger between-group variations than within-group variations are likely to lead to estimation inefficiency for the FE estimator, and thus the pooled OLS estimator may be preferred over the FE estimator when taking account of both estimation bias and inefficiency.

The simulation experiment highlights the importance of understanding the data properties before directly choosing an estimator suggested in theory or in previous applied research. It is important to note that the preferred estimator depends on the data properties which vary with the context of study. Instead of pointing out the unambiguously best estimator, the simulation results provide useful guidelines for estimating pseudo panel data models based on different properties of data.

9.2.2 Contributions to practical urban and transport planning

The research findings of this study also suggest policy implications for urban and transport planning. In Chapter 3, the GWR analysis explores the spatial variability of the explanatory variables and its impact on public transport demand. For example, the results suggest that increasing bus frequency is expected to raise public transport demand more effectively in the outer Sydney areas than in the inner Sydney. Other variables also show moderate spatial variability across space which provides practical information for local urban and transport planning.

The use of the pseudo panel approach, based on repeated cross-sectional household travel surveys conducted in Sydney, extends the applicability of the survey data. The Sydney Household Travel Survey (SHTS) database comprises detailed individual travel information with representative sample observations of the SGMA since 1997, but suffers from the limitation that the individuals are not traced over time. Without the pseudo panel approach, this database is not able to capture the behaviour changes including the lagged demand adjustment. The

pseudo panel approach enables a longitudinal analysis using this database to identify the short-run and long-run demand elasticities for this specific study area. The distinction between short-run and long-run demand has important policy implications in predicting the short-term and long-term demand changes in responses to policy changes such as public transport fare adjustment as discussed in Section 7.4.

The demand elasticities estimated by the dynamic pseudo panel data model identify that public transport demand is more elastic to land use characteristics such as population density and the number of pseudo nodes. According to the sensitivity analysis of the demand forecast in Chapter 8, increasing population density and providing a better connectivity of road network are more likely to increase public transport demand as compared to adjusting the public transport price or frequency. This result highlights the importance of understanding the role land use planning plays in travel behaviour and provides empirical evidence for the integration of transport and urban planning in the SGMA.

The dynamic public transport demand model has also demonstrated a reasonable capability of forecasting demand, given that the forecast demand only differs from the observed demand by around 5 percent to 7 percent in 2009 and 2010, and the forecasted annual growth of demand (1.29 percent) is close to the observed demand growth (2.9 percent) reported by BTS (2012e). The forecast model with the demand elasticities can be a useful tool to forecast demand under different policy scenarios such as adjusting price or increasing population density. This model can possibly be integrated with the Strategic Travel Model run by BTS as the principle transport planning model for the SGMA.

9.3 Limitations of this study

There are several limitations or constraints which need to be discussed to understand the scope of the research, its outcomes and direction for future research in the related field of study.

One of the limitations of this study is related to the data collection. The public transport price variable in the dataset is derived from the reported ticket price in the SHTS. As discussed in Section 3.3, there are various ticket types including periodical tickets and multi-model tickets. This study applies a general multiplier to estimate the average price for each single public transport trip. This approach is not able to distinguish price elasticities among travellers with different ticket types but is applied in this study since the cohort construction of the pseudo panel data mitigates the variation of public transport price across ticket types in its aggregation. The identification of the demand elasticity of price with respect to different ticket types would require a cross-sectional analysis based on the individual data from the SHTS but this is not the focus of this study.

Another limitation with regard to data sources is land use data collection. One of the constraints in retrieving land use variables is the unavailability of household locations. Some land use variables, such as populations and number of pseudo nodes, are analysed at the TZ level by estimating populations and pseudo nodes in a buffer around a TZ centroid. It would be more ideal to use the household locations as the centroids of the buffers. Unfortunately, the household locations are not available for some of the data sources due to confidentiality issues.

The other limitation related to the land use variables is the availability of historical data. Land use data collected from Census are only available every five years, and other land use variables provided by Bureau of Transport Statistics, such as accessibility measures, are available at a single point of time. Although most of the land use characteristics are not expected to substantially change between 1997 and 2009, having the continuously historic data would allow for a more detailed investigation of the short-term and long-term impact of land use changes on public transport demand.

In terms of the pseudo panel methodology applied in this study, the constraint is the lack of public transport observations in the SHTS data due to the low usage of public transport in the SGMA. This in turn reduces the flexibility of constructing the pseudo panel dataset. For example, this analysis attempted to

separate bus and train trips in the pseudo panel data model since the two categories of travellers may have different demand elasticities with respect to some of the determinants such as price and age. However, there were insufficient trips for this study to be able to distinguish bus and train users.

For the dynamic pseudo panel data modelling, this analysis employs a Partial Adjustment Model (PAM) which takes account of the first lag of the dependent variable. The differences between short-run and long-run demand elasticities estimated from the PAM are the same for each explanatory variable because the long-run elasticities are weighted by the coefficient of the lagged dependent variable only. Including lags of independent variables in the PAM is likely to introduce strong multicollinearity and it is inconsistent with the theoretical derivation of the PAM as discussed in Section 7.2. The distinction between short-run and long-run elasticities among the explanatory variables could be possibly investigated by an Error Corrected Model (ECM) by taking account of multiple lags of independent variables in the model, but this modelling approach requires significant time-series variations in variables because it does not take account of cross-sectional variations. Hence, the PAM which estimates both the time-series and cross-sectional variations has been chosen over the ECM given the substantial cross-sectional variance in the pseudo panel dataset of this study.

The demand forecast in Chapter 8 shows plausible results, but the forecast demand is subject to sample selection bias, where only public transport users are selected in the sample. Hence, this forecast demand is not able to capture the demand changes caused by non-public transport users outside of the sample, and this is considered to be one of the factors contributing the prediction difference as compared to the HTS report.

9.4 Directions for future research

The research approaches used in this study are transferable to other study areas where repeated cross-sectional data are available. The refined pseudo panel data approach can be applied to other fields of transport studies. The above discussions on the limitations of this study already point to some directions for

future research. Further potential extensions, based on this research topic but not covered in the scope of this study, are highlighted in the following paragraphs.

The GWR analysis in Chapter 3 uses pooled data between 1997 and 2009 without taking account of the temporal effect. This is a general practice of GWR research as this methodology is very data hungry, which limits its applicability to a panel data analysis. Besides, the current version of GWR package has not incorporated panel data models or time-series models in the algorithms. The incorporation of the temporal effect in the GWR analysis would provide more information for long-term urban and transport planning.

With regard to issues of data, some further work can be done to improve the measurement of some land use characteristics. For example, land use mix is insignificant in the public transport demand models partly because of the level of aggregation and partly because of the appropriateness of the use of land use types in this context. The TZ level of aggregation appears to be too small for land use types to vary within a TZ and thus does not significantly explain the variations in public transport demand changes. Future research could attempt to use other measures such as job categories based on the TZ or household level, or as the basis of aggregating the land use types into a higher level of aggregation.

In accommodating the land use data, the correlation between various land use variables appears to be high. As a result, some variables highly correlated to others had to be removed from the dataset including employment density. The remaining variables still present a certain level of multicollinearity although it is not too substantial to distort the estimation results after model diagnostics. If those variables are of interest for strategic urban planning and they must be included in the model, the correlations will need to be controlled in the estimation process using other modelling approaches.

The other issue in land use data is the self-selection problem as reviewed in Section 2.2.3. It is possible that the self-selection problem may exist in this case

study. However, given that attitudinal data are unavailable, the degree of self-selection cannot be quantitatively identified in the context of this study. This would require separate research investigating this effect and could be an area for future research.

The other direction of future research which may be outside of transport domain is the development of a Best Linear Unbiased Estimator (BLUE) for pseudo panel data model. As shown in Chapter 5, neither of the current estimators developed for panel data model estimation is BLUE. Although the bias and inefficiency can be mitigated by better model specification as demonstrated in Chapter 7, developing an estimator that controls for the unobserved heterogeneity and the lagged dependent variable whilst taking account of cross-sectional variance would be a breakthrough not only for pseudo panel data models but also for other models with similar properties.

Last but not least, the pseudo panel data model has demonstrated its capability of forecasting demand. This demand model can be extended to other modes of travel or multi-modal demand models, especially for car travel which is the major means of travel in the SGMA. Constructing a flexible pseudo panel data model for car travel demand would be easier than for public transport demand only as there are many more observations of car trips in the SGMA. Hence, applying this approach to car travel demand analysis would present a useful tool for urban transport planning.

9.5 Concluding remarks

Research in public transport demand modelling and the interaction between travel behaviour and land use has been extensively evident in the literature, but the gap between these two fields of knowledge has been under-addressed. Longitudinal analysis on public transport demand for a specific study area based on travel survey data is not identified in the literature. This thesis applies and refines the pseudo panel approach to model public transport demand using the SGMA as a case study, whilst incorporating land use variables as well as the temporal effect of travel demand. In spite of some limitations and constraints as

discussed above, this thesis has closed some research gaps to provide a useful reference for future research in this field of study. The research outcomes of the demand elasticities and the interaction between public transport demand and land use also suggest important policy implications for the study area. This research is thus a step forward in travel demand analysis with strong potentials for related future research which is of benefit to urban transport.

REFERENCES

- Anderson, T. W. and Hsiao, C. (1981) "Estimation of dynamic models with error components". *Journal of the American Statistical Association*, Vol. 76, No. 375, pp. 598-606.
- Arellano, M. and Bond, S. (1991) "Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations". *Review of Economic Studies*, Vol. 58, No. 2, pp. 277-297.
- Australian Bureau of Statistics (2008) *Median age by sex, states and territories*. Australian Bureau of Statistics, Australia.
- Australian Bureau of Statistics (2011) *Australian Standard Geographical Classification (ASGC)*. Australian Bureau of Statistics, Australia.
- Australian Bureau of Statistics (2012a) *Average Weekly Earnings, New South Wales - Original*. Australian Bureau of Statistics, Australia.
- Australian Bureau of Statistics (2012b) *CPI: Group, Sub-group and Expenditure Class, Index Numbers by Capital City*. Australian Bureau of Statistics, Australia.
- Balcombe, R., Mackett, R., Paulley, N., Preston, J., Shires, J., Titheridge, H., Wardman, M. and White, P. (2004) *The demand for public transport: a practical guide*. Transport Research Laboratory, Wokingham, the United Kingdom.
- Baltagi, A. H. (2008) *Econometric Analysis of Panel Data*. John Wiley and Sons Ltd., Chichester, the United Kingdom.
- Batley, R., Dargay, J. and Wardman, M. (2011) "The impact of lateness and reliability on passenger rail demand". *Transportation Research Part E*, Vol. 47, No. 1, pp. 61-72.
- Beck, N. and Katz, J. N. (1995) "What to do (and not to do) with time-series cross-section Data". *The American Political Science Review*, Vol. 89, No. 3, pp. 634-647.
- Beck, N. and Katz, J. N. (2011) "Modeling dynamics in time-series-cross-section political economy data". *Annual Review of Political Science*, Vol. 14, No. 17, pp. 331-352.
- Bento, A. M., Cropper, M. L., Mobarak, A. M. and Vinha, K. (2005) "The effects of urban spatial structure on travel demand in the United States". *Review of Economics and Statistics*, Vol. 87, No. 3, pp. 466-478.
- Bernard, J. T., Bolduc, D. and Yameogo, N. D. (2011) "A pseudo-panel data model of household electricity demand". *Resource and Energy Economics*, Vol. 33, No. 1, pp. 315-325.

Bhat, C. R. and Guo, J. Y. (2007) "A comprehensive analysis of built environment characteristics on household residential choice and auto ownership levels". *Transportation Research Part B*, Vol. 41, No. 5, pp. 506-526.

Blundell, R. and Bond, S. (1998) "Initial conditions and moment restrictions in dynamic panel data models". *Journal of Econometrics*, Vol. 87, No. 1, pp. 115-143.

Bresson, G., Dargay, J., Madre, J. L. and Pirotte, A. (2003) "The main determinants of the demand for public transport: a comparative analysis of England and France using shrinkage estimators". *Transportation Research Part A*, Vol. 37, No. 7, pp. 605-627.

Breusch, T. S. and Pagan, A. R. (1980) "The Lagrange Multiplier test and its application to model specifications in econometrics". *Review of Economic Studies* Vol. 47, pp. 239-253.

Bruno, G. S. F. (2005a) "Approximating the bias of the LSDV estimator for dynamic unbalanced panel data models". *Economics Letters*, Vol. 87, No. 3, pp. 361-366.

Bruno, G. S. F. (2005b) "Estimation and inference in dynamic unbalanced panel-data models with a small number of individuals". *The Stata Journal*, Vol. 5, No. 4, pp. 473-500.

Buehler, R. (2011) "Determinants of transport mode choice: a comparison of Germany and the USA". *Journal of Transport Geography*, Vol. 19, No. 4, pp. 644-657.

Bun, M. J. G. and Kiviet, J. F. (2003) "On the diminishing returns of higher order terms in asymptotic expansions of bias". *Economics Letters*, Vol. 79, No. 2, pp. 145-152.

Bureau of Transport Statistics (2011b) *2009/10 Household Travel Survey*. Bureau of Transport Statistics, Sydney, Australia.

Bureau of Transport Statistics (2011c) *2009/10 Household Travel Survey Summary Data*. Bureau of Transport Statistics, Sydney, Australia.

Bureau of Transport Statistics (2011d) *2010/11 Household Travel Survey-Summary Transport Statistics by Local Government Area*. Bureau of Transport Statistics, Sydney, Australia.

Bureau of Transport Statistics (2011e) *Sydney Strategic Travel Model (STM): Modelling future travel patterns*. Bureau of Transport Statistics, Sydney, Australia.

Bureau of Transport Statistics (2012c) *Statistics in TransFigures: Travel Forecasts 2006-2036*. Bureau of Transport Statistics, Sydney, Australia.

Bureau of Transport Statistics (2012d) *Summary Population Forecasts 2006-2046*. Bureau of Transport Statistics, Sydney, Australia.

Bureau of Transport Statistics (2012e) *Summary Transport Statistics by Region*. Bureau of Transport Statistics, Sydney, Australia.

Campos, J., Ericsson, N. R. and Hendery, D. F. (2005) General-to-specific modeling: an overview and selected bibliography. *International Finance Discussion Papers*. Board of Governors of the Federal Reserve System, the United States.

Cao, X., Mokhtarian, P. L. and Handy, S. L. (2007) "Do changes in neighborhood characteristics lead to changes in travel behavior? A structural equations modeling approach ". *Transportation*, Vol. 34, No. 5, pp. 535-556.

Cervero, R. (2002) "Built environments and mode choice: toward a normative framework". *Transportation Research Part D*, Vol. 7, No. 4, pp. 265-284.

Cervero, R. (2006) "Alternative approaches to modeling the travel-demand impacts of smart growth". *Journal of American Planning Association*, Vol. 72, No. 3, pp. 285-295.

Cervero, R. (2007) "Transit-oriented development's ridership bonus: a product of self-selection and public policies". *Environment and Planning A*, Vol. 39, No. 9, pp. 2068 -2085.

Cervero, R. and Kockelman, K. (1997) "Travel demand and the 3Ds: density, diversity, and design". *Transportation Research Part D*, Vol. 2, No. 3, pp. 199-219.

Charlton, M. and Fotheringham, A. S. (2009) *Geographically Weighted Regression-White Paper*. National University of Ireland, Maynooth, County Kildare, Ireland.

Chow, L. F., Zhao, F., Liu, X., Li, M. T. and Ubaka, I. (2006) "Transit ridership model based on geographically weighted regression". *Transportation Research Record: Journal of Transportation Research Board*, No. 1972, pp. 105-114.

Cityrail (2010) A compendium of CityRail travel statistics. Cityrail, Sydney, Australia.

Dargay, J. (2007) "The effect of prices and income on car travel in the UK". *Transportation Research Part A*, Vol. 41, No. 10, pp. 949-960.

Dargay, J., Clark, S., Johnson, D., Toner, J. and Wardman, M. (2010) "A forecasting model for long distance travel in Great Britain". *Proceedings of the 12th World Conference on Transport Research*, Lisbon, Portugal.

Dargay, J. M. (2001) "The effect of income on car ownership: evidence of asymmetry". *Transportation Research Part A*, Vol. 35, No. 9, pp. 807-821.

- Dargay, J. M. (2002) "Determinants of car ownership in rural and urban areas: a pseudo-panel analysis". *Transportation Research Part E*, Vol. 38, No. 5, pp. 351–366.
- Dargay, J. M. and Hanly, M. (2002) "The demand for local bus services in England". *Journal of Transport Economics and Policy*, Vol. 36, No. 1, pp. 73-91.
- Dargay, J. M. and Vythoulkas, P. C. (1999) "Estimation of dynamic car ownership model: a pseudo-panel approach". *Journal of Transport Economics and Policy*, Vol. 33, No. 3, pp. 287-302.
- Deaton, A. (1985) "Panel data from time series of cross-sections ". *Journal of Econometrics*, Vol. 30, No. 1-2, pp. 109-126.
- Douglas, N. J., Franzmann, L. J. and Frost, T. W. (2003) "The estimation of demand parameters for primary public transport service in Brisbane attributes". *Conference Proceedings of the 26th Australasian Transport Research Forum*. Wellington, New Zealand.
- Du, H. and Mulley, C. (2006) "Relationship between transport accessibility and land value: a local model approach with Geographically Weighted Regression". *Transportation Research Record: Journal of Transportation Research Board*, No. 1977, pp. 197-205.
- Estupiñán, N. and Rodríguez, D. A. (2008) "The relationship between urban form and station boardings for Bogotá's BRT". *Transportation Research Part A*, Vol. 42, No. 2, pp. 296-306.
- Ewing, R. and Cervero, R. (2010) "Travel and the built environment: a meta-analysis ". *Journal of the American Planning Association*, Vol. 76, No. 3, pp. 265-294.
- Fotheringham, A. S., Brunson, C. and Charlton, M. E. (2002) *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*. Wiley, Chichester, the United Kingdom.
- Frank, L., Bradley, M., Kavage, S., Chapman, J. and Lawton, K. (2008) "Urban form, travel time, and cost relationships with tour complexity and mode choice". *Transportation*, Vol. 35, No. 1, pp. 37-54.
- Garcia-Ferrer, A., Bujosa, M., Juan, A. and Poncela, P. (2006) "Demand forecast and elasticities estimation of public transport". *Journal of Transport Economics and Policy*, Vol. 40, No. 1, pp. 45-67.
- Gardes, F., Duncan, G., Gaubert, P., Gurgand, M. and Starzec, C. (2005) "Panel and pseudo-panel estimation of cross-sectional and time series elasticities of food consumption: the case of U.S. and Polish data". *Journal of Business & Economic Statistics*, Vol. 23, No. 2, pp. 242-253.

- Gassner, K. (1998) "An estimation of UK telephone access demand using pseudo-panel data". *Utilities Policy*, Vol. 7, No. 3, pp. 143-154.
- Golob, T. F. (2003) "Structural equation modeling for travel behavior research". *Transportation Research Part B*, Vol. 37, No. 1, pp. 1-25.
- Goodwin, P. B. (1992) "A review of new demand elasticities with special reference to short and long run effects of price changes". *Journal of Transport Economics and Policy*, Vol. 26, No. 2, pp.155-169.
- Graham, D. J., Crotte, A. and Anderson, R. J. (2009) "A dynamic panel analysis of urban metro demand". *Transportation Research Part E*, Vol. 45, No. 5, pp. 787-794.
- Greene, W. H. (2000) *Econometric Analysis*. Prentice Hall, New Jersey, the United States.
- Halaby, C. N. (2004) "Panel models in sociological research: theory into practice". *Annual Review of Sociology*, Vol. 30, No. 1, pp. 507-544.
- Hayashi, F. (2000) *Econometrics*. Princeton University Press, New Jersey, the United States.
- Hensher, D. A. (1998) "Establishing a fare elasticity regime for urban passenger transport". *Journal of Transport Economics and Policy*, Vol. 32, No. 2, pp. 221-246.
- Hensher, D. A. (2008) "Assessing systematic sources of variation in public transport elasticities: some comparative warnings". *Transportation Research Part A*, Vol. 42, No. 7, pp. 1031-1042.
- Hensher, D. A. and King, J. (1998) "Establishing a fare elasticity regime for urban passenger transport: time-based fares for concession and non-concession markets segmented by trip length". *Journal of Transportation and Statistics*, Vol. 1, No. 1, pp. 44-57.
- Holmgren, J. (2007) "Meta-analysis of public transport demand". *Transportation Research Part A*, Vol. 41, No. 10, pp. 1021-1035
- Hsiao, C. (1986) *Analysis of Panel Data*. Cambridge University Press, New York, the United States.
- Huang, B. (2007) *The use of pseudo panel data for forecasting car ownership*. Doctoral Dissertation of *Birkbeck College*, University of London, London, the United Kingdom.
- Inoue, A. (2008) "Efficient estimation and inference in linear pseudo-panel data models". *Journal of Econometrics*, Vol. 142, No. 1, pp. 449-466.

Jevons, D., Meaney, N., Robins, N., Dargay, J., Preston, J., Goodwin, P. and Wardman, M. (2005) "How do rail passengers respond to change?". *European Transport Conference*. Strasbourg, France.

Judson, R. A. and Owen, A. L. (1999) "Estimating dynamic panel data models: a guide for macroeconomists". *Economics Letters*, Vol. 65, No. 1, pp. 9-15.

Kitamura, R., Mokhtarian, P. L. and Laidet, L. (1997) "A micro-analysis of land use and travel in five neighborhoods in the San Francisco Bay Area". *Transportation*, Vol. 24, No. 2, pp. 125-158.

Kiviet, J. F. (1995) "On bias, inconsistency, and efficiency of various estimators in dynamic panel data models". *Journal of Econometrics*, Vol. 68, No. 1, pp. 53-78.

Koyck, L. M. (1954) *Distributed Lags and Investment Analysis*. North Holland Publishing Company, Amsterdam, the Netherlands.

Kremers, H., Nijkamp, P. and Rietveld, P. (2002) "A meta-analysis of price elasticities of transport demand in a general equilibrium framework". *Economic Modelling*, Vol. 19, No. 3, pp. 463-485.

Litman, T. (2004) "Transit Price Elasticities and Cross-Elasticities". *Journal of Public Transportation*, Vol. 7, No. 2, pp. 37-58.

Mckenzie, D. J. (2004) "Asymptotic theory for heterogeneous dynamic pseudo-panels". *Journal of Econometrics*, Vol. 120, No. 2, pp. 235-262.

Moffitt, R. (1993) "Identification and estimation of dynamic models with a time series of repeated cross-sections". *Journal of Econometrics*, Vol. 59, No. 1-2, pp. 99-123.

Mokhtarian, P. L. and Cao, X. (2008) "Examining the impacts of residential self-selection on travel behavior: a focus on methodologies". *Transportation Research Part B*, Vol. 42, No. 3, pp. 204-228.

Mulley, C. and Tanner, M. (2009) "The Vehicle Kilometres Travelled (VKT) by private car: a spatial analysis using geographically weighted regression". *Proceedings of the 32nd Australasian Transport Research Forum*. Auckland, New Zealand.

Nerlove, M. (1958) "Adaptive expectations and Cobweb phenomena". *The Quarterly Journal of Economics* Vol. 72, No. 2, pp. 227-240.

NSW Ministry of Transport (2006). *Service Planning Guidelines for Sydney Contract regions*. NSW Ministry of Transport, New South Wales, Australia.

Nickell, S. (1981) "Biases in dynamic models with fixed effects". *Econometrica*, Vol. 49, No. 6, pp. 1417-1426.

Nijkamp, P. and Pepping, G. (1998) "Meta-analysis for explaining the variance in public transport demand elasticities in Europe". *Journal of Transportation and Statistics*, Vol. 1, No. 1, pp. 1-14.

Oum, T. H., Waters, W. G. and Yong, J. S. (1992) "Concepts of price elasticities of transport demand and recent empirical estimates". *Journal of Transport Economics and Policy*, Vol. 26, No. 2, pp. 139-154.

Parks, R. W. (1967) "Efficient estimation of a system of regression equations when disturbances are both serially and contemporaneously correlated". *Journal of the American Statistical Association*, Vol. 62, No. 318, pp. 500-509.

Paulley, N., Balcombe, R., Mackett, R., Titheridge, H., Preston, J., Wardman, M., Shires, J. and White, P. (2006) "The demand for public transport: The effects of fares, quality of service, income and car ownership". *Transport Policy*, Vol. 13, No. 4, pp. 295-306.

Pesaran, M. H. (2004) *General diagnostic tests for cross section dependence in panels*. Cambridge Working Papers in Economics, Faculty of Economics, University of Cambridge, the United Kingdom.

Pinjari, A. R., Pendyala, R. M., Bhat, C. R. and Waddell, P. A. (2007) "Modeling residential sorting effects to understand the impact of the built environment on commute mode choice". *Transportation*, Vol. 34, No. 5, pp. 557-573.

Plümpfer, T. and Troeger, V. E. (2011) "Fixed-effects vector decomposition: properties, reliability, and instruments". *Political Analysis*, Vol. 19, No. 2, 147-164.

Plümpfer, T. and Troeger, V. E. (2007) "Efficient estimation of time-invariant and rarely changing variables in finite sample panel analyses with unit fixed effects". *Political Analysis*, Vol. 15, No. 2, pp. 124-139.

Plümpfer, T. and Troeger, V. E. (2011) "Fixed-effects vector decomposition: properties, reliability, and instruments". *Political Analysis*, Vol. 19, No. 2, pp. 147-164.

Raimond, T. and Hensher, D. A. (1997) "A review of empirical studies and applications". In Golob, T. F., Kitamura, R. and Long, L. (Eds.) *Panels for Transportation Planning*. Kluwer Scientific Publishers, Boston, the United States.

Rajamani, J., Bhat, C. R., Handy, S., Knaap, G. and Song, Y. (2003) "Assessing the impact of urban form measures on nonwork trip mode choice after controlling for demographic and level-of-service effects". *Journal of Transportation Research Board: Transportation Research Record*, No. 1831, pp. 158 - 165.

Ramsey, J. B. (1969) "Tests for specification errors in classical linear least squares regression analysis". *Journal of the Royal Statistical Society*, Vol. 31, No. 2, pp. 350-371.

Reed, W. R. and Ye, H. (2011) "Which panel data estimator should I use?". *Applied Economics*, Vol. 43, No. 8, pp. 985-1000.

Rivera, M. A. I. and Tiglao, N. C. C. (2005) "Modeling residential location choice, workplace location choice and mode choice of two-worker households in metro Manila". *Proceedings of the Eastern Asia Society for Transportation Studies*. Bangkok, Thailand.

Rodriguez, D. A. and Joo, J. (2004) "The relationship between non-motorized mode choice and the local physical environment". *Transportation Research Part D*, Vol. 9, No. 2, pp. 151-173.

Simma, A. and Axhausen, K. W. (2004) "Interactions between travel behaviour, accessibility and personal characteristics: the case of the Upper Austria Region". *European Journal of Transport and Infrastructure Research*, Vol. 3, No. 2, pp. 179-198.

Sohn, K. and Shim, H. (2010) "Factors generating boardings at Metro stations in the Seoul metropolitan area". *Cities*, Vol. 27, No. 5, pp. 358-368.

Souche, S. (2010) "Measuring the structural determinants of urban travel demand". *Transport Policy*, Vol. 17, No. 3, pp. 127-134.

Sung, H. and Oh, J. T. (2011) "Transit-oriented development in a high-density city: Identifying its association with transit ridership in Seoul, Korea". *Cities*, Vol. 28, No. 1, pp. 70-82.

Tsai, C. H. and Mulley, C. (2012) "Identifying short-run and long-run public transport demand change in Sydney: a pseudo panel approach". *13th International Conference of the International Association for Travel Behaviour Research*, Toronto, Canada.

Tsai, C. H., Mulley, C. and Clifton, G. (2012) "The spatial interactions between public transport demand and land use characteristics in the Sydney Greater Metropolitan Area". *Road and Transport Research*, Vol.21, No. 4, pp. 62-73.

Tsai, C. and Mulley, C. Forthcoming (2013) "Identifying Short Run and Long Run Demand Elasticities in Sydney: A Pseudo-Panel Approach". *Journal of Transport Economics and Policy*.

Tsai, C., Leong, W., Mulley, C. and Clifton, G. Forthcoming (2013) "Examining Estimation Bias and Efficiency for Pseudo Panel Data in Travel Demand Analysis". *Transportation Research Record: Journal of Transportation Research Board*.

Verbeek, M. (1992) "Pseudo Panel Data". In Matyas, L. and Sevestre, P. (Eds.) *The Econometrics of Panel Data: Fundamentals and Recent Developments in Theory and Practice*. Kluwer Academic Publishers, the Netherlands.

- Verbeek, M. and Nijman, T. (1992) "Can cohort data be treated as genuine panel data?". *Empirical Economics*, Vol. 17, No. 9, pp. 9-23.
- Verbeek, M. and Vella, F. (2005) "Estimating dynamic models from repeated cross-sections". *Journal of Econometrics*, Vol. 127, No. 1, pp. 83-102.
- Voith, R. (1991) "The long-run elasticity of demand for commuter rail transportation". *Journal of Urban Economics*, Vol. 30, No. 3, pp. 360-372.
- Wang, Y., Kockelman, K. M. and Wang, X. (2011) "Anticipation of land use change through use of geographically weighted regression models for discrete response". *Journal of Transportation Research Board: Transportation Research Record*, No. 2245, pp. 111-123.
- Warunsiri, S. and Mcnown, R. (2010) "The returns to education in Thailand: a pseudo-panel approach". *World Development*, Vol. 38, No. 11, pp. 1616–1625.
- Weis, C. and Axhausen, K. W. (2009) "Induced travel demand: evidence from a pseudo panel data based structural equations model". *Research in Transportation Economics*, Vol. 25, No. 1, pp. 8-18.
- Wooldridge, J. M. (2002) *Econometric Analysis of Cross Section and Panel Data*. The MIT Press, the United States.
- Yee, J. L. and Niemeier, D. (1996) Advantages and disadvantages: longitudinal vs. repeated cross-section surveys. <http://ntl.bts.gov/lib/6000/6900/6910/bat.pdf>, Project Battelle.
- Zhang, L. (2011) "How built environment affects travel behavior: A comparative analysis of the connections between land use and vehicle miles traveled in U.S. cities". *World Symposium on Transport and Land Use Research WSTLUR 2011*, Whistler, Canada.
- Zhang, M. (2004) "The role of land use in travel mode choice: evidence from Boston and Hong Kong". *Journal of the American Planning Association*, Vol. 70, No. 3, pp. 344-360.

APPENDICES

Appendix 1

Table A1.1. Geographical Coverage of the Study Area by Statistical Division and Statistical Subdivision

Statistical Division	Statistical Subdivision
Sydney	Inner Sydney
Sydney	Gosford-Wyong
Sydney	Eastern Suburbs
Sydney	St George-Sutherland
Sydney	Canterbury-Bankstown
Sydney	Fairfield-Liverpool
Sydney	Outer South Western Sydney
Sydney	Inner Western Sydney
Sydney	Central Western Sydney
Sydney	Outer Western Sydney
Sydney	Blacktown
Sydney	Lower Northern Sydney
Sydney	Central Northern Sydney
Sydney	Northern Beaches
Sydney	Central Coast
Hunter	Newcastle
Illawarra	Illawarra SD
Illawarra	Nowra-Bomaderry
Illawarra	Wollongong

Source: Australian Bureau of Statistics (2011)

Appendix 2

Table A2.1. Number of Individual Records in Each Cohort by Group and by Wave in the Final Pseudo Panel Dataset

Group	Wave													Total
	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	
1	130	167	143	143	145	135	55	60	54	85	75	67	39	1,298
2	155	150	95	101	91	84	87	71	88	60	84	69	59	1,194
3	207	212	161	118	168	134	100	103	83	77	58	116	73	1,610
4	118	144	177	138	170	211	103	95	87	99	125	118	87	1,672
5	0	18	31	40	36	66	85	55	110	123	163	178	140	1,045
6	118	126	106	87	76	96	77	51	84	51	47	38	52	1,009
7	106	105	68	77	52	72	71	57	68	47	63	55	46	887
8	85	98	90	79	83	81	97	77	88	66	70	58	49	1,021
9	92	95	70	64	84	107	58	56	63	76	49	50	64	928
10	0	24	20	49	39	39	73	77	108	96	86	80	152	843
11	167	146	124	114	93	54	88	118	87	65	68	80	87	1,291
12	107	112	62	64	119	78	91	83	89	81	44	59	90	1,079
13	109	84	75	92	102	67	87	95	83	81	90	92	86	1,143
14	136	105	91	89	65	76	77	76	66	41	48	62	71	1,003
15	0	21	44	50	38	78	64	64	110	95	134	198	217	1,113
16	112	93	131	116	57	99	48	68	73	42	79	55	67	1,040
17	155	100	67	93	75	80	70	44	52	72	73	75	109	1,065
18	142	100	79	116	93	82	79	90	63	84	96	65	75	1,164
19	161	106	95	75	71	68	60	27	59	45	30	65	58	920
20	0	14	34	71	51	106	70	74	98	90	95	117	177	997
Total	2,100	2,020	1,763	1,776	1,708	1,813	1,540	1,441	1,613	1,476	1,577	1,697	1,798	22,322

Table A2.2. Number of Non-travellers in Each Cohort by Group and by Wave from the Sydney Household Travel Survey Database

Group	Wave													Total
	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	
1	77	74	69	58	59	54	29	34	28	37	32	36	31	618
2	25	33	19	24	22	24	21	18	22	36	22	27	14	307
3	35	35	24	29	28	21	21	17	22	18	25	25	11	311
4	20	26	24	26	23	23	18	19	11	17	21	20	8	256
5	0	1	2	10	6	13	11	5	13	21	15	26	23	146
6	76	99	88	62	60	45	59	52	59	48	60	40	44	792
7	23	28	29	21	25	19	27	36	17	20	29	34	18	326
8	23	34	35	29	21	17	22	28	24	25	21	24	21	324
9	13	17	17	12	21	14	22	15	18	17	8	19	16	209
10	0	4	4	4	7	12	14	17	18	14	18	17	28	157
11	152	160	136	129	120	88	95	99	88	78	102	101	75	1423
12	62	73	55	72	52	39	45	55	44	57	59	62	56	731
13	47	58	56	51	48	36	40	40	42	39	45	45	42	589
14	30	36	25	32	29	24	26	25	25	25	40	36	33	386
15	0	4	7	18	10	9	29	26	36	31	49	50	53	322
16	260	287	253	271	248	200	210	166	154	171	164	157	160	2701
17	124	134	112	124	136	88	97	80	94	123	113	124	119	1468
18	102	114	76	80	97	99	85	73	86	80	85	79	104	1160
19	58	47	50	57	49	49	38	52	41	46	41	43	51	622
20	0	9	11	23	36	31	34	41	50	49	66	61	87	498
Total	1127	1273	1092	1132	1097	905	943	898	892	952	1015	1026	994	13346

Appendix 3

**Stata code for static model simulation (Scenario 4):
(Programming codes for other scenarios can be requested from the author)**

```
global numobs 12
program xtsim1, rclass
version 11.0
drop _all
set obs $numobs
gen id = _n
gen u = rnormal(0,0.2)
gen xi=rnormal(0,0.5)
expand 13
sort id
by id: generate year=_n
gen s=_n
gen nc=rnormal(150,50) if s<156
egen nct=total(nc)
replace nc=23400-nct if s==156
gen c= u+rnormal(0,0.2/sqrt(nc))
gen e = rnormal()
gen x=xi+rnormal(0,0.2)
gen y = 0.2 + 0.8*x + c + e
xtset id year

regress y x
return scalar OLS_b1 = _coef[x]
return scalar OLS_SE = _se[x]

xtreg y x, fe
return scalar FE_b1 = _coef[x]
return scalar FE_SE = _se[x]

xtreg y x, re

return scalar RE_b1 = _coef[x]
return scalar RE_SE = _se[x]

xtpcse y x

return scalar PCSE_b1 = _coef[x]
return scalar PCSE_SE = _se[x]
end

simulate OLS_b1 = r(OLS_b1) OLS_SE = r(OLS_SE) FE_b1 = r(FE_b1) FE_SE =
r(FE_SE) RE_b1 = r(RE_b1) RE_SE = r(RE_SE) PCSE_b1 = r(PCSE_b1)
PCSE_SE = r(PCSE_SE), reps(1000): xtsim1
```

```

egen OLS_bhat=mean(OLS_b1)
egen OLS_MSE=total((OLS_b1-OLS_bhat)^2)
egen OLS_SSE=total((OLS_SE)^2)
gen OLS_CONF=100*sqrt(OLS_MSE)/sqrt(OLS_SSE)

egen FE_bhat=mean(FE_b1)
egen FE_MSE=total((FE_b1-FE_bhat)^2)
egen FE_SSE=total((FE_SE)^2)
gen FE_CONF=100*sqrt(FE_MSE)/sqrt(FE_SSE)

egen RE_bhat=mean(RE_b1)
egen RE_MSE=total((RE_b1-RE_bhat)^2)
egen RE_SSE=total((RE_SE)^2)
gen RE_CONF=100*sqrt(RE_MSE)/sqrt(RE_SSE)

egen PCSE_bhat=mean(PCSE_b1)
egen PCSE_MSE=total((PCSE_b1-PCSE_bhat)^2)
egen PCSE_SSE=total((PCSE_SE)^2)
gen PCSE_CONF=100*sqrt(PCSE_MSE)/sqrt(PCSE_SSE)

gen OLS_BIAS= OLS_bhat-0.8
gen FE_BIAS= FE_bhat-0.8
gen RE_BIAS= RE_bhat-0.8
gen PCSE_BIAS= PCSE_bhat-0.8

gen OLS_RMSE=sqrt(OLS_BIAS^2+OLS_SE^2)
gen FE_RMSE=sqrt(FE_BIAS^2+FE_SE^2)
gen RE_RMSE=sqrt(RE_BIAS^2+RE_SE^2)
gen PCSE_RMSE=sqrt(PCSE_BIAS^2+PCSE_SE^2)

sum OLS_b1 OLS_SE FE_b1 FE_SE RE_b1 RE_SE PCSE_b1 PCSE_SE
OLS_CONF FE_CONF RE_CONF PCSE_CONF OLS_BIAS FE_BIAS RE_BIAS
PCSE_BIAS OLS_RMSE FE_RMSE RE_RMSE PCSE_RMSE

```

Stata code for static model simulation (Scenario 4):

```

global numobs 12
program xtsim2, rclass
version 11.0
drop _all
set obs $numobs
gen id = _n
gen xi=rnormal(0,0.5)
gen u = rnormal(0,0.2)
expand 13
sort id
by id: generate year=_n
gen s=_n
gen nc=rnormal(150,50) if s<156
egen nct=total(nc)

```

```

replace nc=23400-nct if s==156
gen c= u+rnormal(0,0.2/sqrt(nc))
gen e = rnormal()
gen x=xi+rnormal(0,0.2)
gen y = 0.8*x + c + e if year==1
xtset id year
gen y_1=l.y
forvalues year=2/13{
    replace y=0.2*y_1+0.8*x+c+e if year==`year'
    replace y_1=l.y
}

regress y y_1 x
return scalar OLS_b0 = _coef[y_1]
return scalar OLS_SE0 = _se[y_1]
return scalar OLS_b1 = _coef[x]
return scalar OLS_SE1 = _se[x]

xtreg y y_1 x, fe
return scalar FE_b0 = _coef[y_1]
return scalar FE_SE0 = _se[y_1]
return scalar FE_b1 = _coef[x]
return scalar FE_SE1 = _se[x]

xtreg y y_1 x, re

return scalar RE_b0 = _coef[y_1]
return scalar RE_SE0 = _se[y_1]
return scalar RE_b1 = _coef[x]
return scalar RE_SE1 = _se[x]

xtpcse y y_1 x

return scalar PCSE_b0 = _coef[y_1]
return scalar PCSE_SE0 = _se[y_1]
return scalar PCSE_b1 = _coef[x]
return scalar PCSE_SE1 = _se[x]
end

simulate OLS_b0 = r(OLS_b0) OLS_SE0 = r(OLS_SE0) FE_b0 = r(FE_b0)
FE_SE0 = r(FE_SE0) RE_b0 = r(RE_b0) RE_SE0 = r(RE_SE0) PCSE_b0 =
r(PCSE_b0) PCSE_SE0 = r(PCSE_SE0) OLS_b1 = r(OLS_b1) OLS_SE1 =
r(OLS_SE1) FE_b1 = r(FE_b1) FE_SE1 = r(FE_SE1) RE_b1 = r(RE_b1)
RE_SE1 = r(RE_SE1) PCSE_b1 = r(PCSE_b1) PCSE_SE1 = r(PCSE_SE1),
reps(1000): xtsim2

egen OLS_b0hat=mean(OLS_b0)
egen OLS_b0MSE=total((OLS_b0-OLS_b0hat)^2)
egen OLS_b0SSE=total((OLS_SE0)^2)
gen OLS_b0CONF=100*sqrt(OLS_b0MSE)/sqrt(OLS_b0SSE)

```



```

egen FE_b0hat=mean(FE_b0)
egen FE_b0MSE=total((FE_b0-FE_b0hat)^2)
egen FE_b0SSE=total((FE_SE0)^2)
gen FE_b0CONF=100*sqrt(FE_b0MSE)/sqrt(FE_b0SSE)

egen RE_b0hat=mean(RE_b0)
egen RE_b0MSE=total((RE_b0-RE_b0hat)^2)
egen RE_b0SSE=total((RE_SE0)^2)
gen RE_b0CONF=100*sqrt(RE_b0MSE)/sqrt(RE_b0SSE)

egen PCSE_b0hat=mean(PCSE_b0)
egen PCSE_b0MSE=total((PCSE_b0-PCSE_b0hat)^2)
egen PCSE_b0SSE=total((PCSE_SE0)^2)
gen PCSE_b0CONF=100*sqrt(PCSE_b0MSE)/sqrt(PCSE_b0SSE)

gen OLS_b0BIAS= OLS_b0hat-0.2
gen FE_b0BIAS= FE_b0hat-0.2
gen RE_b0BIAS= RE_b0hat-0.2
gen PCSE_b0BIAS= PCSE_b0hat-0.2

gen OLS_b0RMSE=sqrt(OLS_b0BIAS^2+OLS_SE0^2)
gen FE_b0RMSE=sqrt(FE_b0BIAS^2+FE_SE0^2)
gen RE_b0RMSE=sqrt(RE_b0BIAS^2+RE_SE0^2)
gen PCSE_b0RMSE=sqrt(PCSE_b0BIAS^2+PCSE_SE0^2)

egen OLS_b1hat=mean(OLS_b1)
egen OLS_b1MSE=total((OLS_b1-OLS_b1hat)^2)
egen OLS_b1SSE=total((OLS_SE1)^2)
gen OLS_b1CONF=100*sqrt(OLS_b1MSE)/sqrt(OLS_b1SSE)

egen FE_b1hat=mean(FE_b1)
egen FE_b1MSE=total((FE_b1-FE_b1hat)^2)
egen FE_b1SSE=total((FE_SE1)^2)
gen FE_b1CONF=100*sqrt(FE_b1MSE)/sqrt(FE_b1SSE)

egen RE_b1hat=mean(RE_b1)
egen RE_b1MSE=total((RE_b1-RE_b1hat)^2)
egen RE_b1SSE=total((RE_SE1)^2)
gen RE_b1CONF=100*sqrt(RE_b1MSE)/sqrt(RE_b1SSE)

egen PCSE_b1hat=mean(PCSE_b1)
egen PCSE_b1MSE=total((PCSE_b1-PCSE_b1hat)^2)
egen PCSE_b1SSE=total((PCSE_SE1)^2)
gen PCSE_b1CONF=100*sqrt(PCSE_b1MSE)/sqrt(PCSE_b1SSE)

gen OLS_b1BIAS= OLS_b1hat-0.8
gen FE_b1BIAS= FE_b1hat-0.8
gen RE_b1BIAS= RE_b1hat-0.8

```

```

gen PCSE_b1BIAS= PCSE_b1hat-0.8

gen OLS_b1RMSE=sqrt(OLS_b1BIAS^2+OLS_SE1^2)
gen FE_b1RMSE=sqrt(FE_b1BIAS^2+FE_SE1^2)
gen RE_b1RMSE=sqrt(RE_b1BIAS^2+RE_SE1^2)
gen PCSE_b1RMSE=sqrt(PCSE_b1BIAS^2+PCSE_SE1^2)

sum OLS_b0 OLS_SE0 FE_b0 FE_SE0 RE_b0 RE_SE0 PCSE_b0 PCSE_SE0
OLS_b1 OLS_SE1 FE_b1 FE_SE1 RE_b1 RE_SE1 PCSE_b1 PCSE_SE1
OLS_b0CONF FE_b0CONF RE_b0CONF PCSE_b0CONF OLS_b0BIAS
FE_b0BIAS RE_b0BIAS PCSE_b0BIAS OLS_b0RMSE FE_b0RMSE
RE_b0RMSE PCSE_b0RMSE OLS_b1CONF FE_b1CONF RE_b1CONF
PCSE_b1CONF OLS_b1BIAS FE_b1BIAS RE_b1BIAS PCSE_b1BIAS
OLS_b1RMSE FE_b1RMSE RE_b1RMSE PCSE_b1RMSE

global numobs 12
program xtsim3, rclass
version 11.0
drop _all
set obs $numobs
gen id = _n
gen xi=rnormal(0,0.5)
gen u = rnormal(0,0.2)
expand 13
sort id
by id: generate year=_n
gen s=_n
gen nc=rnormal(150,50) if s<156
egen nct=total(nc)
replace nc=23400-nct if s==156
gen c= u+rnormal(0,0.2/sqrt(nc))
gen e = rnormal()
gen x=xi+rnormal(0,0.2)
gen y = 0.8*x + c + e if year==1
xtset id year
gen y_1=l.y
forvalues year=2/13{
    replace y=0.2*y_1+0.8*x+c+e if year==`year'
    replace y_1=l.y
}

xtabond2 y y_1 x, gmmstyle(y_1, lag(2 2)) ivstyle(x) twostep robust orthogonal
noleveleq
return scalar DGMM_b0 = _coef[y_1]
return scalar DGMM_SE0 = _se[y_1]
return scalar DGMM_b1 = _coef[x]
return scalar DGMM_SE1 = _se[x]

```

```

xtabond2 y y_1 x, gmmstyle(y_1, lag(2 2)) ivstyle(x) twostep robust orthogonal
nocons
return scalar SGMM_b0 = _coef[y_1]
return scalar SGMM_SE0 = _se[y_1]
return scalar SGMM_b1 = _coef[x]
return scalar SGMM_SE1 = _se[x]

end

simulate DGMM_b0 = r(DGMM_b0) DGMM_SE0 = r(DGMM_SE0) SGMM_b0 =
r(SGMM_b0) SGMM_SE0 = r(SGMM_SE0) DGMM_b1 = r(DGMM_b1)
DGMM_SE1 = r(DGMM_SE1) SGMM_b1 = r(SGMM_b1) SGMM_SE1 =
r(SGMM_SE1), reps(1000): xtsim3

egen DGMM_b0hat=mean(DGMM_b0)
egen DGMM_b0MSE=total((DGMM_b0-DGMM_b0hat)^2)
egen DGMM_b0SSE=total((DGMM_SE0)^2)
gen DGMM_b0CONF=100*sqrt(DGMM_b0MSE)/sqrt(DGMM_b0SSE)
egen SGMM_b0hat=mean(SGMM_b0)
egen SGMM_b0MSE=total((SGMM_b0-SGMM_b0hat)^2)
egen SGMM_b0SSE=total((SGMM_SE0)^2)
gen SGMM_b0CONF=100*sqrt(SGMM_b0MSE)/sqrt(SGMM_b0SSE)

gen DGMM_b0BIAS= DGMM_b0hat-0.2
gen SGMM_b0BIAS= SGMM_b0hat-0.2

gen DGMM_b0RMSE=sqrt(DGMM_b0BIAS^2+DGMM_SE0^2)
gen SGMM_b0RMSE=sqrt(SGMM_b0BIAS^2+SGMM_SE0^2)

egen DGMM_b1hat=mean(DGMM_b1)
egen DGMM_b1MSE=total((DGMM_b1-DGMM_b1hat)^2)
egen DGMM_b1SSE=total((DGMM_SE1)^2)
gen DGMM_b1CONF=100*sqrt(DGMM_b1MSE)/sqrt(DGMM_b1SSE)

egen SGMM_b1hat=mean(SGMM_b1)
egen SGMM_b1MSE=total((SGMM_b1-SGMM_b1hat)^2)
egen SGMM_b1SSE=total((SGMM_SE1)^2)
gen SGMM_b1CONF=100*sqrt(SGMM_b1MSE)/sqrt(SGMM_b1SSE)

gen DGMM_b1BIAS= DGMM_b1hat-0.8
gen SGMM_b1BIAS= SGMM_b1hat-0.8

gen DGMM_b1RMSE=sqrt(DGMM_b1BIAS^2+DGMM_SE1^2)
gen SGMM_b1RMSE=sqrt(SGMM_b1BIAS^2+SGMM_SE1^2)

sum DGMM_b0 DGMM_SE0 SGMM_b0 SGMM_SE0 DGMM_b1 DGMM_SE1
SGMM_b1 SGMM_SE1 DGMM_b0CONF SGMM_b0CONF DGMM_b0BIAS
SGMM_b0BIAS DGMM_b0RMSE SGMM_b0RMSE DGMM_b1CONF
SGMM_b1CONF DGMM_b1BIAS SGMM_b1BIAS DGMM_b1RMSE
SGMM_b1RMSE

```