Reinterpreting Vehicle Ownership in the Era of Shared and Smart Mobility

by

Rounaq Basu

B.Tech. in Civil Engineering
Indian Institute of Technology Bombay (2016)

Submitted to the Department of Urban Studies and Planning
in partial fulfillment of the requirements for the degrees of

Master in City Planning (MCP)

and

Master of Science in Transportation (MST)

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2019

Author .....
Department of Urban Studies and Planning
May 21, 2019

Certified by .....
Joseph Ferreira
Professor of Urban Planning & Operations Research
Thesis Supervisor

Accepted by .....
Ceasar McDowell
Professor of the Practice
Co-Chair, MCP Committee

Accepted by .....
Heidi Nepf
Donald and Martha Harleman Professor of Civil and Environmental Engineering
Chair, Graduate Program Committee

Reinterpreting Vehicle Ownership in the Era of Shared and Smart Mobility

by

Rounaq Basu

Submitted to the Department of Urban Studies and Planning
on May 21, 2019, in partial fulfillment of the
requirements for the degrees of
Master in City Planning (MCP)
and
Master of Science in Transportation (MST)

Abstract

Emerging transportation technologies like autonomous vehicles and services like on-demand shared mobility are casting their shadows over the traditional paradigm of vehicle ownership. Several countries are witnessing stagnation in overall car use, perhaps due to the proliferation of access-based services and changing attitudes of millennials. Therefore, it becomes necessary to revisit this paradigm, and reconsider strategies for modeling vehicle availability and use in this new era. This thesis attempts to do that through three studies that contribute to the methodological, conceptual, and praxis literatures.

The first study proposes a hybrid modeling methodology that leverages machine learning techniques to enhance traditional behavioral discrete choice models used in practice. The usefulness of this model to predict market shares of unforeseen choices like new mobility services is illustrated through an application to the off-peak car in Singapore. Our model significantly improves upon the market shares predicted by traditional models through an average reduction of 60% in RMSE.

The second study shifts the focus from vehicle ownership to vehicle availability in the form of mobility bundles. We leverage Singapore's unique policy environment to empirically model households' preferences for unique mobility bundles that are constructed in an ordinal fashion. This is followed by an examination of car usage within the household. Significant intra-household interaction effects are found with respect to job location, in addition to the observation of gender biases in the decision-making process.

The third study evaluates the effectiveness of car-lite policies that seek to replace private vehicle usage with shared and smart mobility services. Behavioral responses to the policy and associated market effects are modeled using an integrated land use transport simulator calibrated for Singapore. Initially favorable aggregate outcomes tend to disappear as short-term market effects set in. Although outcomes stabilize to a certain extent over the long-term, the initial characteristics of the study area are found to strongly influence the success of such policies.

Thesis Supervisor: Joseph Ferreira
Title: Professor of Urban Planning & Operations Research

Acknowledgments

First and foremost, I would like to express my gratitude to my supervisor, Prof. Joe Ferreira. In addition to the official roles of academic advisor and research supervisor, he has been a mentor to me throughout my time at MIT in every sense of the word. His guidance, advice, and trust have been truly invaluable. Thanks to Joe, I have been able to engage in exciting research opportunities, challenge myself in rigorous academic pursuits, and enjoy a rewarding teaching experience. His encouragement and support, even as I doubted myself over the ambitious dream of finishing three studies in the limited time-frame of a Master's degree, have been unwavering and critical for this thesis.

I would like to thank my reader Prof. Chris Zegras for providing valuable feedback throughout the research process. Chris supported me during an important move in my life, for which I am truly grateful. I would also like to express my appreciation for the opportunities he provided, both in and outside academia, that helped me think critically about connecting academic research and public policy. I am grateful to Prof. Amy Glasmeier for the care and kindness she has bestowed on me. It has been an absolute pleasure to work with her, and learn from her experience and knowledge.

I thank my lab-mates Roberto Ponce Lopez and Jingsi Shaw for their friendship and camaraderie, which helped me keep my head up in the academic whirlwind. I appreciate the support from my colleagues in Singapore, Chetan Rogbeer and Xiaohu Zhang, in particular. I also thank Mazen Danaf and Mathew Swisher, who were very patient and generous with their time and feedback when I wanted to bounce my ideas off them. A note of thanks is also due to Ellen Rushman and Sandy Wellford at DUSP, and Kiley Clapper at CEE, for their help with administrative issues.

I appreciate the support of the Singapore Urban Redevelopment Authority (URA), the Singapore Land Authority (SLA), the Singapore Land Transport Authority (LTA), and the Housing Development Board (HDB) in providing data and other helpful information. The research presented in this thesis was supported by the National Research Foundation Singapore through the Singapore MIT Alliance for Research and Technology's Future Urban Mobility IRG research program.

Talking of causal effects, I would be remiss to not thank my parents for their love and support over all the years of my existence. It would not be necessary to use a model to prove that they have had a strongly positive and statistically significant impact on my life.

THIS PAGE INTENTIONALLY LEFT BLANK

Contents

List of Figures

THIS PAGE INTENTIONALLY LEFT BLANK

List of Tables

Chapter 1

Introduction

Car ownership and use is expanding throughout the world. Historically, economic development has been strongly associated with an increase in the demand for transportation and particularly in the number of road vehicles. Vehicle ownership may promote work if employment opportunities and job searches are enhanced by reliable transportation (Baum, 2009). For example, vehicles may serve to reduce potential physical isolation from employment opportunities. Since the growth in vehicle ownership is continuing hand-in-hand with rapid urbanization, the strains are particularly severe in cities. Rising vehicle ownership trends have led to significant increases in negative externalities associated with transportation such as pollution and congestion. This has motivated policy-makers and researchers to examine vehicle ownership trends more closely over the past couple of decades. The primary reason for this is understandable, since having access to a vehicle increases an individual's (or their household's) travel options, leading to greater mobility. Secondary reasons for this scrutiny include the need to predict future transport investment in road infrastructure and the commercial demand for new vehicles.

1.1 Background & Motivation

National governments use car ownership models to forecast tax revenues and the regulatory impact of changes in the level of taxation (Giblin and McNabola, 2009). This type of model system examines configuration changes in the automobile market, life cycle CO_2 emissions from automobile transport, and tax revenues due to different taxation policies. Vehicle ownership models are also used by policy-makers to identify factors that affect vehicle usage

in the form of vehicle miles traveled (VMT), and therefore address the problems related to traffic congestion, gas consumption and air pollution (Millard-Ball and Schipper, 2011). Models for car ownership growth are also important to estimate the implications on energy demand and price, and on global CO_2 emissions.

We next examine how vehicle ownership is a financial burden upon low-income households, whose mobility needs are severely constrained without the ownership of private vehicles. Attitudinal factors also play a role in impacting vehicle ownership and usage. Psycho-social pressures such as automobility culture and car pride are discussed briefly. Recent studies have observed a stagnation of total vehicle usage at the national level in several industrialized countries, especially in the global North. This "peak car" phenomenon is discussed along with possible explanations such as macro-economic factors and changing behavioral patterns of millennials. Finally, we examine how emerging transportation technologies and services are rendering the notion of traditional vehicle ownership less effective, and motivating modelers to conceptualize a shared mobility paradigm.

1.1.1 Costs of vehicle ownership

While vehicle ownership is a readily available and measurable outcome, it also represents a response to the gap between desired mobility and actually realized mobility using non-car alternatives. As traditional economic factors such as household income, vehicle prices, and fuel costs affect the decision to purchase a private car, it is not hard to imagine that lower-wealth households are the worst affected by unmet mobility needs. Car-less households may react by purchasing a vehicle down the road, when they can afford it. Psycho-social pressures are also strong drivers of motorization among lower- and lower-middle income households.

Despite low incomes and financial problems, cars may be a necessity for certain urban households. Evidence for this phenomenon was shown by Curl et al. (2018), who found that car ownership rates increased in Glasgow between 2006 and 2011 undeterred by the global recession in 2008. Around 8.5% of households did not relinquish their car in spite of financial difficulties, and were termed as '*forced car owners*'. Walks (2018) found a significant relationship between automobile dependence and the relative burden of household indebtedness, particularly for automobile loans. Their examination of seven Canadian metropolitan areas point towards the increase in households' debt levels and changes in financing mechanisms for automobile purchases.

Blumenberg et al. (2018) differentiated between auto-deficit households (less than one car per driver) and zero-vehicle households in an effort to examine the cause of car deficits. Using data from the California Household Travel Survey, their study found that low-income auto-deficit households manage their travel needs by carefully negotiating the use of their privately owned vehicles, which restricts their mobility benefits. Further differentiation between choice and constraint among zero-vehicle households was examined by Brown (2017), who reported that 79% of zero-vehicle households do not own a car because of economic or physical constraints. A study in Australia by Delbosc and Currie (2012) implied that restricted and unfulfilled mobility needs due to financial constraints in purchasing vehicles would lead to activity space restriction, which in turn might cause lower psychological well-being and fewer social support networks.

Rising costs of transport, energy, and housing in metropolitan areas will certainly affect the choice of the housing-mobility bundle. To that effect, Li et al. (2018) found that rising fuel prices in Brisbane had a significant impact on residential location choice, which might affect urban spatial structure in the future. Ritter and Vance (2013) examined alternative scenarios pertaining to demographic change in an effort to examine how decreases in household size might translate into changes in car ownership at the national level. Based in Germany, their study found that the number of cars would continue to increase despite decreases in population, even in the baseline scenario.

1.1.2 Attitudes towards vehicle ownership

There is no denying the fact that attitudinal preferences and automobility cultures play a decisive role in reinforcing the paradigm of car ownership. Wells and Xenias (2015) impressed on the continued appeal of cars as personal space, which may even be enhanced by future technological changes such as AVs. Considering the housing-mobility bundle, De Vos et al. (2012) reported that residential dissonance or mismatch can occur when the preferred residential neighborhood does not match with the actual location. This affects households in realizing their preferred travel behavior patterns. If their preferred modes are unavailable in their actual neighborhood, they are forced to purchase cars despite preferring sustainable alternatives. Despite policies geared towards curbing car usage such as expensive licenses, psycho-social factors can influence vehicle purchase decisions. Using such a setting in Shanghai, Zhao and Zhao (2018) found that car pride correlates with car use and owning newer,

more expensive, and luxury cars.

Are people hard-wired to buy cars as soon as they can afford it? Not always! Focusing on a sub-section of such households with complex mobility needs (dual-income families with children) living in Gothenburg, Lagrell et al. (2018) find that these households adjust their space-time prisms for mandatory activities using proximity-oriented strategies. However, the apparent friction and perceived inconvenience due to restricted prisms for free-time activities lead the authors to question the long-term persistence of voluntary carlessness, especially with changes in family life situations.

1.1.3 The "peak car" phenomenon

Several countries in the global North have started to witness a plateau and subsequent decrease of private vehicle usage in the past couple of decades. Bastian et al. (2016) use macro-economic variables such as GDP and fuel price to explain the observed trends in car traffic in the USA, France, UK, Sweden, Australia and Germany. Kuhnimhof et al. (2013) delve into examining the international heterogeneity of causal factors in their study the USA, UK, France and Germany. Consistent with the peak car phenomenon, they find that the 1970s to the mid-1990s was a period of per-capita car travel growth, followed by a period of stagnation or decreases beginning around the turn of the millennium. While the total travel demand by drivers decreased in France and the USA, availability of modal alternatives played a much larger role in Germany and the UK. Studying trends in eight industrialized countries, Millard-Ball and Schipper (2011) find that total activity growth has halted relative to GDP, which provides hope for an accelerated decline in the energy intensity of car travel, stagnation in total travel per capita, and modal shifts back to public transport.

1.1.4 Behavioral shifts in millennials

In addition to the macro-economic changes outlined above, several studies point to the emergence of millennials (those born in the 1980s and 1990s) as an explanation of the peak car phenomenon. Anecdotes in the popular press largely agree on millennials being more likely to live in urban settings, marry later in life, have fewer kids, purchase fewer vehicles (if at all), and use alternative modes, compared to previous generations. These choices manifest themselves in different activity-travel patterns, and are important to understand

for shaping future mobility systems. Kuhnimhof et al. (2012) reported similar findings in their examination of national-level measures of car availability and car travel across the USA, UK, Germany, France, Japan, and Norway. A study by Lavieri et al. (2017) found that age, parenting status, and residential location have substantially significant effects on automobile-oriented mobility choices. While increasing urbanization and postponement of parenthood might reduce future car ownership among young adults, the increasing number of young families moving to more urbanized areas can counteract that effect (Oakil et al., 2016).

Are young adults truly turning their back on the car, or simply delaying the inevitable transition to a car-dependent lifestyle? Delbosc (2017) use the rate of driver licensing as a proxy measure for examining the car-dependent behavior of young adults in developed countries. While they find evidence of changes in youth driver licensing, supporting this trend through policy and planning is essential for continued pressure on the regime of car dominance. Knittel and Murphy (2019) advocate for subdued optimism regarding millennials' vehicle ownership and use preferences. They find little difference in preferences for vehicle ownership between millennials and previous generations after controlling for confounding variables, along with evidence of increased vehicle miles traveled (VMT). While millennials are certainly altering endogenous life choices, these choices are found to have a minimal effect on vehicle ownership.

Klein and Smart (2017) too caution planners to temper their enthusiasm in this aspect, as they find evidence of millennial decisions being driven by macro-economic factors, which could reverse in the near future. High fixed costs are the main barrier to the entry of young adults in the private automobile market. A study of South African students by Luke (2018) corroborate this and find that most students intend to purchase a car as soon as they can afford it. The view of cars being a necessity is further reinforced by the poor valuation of public and alternative transportation modes. While car-sharing services are growing in popularity in the USA and major European countries, operators have been reluctant to enter the market in most countries in the global South. McDonald (2015) warns planners and policy-makers that managing increases in automobility as millennials age and experience improvements in their economic fortunes, despite their comparatively different travel choices in the present, is going to be a challenge.

1.1.5 Moving towards a shared mobility paradigm

The emergence of services like car-sharing, bike-sharing, e-scooters and transportation network companies (TNCs) who offer ride-sharing services is transforming the traditional vehicle ownership paradigm into a shared mobility paradigm. Mobility-as-a-Service (MaaS), which is a manifestation of such a paradigm, has already been implemented in several cities and is starting to grow in popularity, especially in Europe. MaaS overcomes market segmentation by offering multiple modal alternatives at their respective marginal cost for each trip. By sharing the fixed costs among a large market base, operators can realize their investments by further encouraging the use of shared modes. While there are several studies seeking to examine the impact of one of these shared services on travel behavior, we should be concerned about studying the overall mobility landscape in its entirety. Microsimulation allows us to consider multiple modes jointly through a city-scale transport system. Using MATSim, Becker et al. (2018) find that MaaS can reduce transport-related energy consumption by 25% and increase system efficiency by up to 7%. If shared mobility acts as a substitute for public transport in lower-density regions, efficiency gains may reach up to 11%.

Firnkorn and Müller (2012) considered the perspective of an automaker entering the car-sharing market. Using the case of car2go which was launched by Daimler in 2009, they find that allocation of public space to car-sharing systems could result in a net gain of space in cities on top of environmental benefits of shared car usage. Moreover, municipal support regarding parking spaces could encourage car manufacturers to launch more schemes related to car-sharing. In general, access-based services allow consumers to avoid the burdens of ownership. Schaefers et al. (2016) used risk perception theory to examine the effects of different dimensions of risk (such as financial, performance, and social) on the intensity of access-based service usage, as well as their influence on ownership reduction. Their study of car-sharing users revealed that higher usage of access-based services were linked positively with risk perceptions, and consequently increased the likelihood of vehicle ownership reduction.

1.2 Research questions

We are motivated by the above discussion to reinterpret the meaning of vehicle ownership in this new era of shared and smart mobility. This thesis reconsiders modeling strategies

for vehicle availability and usage based on this premise. We conduct three studies that contribute to the methodological, conceptual, and praxis literatures.

1. **How can we predict market shares for new mobility services without observed data?**

Recent years have witnessed the emergence of large and rich data sets in the public realm along with the development of more accurate predictive techniques through machine learning. However, predicting the adoption of technologies like AVs and mobility-on-demand (MoD) services is challenging as these technologies are unforeseen and do not exist in current choice sets, which is where econometrics is particularly useful for applications of behavioral theory. Therefore, it is worth examining how traditional econometric frameworks can be augmented with machine learning techniques.

2. **What factors influence the choice of a mobility bundle? Moreover, how do intra-household dynamics influence who is the primary user of the car?**

While several studies have examined the determinants of household vehicle ownership, it is still unclear how these determinants change when we try to model the choice of a mobility bundle. The shift away from traditional vehicle ownership is necessitated by the proliferation and growing popularity of access-based services. Therefore, an examination of a bundle of mobility options becomes imperative to keep up with the modern era. Since households now have multiple mobility alternatives available for use, it is also important to explore the intra-household dynamics that influence the allocation of the car to a primary user. This becomes especially important in contexts where cars are extremely expensive or for lower-income households with only one car, as the car is likely to be used for trip-chaining in such scenarios.

3. **Will increased accessibility stemming from smart and shared mobility lead to a reduction in the choice of car-inclusive mobility bundles?**

It is widely agreed upon that access-based services and smart mobility will cause an increase in accessibility. However, it is unclear whether this will lead to a reduction in vehicle ownership. Some studies have examined the effect of specific services such as car-sharing or MoD on vehicle ownership in isolation, but they rely on very strong assumptions related to complete substitution. We argue that a comprehensive consideration of the gamut of the mobility spectrum is necessary, along with an examination of effects at the household

level which can then be aggregated to the metropolitan area scale. This study attempts to use behavioral econometric models embedded in an agent-based microsimulation to evaluate this research question.

1.3 Thesis organization

The remainder of this thesis is organized as follows. Chapter 2 provides a summary of several avenues of research related to vehicle ownership, and touches upon both the traditional and shared mobility paradigms. The contextual setting and data sources used in this thesis are discussed in Chapter 3. The first study of this thesis proposes a hybrid methodology to predict market shares of new mobility services in Chapter 4. The second study examines the determining factors behind the choice of a mobility bundle and usage of a car in Chapter 5. The third study evaluates the effectiveness of car-lite policies in reducing the choice of car-inclusive mobility bundles in Chapter 6. The thesis is concluded by summarizing key findings and limitations of the three studies, and suggesting avenues for future research in Chapter 7.

Chapter 2

Literature Review

This chapter provides a review of the literature on vehicle ownership along various fronts. We first present a discussion on the theoretical frameworks that some authors have used to conceptualize vehicle ownership decisions. Second, we provide an overview of different statistical methods that have been used to model vehicle ownership. This is followed by a discussion of empirical findings related to the determinants of vehicle ownership. Recognizing the emergence of shared and smart mobility, we summarize studies that examine the impacts of car-sharing systems on vehicle ownership decisions in the next sub-section. This is followed by a brief pointer to studies analyzing the impacts of autonomous vehicles on these decisions.

2.1 Conceptual frameworks

The *life-cycle approach*, or the *life-oriented approach*, argues that the consideration of inter-dependencies between long-term decisions is essential to understand various life choices (Zhang, 2017). These long-term decisions involve *location choices* (such as places of residence, education, and employment), *mobility ownership choices* (such as cars v/s long-term public transport season passes), and *personal and familial choices* (such as marriage and birth of children). This approach posits that changes in long-term choices are outcomes of a continuous process of development over the life-cycle, rather than discrete decisions.

Ownership of long-term mobility tools such as private cars or motorcycles involve a trade-off between large one-time costs for a low marginal cost at the time of usage. While the ownership of certain tools or bundles manifests itself in commitment to a certain pattern of travel behavior, this decision is influenced by other long-term choices. Studying these

dynamics and the influence of turning points in life is incredibly challenging, especially with cross-sectional data, as the direction of causality cannot be established. Therefore, one proposed data collection method is the use of *mobility biographies*.

A mobility biography is a survey that links different dimensions of life together by providing a longitudinal perspective on the dynamics of long-term mobility decisions. While data can be collected in real-time (i.e., respondents are asked to fill in the survey with a certain frequency for the entire observation period), retrospective surveys are good substitutes for shorter periods of time. Respondents are asked to recall and record their long-term decisions over the past couple of years or so in a retrospective survey, while a mobility biography is typically conducted over ten or more years. As retrospective surveys ask respondents to trace back their steps, this reliance on respondent memory holds credibility only over relatively short periods. Recollections over longer periods, such as five years or more, would be subject to recall bias, in addition to missing out on finer details associated with these decisions.

The necessity of the mobility biography approach has witnessed positive reinforcement in several contexts. Using a 20-year longitudinal survey spanning from 1985 to 2004 in Zurich, Beige and Axhausen (2012) found strong inter-dependencies between turning points in life and long-term mobility decisions during the life course. Extending this work, several authors used panel data to model the change in car ownership level in response to long-term choices. Clark et al. (2016) proposed a conceptual framework to explain the process through which the number of cars owned by households changes over time. They believe that *life events*, or turning points in life, cause a discrepancy between the satisfaction associated with the current car ownership level and a more desirable alternative. This discrepancy is modeled through a condition of *stress*, which causes travel behavior adaptation and consideration of change in car ownership level. These adjustments result in a *propensity to change car ownership level*, which can also be influenced by sudden external stimuli (such as rise in parking costs).

The same group of authors tried to validate their proposed framework by examining the factors associated with different types of car ownership level change (zero to one car, one to two cars, and vice versa). Using the first two waves (2009-2011) of the UK Household Longitudinal Survey, Clark et al. (2016) found that changes to composition of households and to driving license availability were the strongest predictors of car ownership level changes,

followed by employment status and income changes. The stressor framework was also validated by a study in the Netherlands by Oakil et al. (2014). Using a panel data set spanning 20 years, they found that strong and simultaneous relationships exist between car ownership changes, and household formation and dissolution processes.

The empirical studies highlighted in this section focus on the relationship between life events (such as birth of children and buying a car), rather than explaining the current car ownership status from a set of stationary explanatory variables. While this topic is certainly an avenue for future research, most studies relied on retrospective surveys that spanned around 20 years, which is a major drawback due to the recall bias limitation discussed earlier. Therefore, in addition to extending work on this topic, future researchers should be cautious about the survey mechanism and data collection strategies they seek to employ.

2.2 Statistical research methods

While a wide variety of vehicle ownership models can be constructed, planners and policy-makers are more concerned about models that are relevant to public sector planning. Jong et al. (2004) classified such models into nine categories focusing on the literature from 1995 to 2002.

1. **Aggregate Time Series Models:** These models use a sigmoid-shape function to represent the growth of car ownership over time as a function of income or GDP. Examples of such functions may include but are not limited to logarithmic and logistic curves. Diffusion theories related to product life cycles form the economic rationale behind the use of the S-curve.

2. **Aggregate Cohort Models:** Cohorts are created using the birth years of individuals in the current population. Evidence of the cohort effect has been reported in several Western European countries, thereby lending weight to the practice of using cohort-based models.

3. **Aggregate Car Market Models:** This category distinguishes between demand and supply of cars in the car market. Usually based on time-series data of car registration and price, annual forecasts of vehicle stock size and composition are estimated using these models.

4. **Heuristic Simulation Models:** These models use the assumption of stability of the

household's monetary budget (as a proportion of net income) over time. While they have proven useful for policy simulations, a drawback of these models is the reliance on only car costs to model car type choice without accounting for latent preferences.

5. **Static Disaggregate Car Ownership Models:** Discrete choice models dealing with the number of cars owned by a household fall in this category. It is typical of such models to consider only the demand side of the car market. Random utility theory is frequently used to construct such models through the binary and multinomial logit structures, among others.

6. **Joint Discrete-Continuous Models:** Such models involve two different disaggregate models, one for explaining car ownership and the other for explaining car use conditional on car ownership. The indirect utility model is a classic example of this family, where micro-economic theory postulates a relationship between indirect utility functions for different car ownership states and demand functions for car use.

7. **Static Disaggregate Car-type Choice Models:** Discrete choice models dealing with the choice of car type of the household, conditional on car ownership, fall in this category. While disaggregate models for the number of cars per household are used to provide inputs for multimodal transportation modeling systems, these models are used to forecast the size and composition of the car fleet.

8. **Pseudo-panel Methods:** Panel data sets are used to incorporate temporal and lagged effects in traditional discrete choice models. Inclusion of lagged variables help account for state dependence and individual-specific error components, which arise due to unobserved heterogeneity across households. Pseudo-panel data, which are based on cohort averages of repeated cross-sections, often help in overcoming problems associated with panel data, such as attrition.

9. **Dynamic Car Transactions Models with Vehicle Type Conditional on Transaction:** These models aim to extend the disaggregate modeling approach for the size and composition of the car market into the domain of dynamic models. Static disaggregate vehicle holding models provide a time path for the car fleet, which are then enhanced using hazard or Markov models.

A more recent review by Anowar et al. (2016) outlines four alternative modeling approaches that have gained widespread prominence over the past couple of decades. Their review focuses specifically on disaggregate household-level studies of vehicle ownership and use.

1. **Exogenous Static Models:** These models predict vehicle holdings at a particular point in time without considering the dynamics of vehicle evolution. This family includes standard discrete choice models (such as the binary, ordered, multinomial, and nested probit/logit models), count models (such as Poisson, negative binomial, and Poisson-lognormal regression models), and advanced discrete choice models (such as the latent class ordered and multinomial logit models).

2. **Endogenous Static Models:** Vehicle ownership or type is jointly modeled with other related decision processes, such as vehicle usage. Standard discrete choice methods can be used to analyze joint choices by defining choice alternatives as combinations of various choice levels. Another approach is to use methods that incorporate unobserved correlations across choice processes. Examples of this family include standard discrete choice models, mixed multi-dimensional choice models, discrete-continuous models, Bayesian models, copula-based models, and structural equation models.

3. **Exogenous Dynamic Models:** These models examine evolution in vehicle ownership decisions through the use of panel data sets that incorporate both cross-sectional and time-series dimensions. As mentioned earlier, the difficulty in obtaining panel data can be circumvented by combining multiple cross-sectional data sets into a pseudo-panel. Examples of such models include standard discrete choice models, duration models, and random effects models.

4. **Endogenous Dynamic Models:** Models that consider both the endogeneity between household fleet size or composition and usage decisions, and dynamics associated with the vehicle acquisition process fall in this category. Examples include copula-based joint GEV-based logit regression models, multinomial probit models, structural equation models, and simultaneous equation models.

Most studies captured in these reviews as well as our review of recent literature employ standard discrete choice models. Since this thesis considers vehicle ownership through an

exogenous static modeling framework, we will highlight findings from a couple of relevant studies that motivate our model selection in later chapters. Noting that both ordered and unordered response choice mechanisms can exist, Bhat and Pulugurta (1998) considered the multinomial logit (MNL) and the ordered logit (OL) in a comparative exercise. Their study used these models on different data sets to conclude that the MNL is both theoretically and statistically better than OL. It is worth keeping in mind that ordered responses are not consistent with global utility maximization, as they hypothesize that a latent variable representing car owning propensity drives the household's observed choices. In addition to being consistent with the global utility maximization theory, the MNL model also allows for greater flexibility in specifying parameters. Potoglou and Susilo (2008) extended this comparative analysis by also considering the ordered probit model, and reached the same conclusion that the MNL performs the best when household vehicle ownership is modeled as an exogenous process with static data. However, the increased flexibility of MNL comes at the cost of increased computation time and effort in interpreting outcomes, both of which can be quite significant in applications with a large number of covariates.

2.3 Determinants of vehicle ownership

We will focus only on discussing the determinants of vehicle ownership at a disaggregate level, as this thesis deals with household vehicle ownership rather than national vehicle fleet size[1]. The determinants of household vehicle ownership can be categorized as follows.

$$VO_{hh} = f(\boldsymbol{X_{dem}}, \boldsymbol{X_{acc}}, \boldsymbol{X_{LU}}, \boldsymbol{X_{BE}}, \boldsymbol{X_{housing}}, \boldsymbol{X_{job}}, \boldsymbol{X_{psych}}) \tag{2.1}$$

1. **Household demographics ($\boldsymbol{X_{dem}}$):** Household socio-demographic characteristics such as income, household size and composition are important determinants of vehicle ownership. Individual demographic attributes such as age, gender, education, and employment status can also play a crucial role in this decision. We discussed earlier how life events can influence the decision to purchase a car, thereby reinforcing the importance of considering the presence of children. We believe that it would be worthwhile to also explore the age of the children, as younger children are dependent on adults for transportation while teenagers are relatively

[1] A variety of macro-economic factors such as GDP, population density, growth in road density, etc. are used to model and forecast the growth of the national fleet size.

more self-sufficient. Education status can also influence this decision, as higher educational qualifications can possibly make individuals more conscious of the environmental impact of their vehicle usage. Finally, the employment status can also play a major role, as full-time workers who need to commute regularly would be more likely to purchase a car. Unemployed individuals or those who can work from home do not have similar mobility needs.

2. **Accessibility ($\boldsymbol{X_{acc}}$):** Accessibility can be considered at both the micro- and macro-levels. Micro-accessibility can be constructed with the use of built environment measures, such as the local distances to different types of amenities. These amenities may include bus stops, MRT stations, shopping malls, schools, etc. On the other hand, macro-accessibility is calculated at a more aggregate scale and reflects a zonal measure of accessibility to opportunities, usually in the form of jobs. Some techniques to calculate macro-accessibility are the cumulative opportunities, gravity-based, trip-based, and activity-based models.

3. **Land Use ($\boldsymbol{X_{LU}}$):** The level of urbanization can play a role in determining access to transport infrastructure. We would expect households in highly urbanized areas to be less likely to own a car. Moreover, land use also affects access to transport infrastructure. Low-density suburbs are likely to be isolated from public transport, while high-density mixed-use neighborhoods are more likely to experience multi-modal transportation options in close proximity.

4. **Built environment ($\boldsymbol{X_{BE}}$):** The built environment can influence the decision to purchase a car to a significant extent, as it determines the ease and availability of alternative mobility options. Measures of the built environment can include proximity to public transport stops and stations, in addition to urban form and street network characteristics.

5. **Housing ($\boldsymbol{X_{housing}}$):** While the housing location certainly matters in these decisions, the type of housing can provide insights into the wealth of the household. Respondents usually only report their individual and household incomes on surveys, but there might be other assets that result in household wealth. This is especially true of retirees, who might have negligible income but have accumulated wealth over their working life. One way to obtain a proxy for wealth is to observe attributes of the housing unit, mainly in the form of ownership (public or private), tenure (rented or owned), age, and market value. This information can be augmented with the income data to provide a measure of household

wealth, which would be very useful to construct a household budget constraint.

6. **Jobs ($\boldsymbol{X_{job}}$):** In addition to the employment status of individuals, the job location plays a role in determining car ownership and use. Attributes of the job location can be measured in both absolute and relative terms. Absolute variables would include built environment and accessibility measures at the job location, while relative variables would measure the inconvenience in making trips from the residential location to the job location. Examining characteristics of daily commute trips and considering intra-household interactions effects can help determine the primary user of the car, which is an important decision in households with multiple workers but only one car.

7. **Psychological factors ($\boldsymbol{X_{psych}}$):** Psychological factors, such as perceptions, attitudes, and habits, can also be used to explain car ownership and use. Social prestige and car pride are important determinants of such decisions, especially in certain socio-cultural contexts. Newer generations are more likely to be environment-conscious, which could influence their decisions in this regard and induce a shift to electric cars and public transport modes.

We would like to point out that it is possible for these determinants to overlap while constructing covariates, as in the case of quantifying job location effects. We discuss the above determinants from a conceptual standpoint. While operationalizing vehicle ownership models, the modeler must exercise good sense in capturing all the above effects through the use of economically relevant variables. Keeping this in mind, we summarize recent studies that examined household vehicle ownership in an exogenous fashion across a wide variety of socio-cultural contexts in Table 2.1.

Table 2.1: Summary of recent empirical studies examining determinants of household car ownership

Study	Data source	Data type	Dependent variable(s)	Independent variables	Modeling method
Clark (2009)	England & Wales	Cross-section	Own (0, 1, 2+)	HH dem, Housing	NB, DT, NN
Nolan (2010)	Ireland	Panel	Own (0, 1+)	HH dem	Dynamic random effects probit
Caulfield (2012)	Ireland	Cross-section	Own (0, 1, 2, 3+)	HH dem, LU, Commute mode	MNL
Brownstone and Fang (2014)	USA	Cross-section	Own (0, 1, 2+) & use (VMT)	HH dem, LU	Bivar OP (own) & censored Tobit (use)
Combs and Rodríguez (2014)	Bogotá (Colombia)	Pseudo-panel	Own (0, 1+)	BRT acc, BE, LU	BL
Anowar et al. (2016)	Montreal (Canada)	Pseudo-panel	Temporal change in Own (0, 1, 2, 3+)	HH dem, PT acc, LU	Gen-OL
McBride et al. (2016)	Seattle (USA)	Panel	Own (0, 1, 2, 3, 4+) & trip generation	HH dem, LU, Attitudes	LCMC
van Eggermond et al. (2016)	Switzerland	Cross-section	Own (0, 1, 2, 3) & use (VKT)	HH dem, BE, Macro acc, LU	MNL (own) & OLS (use)
Shen et al. (2016)	Shanghai (China)	Cross-section	Own (0, 1+) & commute mode	HH dem, BE, Attitudes	BL (own) & NL (mode)
Choudhary and Vasudevan (2017)	India	Cross-section	Own (0, 2W, 4W, Mix)	HH dem	MNL
Clark and Rey (2017)	Great Britain	Panel	Temporal change in Own (0, 1, 2+)	-	Spatial MC
Jiang et al. (2017)	Jinan (China)	Cross-section	Own (0, 1, 2+) & use (VKT)	HH dem, LU, BE, PT acc	MNL (own) & DHM (use)
Maltha et al. (2017)	Netherlands	Pseudo-panel	Own (0, 1, 2)	HH dem, LU	Fixed effects OL
Soltani (2017)	Shiraz (Iran)	Cross-section	Own (0, 1, 2, 3+)	HH dem, LU	NL (Car v/s no car)
Ao et al. (2018)	Sichuan (China)	Cross-section	Own (0, LC, HC, Mix)	HH dem, BE	MNL
Bansal et al. (2018)	3 cities (India)	Cross-section	Own (0, 2W, 4W, Mix) & use (VKT)	HH dem	MNL (own) & OLS (use)
Ding et al. (2018)	Washington (USA)	Cross-section	Own (0, 1, 2, 3+) & commute mode	HH dem, BE	Multilevel MNL-SEM
Le Vine et al. (2018)	China	Cross-section	Own (0 vs 1+, 1 vs 2+)	HH dem, Housing, Attitudes	BL

It is also worth mentioning that macro-economic factors such as vehicle prices and fuel costs can also influence decisions of vehicle ownership and usage. Moreover, we would expect vehicle and fuel price elasticity to vary by household type. For example, greater car dependency in rural regions can be associated with lower price elasticity of car demand. Fluctuations in these costs would also have a smaller impact on car ownership in rural regions, primarily because of the lack of suitable alternative mobility options.

Inferring general conclusions from different studies remains a challenge. Some of these challenges arise from difference in scales of analysis, varying built environment measures, different sources of travel behavior data, dissimilar control variables (if used at all), and different measures of final outcomes (Zegras, 2010). The issue of self-selection is another challenge. Households may choose a residential location according to their travel preferences, and this would influence their travel behavior endogenously. Therefore, some studies seek to model endogenous choices such as residential location and car ownership in a joint manner. Further discussion on joint modeling of vehicle ownership with other choice dimensions is provided in Section 5.2.

2.4 Impact of car-sharing systems

Car-sharing services offer users short-term access to vehicles based on a flat membership fee and a usage rate pertaining to the total time and distance traveled. The membership fee can include parking, fuel, insurance, cleaning, and maintenance and inspection expenses. These vehicles can be reserved from a fleet, and are usually parked at central locations across the city, which may be in stand-alone parking lots or near transportation hubs such as metro and train stations. One of the major advantages of car-sharing is that the fixed costs are significantly lower than those associated with vehicle ownership, thereby making such services particularly lucrative for lower-income drivers with sporadic vehicle needs and households with multiple drivers but only one car.

The modified cost structure of vehicle availability has direct implications for car ownership and usage. Katzev (2003) found that car-sharing members drive considerably less after becoming members of the service. It is also hypothesized that the use of a car-sharing service usually leads to a reduction in vehicles per capita among member households. Having enjoyed considerable success over the last decade with the proliferation of organizations like

ZipCar and car2go (among others), car-sharing systems tend to be staffed with fully electric vehicles (EVs) and plug-in hybrid electric vehicles (PHEVs), which provide several advantages including but not limited to noise reduction, zero tail-pipe emissions, and improved vehicle efficiency. Reductions of 47% in energy consumption and 65% in CO_2 emissions due to the use of EVs and PHEVs were reported by Baptista et al. (2014) using a case study of the MobCarsharing operator in Lisbon, Portugal.

Car-sharing services can be differentiated into two categories based on vehicle parking locations. The first category is the *station-based system*, an example of which is Zipcar. Dedicated parking spots are available at strategic locations throughout the city, where users can pick-up and drop-off the shared cars. The second category, as exemplified by car2go, is the *free-floating system*, where users can pick-up and drop-off cars at any location keeping local parking and traffic ordinances in mind. Further differentiation can be carried out based on vehicle return policies. In the *one-way model*, vehicles do not need to be returned to their pick-up point; however, the converse is true for the *round-trip model*. Owing to rebalancing constraints, advanced booking is available only for the round-trip model.

Car-sharing has started to receive a lot of academic attention in recent years, especially with regard to its impact on private vehicle ownership. We summarize the results of selected empirical studies on this topic in Table 2.2. Three relevant metrics are calculated:

- **Reduction in car ownership:** What share of car-sharing members have reduced their private vehicle holdings due to their membership in the car-sharing scheme?
- **Forgone purchase of private car:** How many members would buy a private vehicle if the car-sharing scheme was unavailable?
- **Replacement rate:** How many private vehicles are replaced per car-sharing vehicle?

Becker et al. (2018) argued that the first two metrics are often added to obtain combine impact on vehicle holdings in the literature, which is erroneous because that would result in double-counting of households that would shed a car due to their car-sharing membership and would, therefore, be likely to buy a car in the absence of the scheme. Following their proposed procedure, we try to interpret results reported in these studies according to the three metrics, where possible. In certain cases, we calculate these metrics based on other metrics reported in the articles.

We find significant differences in the magnitude of the impact of car-sharing on private vehicle ownership from Table 2.2. These differences occur both across system categories and spatial settings. Free-floating systems have a lower impact compared to station-based systems, perhaps owing to the uncertainty in obtaining access to a car-share vehicle in a free-floating system. While free-floating systems allow for relatively unrestricted parking locations at the end of a trip, this flexibility is unable to completely counter the trepidation and uncertainty that users face while reserving a vehicle for a trip. Unrestricted parking might also result in vehicles accumulating at trip destination hotspots such as the CBD or shopping malls, while trip origin hotspots would be closer to residential clusters where the demand for vehicles is more pronounced.

Table 2.2: Summary of empirical studies examining the effect of car-sharing on car ownership

Study	Data source	Data type	Reduction in car ownership	Foregone purchase of private car	Replacement rate
Free-floating car-sharing system					
Firnkorn and Müller (2011)	Ulm (Germany)	Stated Preference	4%	10%	-
Stasko et al. (2013)	Ithaca (USA)	Cross-section	-	18%	1:15
Martin and Shaheen (2016)	US & Canada	Cross-section	2-5%	7-10%	1:7-11
Giesel and Nobis (2016)	Berlin & Munich (Germany)	Cross-section	7%	-	-
Le Vine and Polak (2017)	London (UK)	Cross-section	37%	31% [a]	-
Becker et al. (2017)	Basel (Switzerland)	Cross-section	19%	27%	1:13
Becker et al. (2018)	Basel (Switzerland)	Panel with tracking	6%	-	-
Namazu and Dowlatabadi (2018)	Vancouver (Canada)	Cross-section	9% [a]	-	-
Station-based car-sharing system					
Katzev (2003)	Portland (USA)	Cross-section	26%	53%	-
Cervero and Tsai (2004)	San Francisco (USA)	Panel with tracking	22%	28%	1:4-6
Lane (2005)	Philadelphia (USA)	Cross-section	25%	29%	1:23
Millard-Ball (2005)	USA & Canada	Cross-section	20%	-	1:5-6
Martin et al. (2010)	USA & Canada	Cross-section	22%	25%	1:9-13
Loose (2010)	Europe	Cross-section	15.7% - 31.6%	33% - 35%	-
Engel-Yan and Passmore (2013)	Toronto (Canada)	Cross-section	29%	55%	-
Louvet (2014)	France (Paris)	Cross-section	23%	-	-
Klincevicius et al. (2014)	Montreal (Canada)	Cross-section	13% [a]	-	-
Mishra et al. (2015)	San Francisco (USA)	Cross-section	4% - 24% [a]	-	-
Giesel and Nobis (2016)	Berlin & Munich (Germany)	Cross-section	15%	-	-
Clewlow (2016)	San Francisco (USA)	Cross-section	30% [a]	37% [a]	-
Namazu and Dowlatabadi (2018)	Vancouver (Canada)	Cross-section	47% [a]	-	-
Liao et al. (2018)	Netherlands	Stated Preference	40%	20%	-

[a] Calculated by author based on results reported in article

We first examine the observed or revealed impact of car-sharing in bringing about a reduction in car ownership. While free-floating systems in Germany have a modest impact of 4-7%, station-based systems have more than twice that impact to the tune of 15%. Only free-floating systems were observed in Switzerland, but their impact stands at a considerable rate of 6-19%. The UK has the highest impact of 37% for European free-floating systems, while the Netherlands tops the leaderboard at 40% for the station-based equivalent. Examination of station-based systems across Europe yields an impact ranging from 16% to 32%. Parking restrictions in North American downtowns, which allow for only station-based systems, moderate the impacts compared to their European counterparts. We find that several studies set in American and Canadian cities estimate the impact to be in the range of 20-30%, with Vancouver being an outlier at 47%.

The stated impact of car-sharing is next examined through the second metric, where users are asked whether they would purchase a private vehicle in lieu of the car-sharing scheme. While most studies report consistent impacts across the revealed and stated metrics, there are a few outliers. North American cities like Portland, San Francisco, and Toronto have relatively high stated impacts, which are almost double the revealed impacts. This may be a manifestation of the dependence of individuals living in these cities on private cars, owing to the lack of reliable transportation alternatives among other reasons. Car-sharing services, although recently launched, provide a reliable and affordable alternative means to mobility, which explains the revealed impacts of 26-30%. However, these members would be forced to buy cars if the car-sharing service disappeared, which might force them into urban poverty. On the other hand, Netherlands has a stated impact (20%) that is half the revealed impact (40%), which can be explained by the availability of reliable transportation alternatives such as bicycles and public transport.

It is also worthwhile to examine the influence of exogenous socio-economic and demographic variables on the frequency of use of car-sharing services. Using data from the 2014-15 Puget Sound Regional Travel Study, Dias et al. (2017) found that users of these services tend to be young, well-educated, higher-income, working individuals residing in high-density areas. The importance of attitudinal preferences for environment-friendly mobility was highlighted by Clewlow (2016). These results are corroborated by Mishra et al. (2015), who also account for self-selection bias due to differences in observed characteristics of respondents. Through an examination of the effect of car-sharing service type on vehicle

ownership in Vancouver, Namazu and Dowlatabadi (2018) reported that members of round-trip or two-way services were five times more likely to reduce car ownership compared to one-way services. One-way services are more likely to be used as complements for other transportation modes, while two-way services are more likely to be used as a substitute for private cars. Results from Becker et al. (2018) also indicate that different car-sharing service types can both complement and compete with other transport modes.

2.5 Impact of autonomous vehicles (AVs)

Fully autonomous vehicles (AVs) promise a fundamental revolution in mobility by substantially reducing the generalized cost of travel. Travel is expected to become safer, cheaper, more comfortable, and more sustainable. Car travel can become more accessible to population cohorts like children, the elderly, and the disabled. While positive benefits may include a reduction of the total vehicle fleet and significant gains in road capacity, unintended consequences can result in additional travel demand (Gucwa, 2014) and a new wave of suburbanization and urban sprawl (Glaeser and Kahn, 2004). AVs are expected to impact private vehicle ownership decisions, along with other long-term decisions like location choices at both residential and firm levels. We discuss these aspects in greater detail in Section 6.2 that provides an important preface to our study exploring the impact of car-lite policies, which advocate for increased and better accessibility, on the housing-mobility bundle.

THIS PAGE INTENTIONALLY LEFT BLANK

Chapter 3

Data & Context

This chapter describes the contextual setting of the three studies undertaken in this thesis. We outline the spatial context of Singapore along with discussing relevant policy efforts undertaken by the government to address vehicle ownership and usage. The second subsection describes the various data sources used in this thesis. Additional details related to construction of variables from these data sources are also provided.

3.1 Contextual Setting

The three studies in this thesis are contextually and empirically set in Singapore, an island city-state located in South-east Asia. Although it gained independence relatively recently in 1965, it is considered as the world's most "technology-ready" nation by the World Economic Forum and the best-performing smart city on a global scale[1]. However, it is also ranked as the most expensive city to live in[2] and has a fairly high Gini income inequality index of 0.46[3]. Home to almost 6 million people (as of 2018), the Singaporean economy has continued to grow in strength with an estimated GDP of about 324 billion USD (as of 2017)[4].

The Singaporean government recognized the importance of transportation management quite early on, and instituted a three-pillar systems decades earlier that they continue to revise and improve. The vehicle ownership control system was instituted in 1972, followed by

[1] https://www.channelnewsasia.com/news/singapore/singapore-best-performing-smart-city-globally-study-10038722

[2] https://www.economist.com/graphic-detail/2018/03/15/asian-and-european-cities-compete-for-the-title-of-most-expensive-city

[3] https://www.cia.gov/library/publications/the-world-factbook/rankorder/2172rank.html

[4] https://tradingeconomics.com/singapore/gdp

the adoption of congestion pricing in 1974, and continuous expansion of the public transport system since 1987. Their proactive efforts in providing affordable and accessible transportation to all residents has resulted in the top global ranking in a study by McKinsey[5]. Using buses, Mass Rapid Transit (MRT) and Light Rail Transit (LRT) systems, the public transport network in Singapore is quite extensive and accounts for just over 67% of all trips (as of last year). Heavy investment towards betterment and extension of public transport infrastructure is a key characteristic of government policy, as they aim to reach a public transport mode share of 75% by 2030[6]. The bus and MRT networks are shown in Figures 3-1 and 3-2 respectively. Proposed extensions to the MRT network (as of 2012) have been highlighted in the figure. It is worth noting that some of those extensions, such as the Downtown Line, have been completed and are fully operational at the time of writing this thesis.

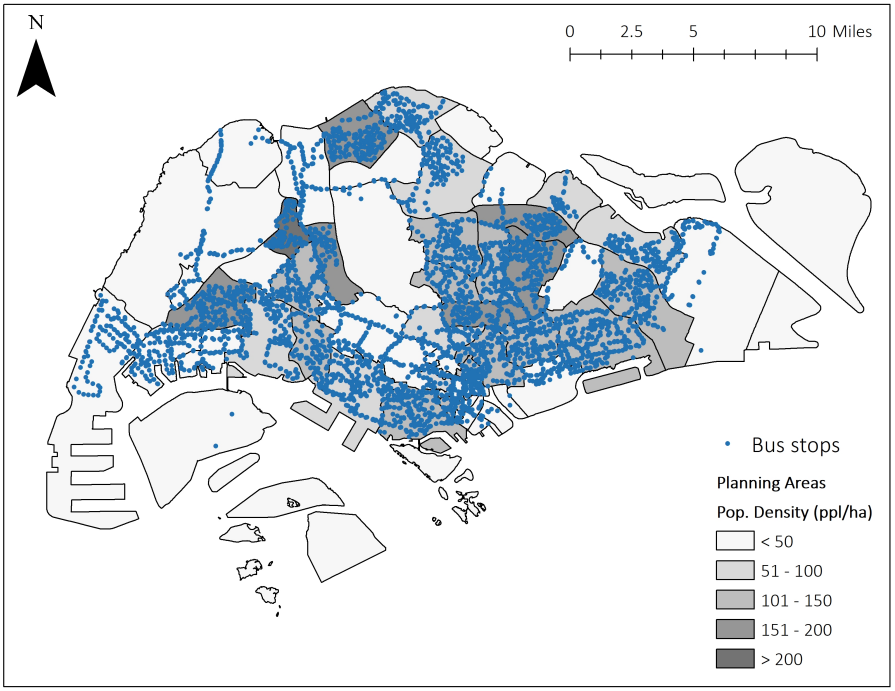


Figure 3-1: Bus network in Singapore (as of 2012)

3.1.1 Private vehicle ownership policies

The Singaporean government seeks to control private vehicle ownership through two major policies: (a) the *Additional Registration Fee* (ARF), and (b) the *Vehicle Quota System*

[5]https://www.straitstimes.com/singapore/transport/spore-public-transport-system-tops-global-list
[6]https://www.mot.gov.sg/about-mot/land-transport/public-transport

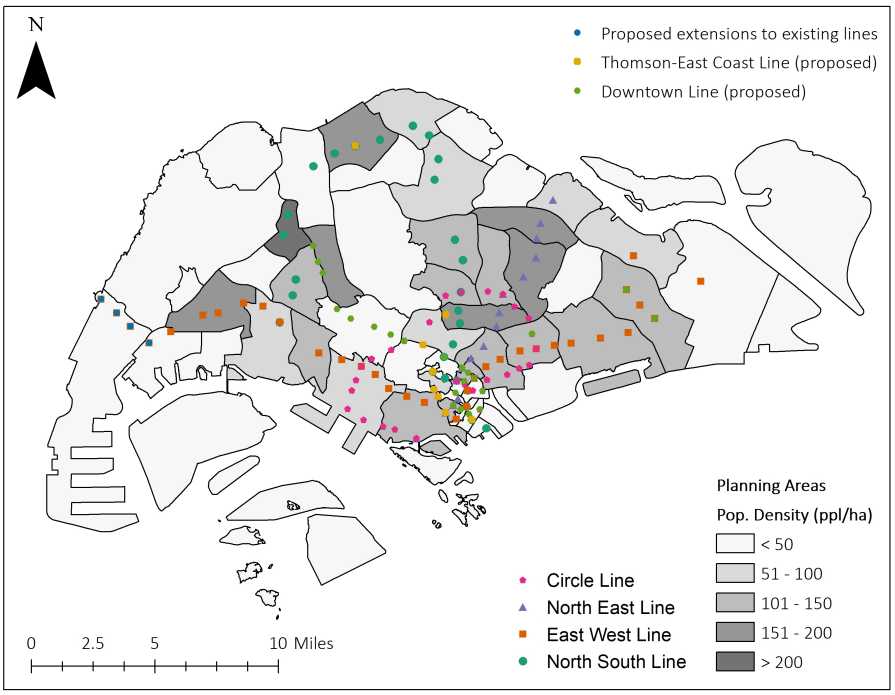


Figure 3-2: Mass Rapid Transit (MRT) network in Singapore (as of 2012)

(VQS). While the former limits fleet growth in addition to generating revenue, the latter scheme controls the total number of active vehicles on an annual basis. We discuss these policies in further detail in the following sub-sections.

Additional Registration Fee (ARF)

The Additional Registration Fee (ARF) was instituted as a share of the Open Market Value (OMV) of the vehicle in 1972. The ARF grew from 15% in 1972 to 150% in 1980 and 175% in 1983. The current tax structure is even more steep [7]. Cars registered on or after September 1, 1998 are subject to a registration fee of 220 SGD. The ARF is imposed on top of this registration fee using the following tiered scheme:

- First 20,000 SGD of OMV: 100% ARF rate
- Next 30,000 SGD of OMV: 140% ARF rate
- Above 50,000 SGD of OMV: 180% ARF rate

As an example, the ARF payable for a car with an OMV of 75,000 SGD stands at $\big[(100\% * 20,000) + (140\% * 30,000) + (180\% * 25,000) = \big]$ 107,000 SGD. Thus, we can see how the ARF

[7] https://www.lta.gov.sg/content/ltaweb/en/roads-and-motoring/owning-a-vehicle/costs-of-owning-a-vehicle/tax-structure-for-cars.html

could easily exceed the OMV for certain vehicle models. Road taxes are also levied based on engine and fuel types, while rebates are offered for older cars and amount of vehicular CO_2 emissions.

Vehicle Quota System (VQS)

Registration of every new vehicle must be preceded by the acquisition of a Certificate of Entitlement (COE), which represents the right to own and use a vehicle for 10 years. At the end of the 10-year period, vehicle owners can de-register their vehicle or choose to renew their COE for another 5- or 10-year period by paying the Prevailing Quota Premium. If the vehicle is de-registered before the expiration of its COE, a rebate is prorated to the number of days remaining on the COE. Users can apply for five categories of COE — small cars, large cars, goods vehicles and buses, motorcycles, and an open category[8]. COEs are allocated through an open bidding process, which is conducted twice a month. The number of available COEs is dependent on the limits set by the Vehicle Quota System (VQS). As the supply of COEs is highly regulated, prices are extremely volatile depending on the current levels of demand. COE prices have varied between 25,000 - 100,000 SGD for cars and 2,000 - 10,000 SGD for motorcycles over the past decade[9]. COEs are also transferable along with the vehicle in the case of private vehicle transactions between individuals.

Introduced in May 1990, the Vehicle Quota System regulates the rate of growth of vehicles on Singaporean roads, at a rate that can be sustained by developments in their land transport infrastructure. Calculation of the vehicle quota is carried out every three months, and is determined by the number of vehicles de-registered, an allowable growth rate in vehicle population (set by the government), and certain adjustments to account for changes in taxi population, replacements, and expired or cancelled temporary COEs[10]. The allowable growth rate is set to be very restrictive. While it started out at 3% per annum prior to 2009, stringent decreases since then have resulted in the current rate standing at 0.25% per annum (up to January 2018)[11].

Policies under the Vehicle Quota System have been effective in controlling the overall

[8] https://www.lta.gov.sg/content/ltaweb/en/roads-and-motoring/owning-a-vehicle/vehicle-quota-system/certificate-of-entitlement-coe.html

[9] https://coe.sgcharts.com/

[10] https://www.lta.gov.sg/content/ltaweb/en/roads-and-motoring/owning-a-vehicle/vehicle-quota-system/overview-of-vehicle-quota-system.html

[11] https://www.lta.gov.sg/content/ltaweb/en/roads-and-motoring/owning-a-vehicle/vehicle-quota-system.html

number of vehicles added to the national fleet. Moreover, high taxes and financial regulations have transformed car ownership into a luxury good in Singapore. Apart from the COE and the ARF, users also have to pay a registration fee, a 20% excise duty on the OMV, and a 7% Goods & Services Tax in addition to operating costs. For example, a new Toyota Corolla Altis, whose OMV was 104,998 SGD as of April 2017, is estimated to cost 163,950 SGD over a 10-year period[12]. Annual operating costs would include an average of 1,473 SGD on car insurance, 621 SGD for maintenance costs, 742 SGD in road tax, and 2,341 SGD in petrol costs[13]. In another comparison using the Volkswagen Golf 1.4 as an example, the total cost in Singapore is reported to be around 110,500 USD, which is almost 2.5 times the price in neighboring Malaysia[14].

3.1.2 Private vehicle usage policies

Policies to control private vehicle usage in Singapore evolved in three phases. First, the Area Licensing Scheme (ALS) was introduced in 1975 that charged drivers entering the CBD. However, the initial success of this policy could not be sustained over time as employment opportunities in the CBD rose. Second, the ALS was extended to major expressways outside the CBD in 1995 through the Road Pricing Scheme (RPS). Finally, the Electronic Road Pricing (ERP) scheme, which integrated the ALS and RPS strategies in an automated manner, was introduced in 1998. The government hopes to reduce on-road congestion by harnessing technological advances and adopting a "pay-as-you-use" principle on road usage.

Electronic Road Pricing (ERP)

Motorists are charged when they use certain roads during peak hours according to the ERP scheme. ERP rates vary according to the location of the road and time period, and are dependent on local traffic conditions. These rates are determined by a quarterly review of traffic speeds of priced roads, and during the June and December school holidays. These calculations are based on an optimal speed range of 20-30 kmph on arterial roads and 45-65 kmph on expressways[15].

[12] https://dollarsandsense.sg/2018-edition-cost-owning-car-singapore-10-years/
[13] https://www.valuechampion.sg/costs-car-ownership-singapore
[14] https://blog.seedly.sg/factors-cost-car-price-singapore/
[15] https://www.lta.gov.sg/content/ltaweb/en/roads-and-motoring/managing-traffic-and-congestion/electronic-road-pricing-erp.html

Off-Peak Car (OPC)

The Off-Peak Car (OPC) is a government scheme in Singapore that aims to curb rush hour traffic by allowing the purchase of a car that can only be used during weekday off-peak hours (7PM - 7AM) at a reduced price. The use of the OPC is unrestricted during weekends and public holidays. Additional incentives are provided in the form of reduced toll costs and road taxes, and lower Certificate of Entitlement (COE) charges. This scheme was introduced by the Land Transport Authority (LTA) in 1994. While the original scheme is no longer available for registration or conversion, car owners can register for or convert to the Revised Off-Peak Car (ROPC) scheme that was implemented in 2009. New cars registered under the ROPC scheme can enjoy up to a 17,000 SGD rebate on the Quota Premium for a COE and the ARF. Moreover, there is a flat discount of up to 500 SGD on annual road tax[16]. There were around 34,000 OPCs in Singapore (as of July 2015), which accounts for only 5.8% of all private cars[17]. One of the reasons behind the low uptake may be the temporal restriction that prohibits drivers from using it as a commute mode for day-shifts. Therefore, we can hypothesize that workers with night-shift jobs (such as blue-collar workers in certain professions) and households that use this car as a secondary vehicle for weekend trips would be more likely to avail of this scheme.

3.2 Data Sources

The primary data source used in this thesis is the Household Interview Travel Survey (HITS). Along with using both the 2008 and 2012 versions, we augment our data set with information related to the built environment and land use. Finally, we use travel skims that provide detailed travel time and cost information between an origin-destination (OD) pair at the level of Traffic Analysis Zones (TAZs).

3.2.1 Household Interview Travel Survey (HITS)

The HITS is a pen-and-paper survey that is carried out once every four years and is used to collect data about households, individuals and their travel patterns for one observed working day. The survey contains three sections — *household particulars*, *individual particulars* and

[16]https://www.lta.gov.sg/content/ltaweb/en/roads-and-motoring/transport-options-for-motorists/revised-off-peak-car-and-opc-and-weekend-car.html

[17]https://dollarsandsense.sg/do-people-still-get-off-peak-cars/

trip particulars. Socio-demographic characteristics of a random 1% of households (∼9,500) and individuals (∼35,000) were recorded in the first two sections. The final section contains data about each stage of a trip that each individual undertook with trip details such as point of origin/destination, travel time, mode, purpose, etc. We summarize the HITS questionnaire in Table 3.1. It is worthwhile to mention that there are only minor differences between the HITS 2008 and 2012 questionnaire, such as the inclusion of household ethnicity and child's school location in 2012. For the sake of maintaining consistency, we consider the same set of variables from 2008 and 2012 in our first study. However, the second and third studies use variables from the 2012 version, as additional information is available and temporal comparison is not necessary for those studies.

Since individual income was provided as a categorical variable, we created a log-normal distribution for income and randomly sampled from this distribution to obtain a continuous income value for each individual. The sampling procedure was constrained to ensure that the randomly picked value corresponded to the category mentioned in HITS. Missing or refused incomes (for adults with jobs) were imputed based on other individual characteristics. More details on the income imputation procedure can be found in Appendix A.1.

We also noticed that the HITS sample underestimated taxi ownership in Singapore by a significant amount. This is a particularly critical flaw in the sampling technique because households with a taxi are unlikely to be owning an additional private car. After using sampling weights, the total taxi count in the weighted HITS population is around 18,500, while the actual taxi count in Singapore in 2012 was close to 26,000. Therefore, households that would be most likely to own a taxi were identified based on employment, occupation and industry of individuals' jobs. We then used an imputation method to randomly assign taxis to a subset of these selected households through a weighted iterative proportional fitting procedure such that the total taxi count reached close to 26,000. More details on the taxi imputation procedure can be found in Appendix A.2.

3.2.2 Built environment & Land use

It is pertinent to mention here that postcodes in Singapore usually represent individual buildings, so the residential location of each household is specified at the building level. Similarly, the job location of each individual and the origin/destination of each trip are also specified at the postcode level. Evidence of the fine-grained detail of location information

Table 3.1: Description of HITS questionnaire

Section	Variable	Original Encoding	Examples
Household (HH)	Dwelling type	Categorical	HDB 1-room, private flat, etc.
	Dwelling location	Categorical	Postcode, if available
	Household size	Continuous	1, 2, etc.
	Available vehicles	Categorical	Normal car, Taxi, Bike, etc.
	Vehicle properties	Categorical	Registered, Rental, etc.
	Sampling weight	Continuous	50, 100, etc. (can be fractional)
Individual (IND)	Age	Categorical	6-9 years, 10-14 years, etc.
	Resident status	Categorical	Citizen, Permanent resident, etc.
	Gender	Categorical	Male, Female
	Driving license	Categorical	Car, Motorcycle, etc.
	Employment status	Categorical	Employed full-time, Student, etc.
	Occupation	Categorical	Professional, Service and sales, etc.
	Industry	Categorical	Manufacturing, Construction, etc.
	Job location	Categorical	Postcode, if available
	Monthly income	Categorical	No income, $1-$1000, Refused, etc.
	Sampling weight	Continuous	50, 100, etc. (can be fractional)
Trip (TP)	Origin	Categorical	Postcode, if available
	Destination	Categorical	Postcode, if available
	Start time	Continuous	0001-2400
	End time	Continuous	0001-2400
	Mode	Categorical	Car, Bus, MRT, etc.
	Purpose	Continuous	Work, Pick-up/drop-off, etc.
	Sampling weight	Continuous	50, 100, etc. (can be fractional)

can be seen from Figure 3-3.

We utilized this fine-grained detail to augment postcode information with spatial data related to the built environment and land use. First, we incorporated built environment data by calculating distances to amenities like bus stops, MRT stations, primary schools, shopping malls, expressways, and the CBD for each postcode. Next, we conducted a spatial join between postcodes and parcels to augment our data set with land use data. The land use map of Singapore is shown in Figure 3-4 at the parcel level.

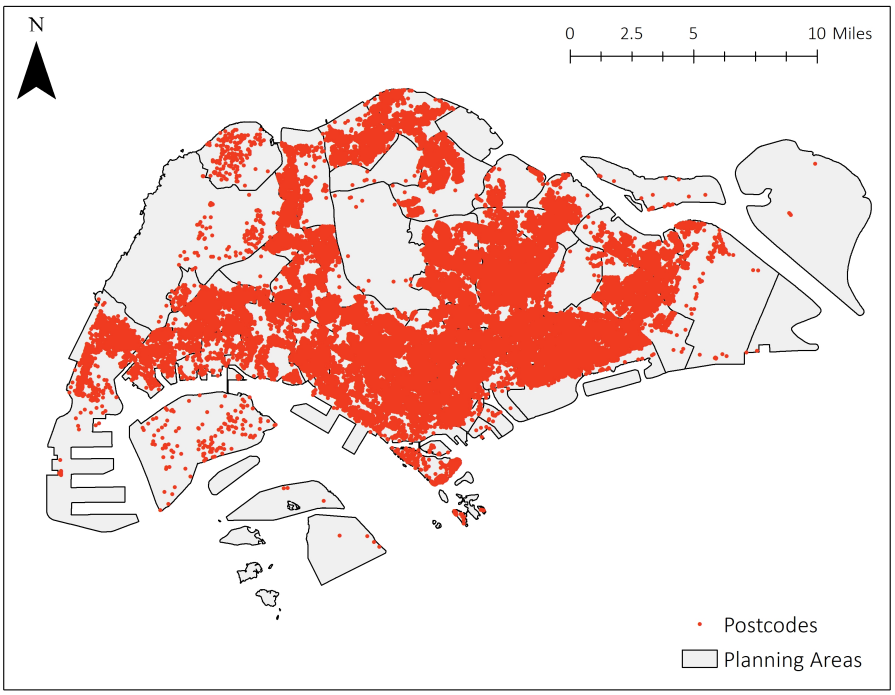


Figure 3-3: Postcodes (buildings) in Singapore (as of 2012)

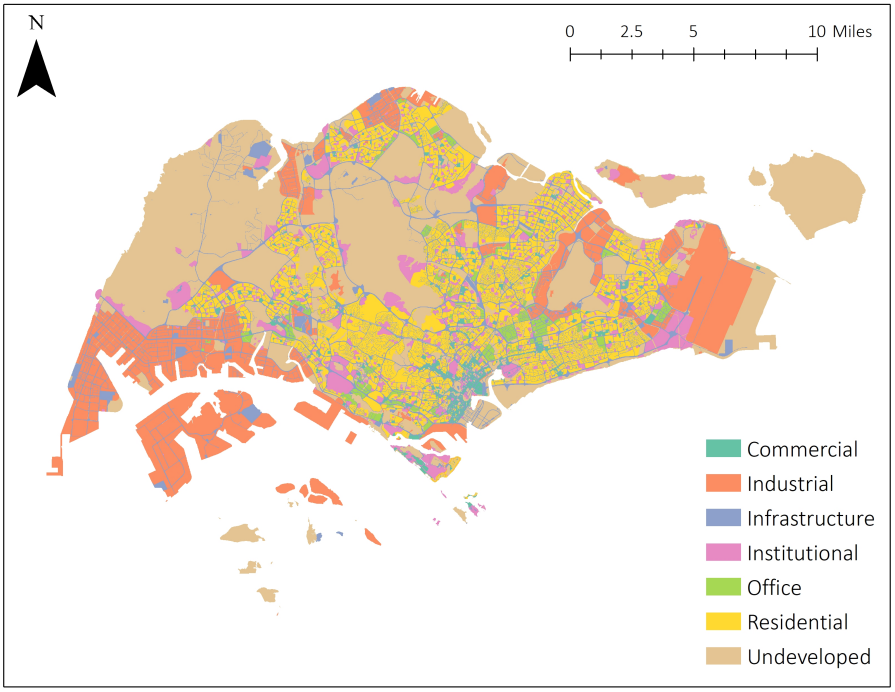


Figure 3-4: Parcel-level land use map of Singapore (as of 2012)

Several land use metrics were calculated for each postcode:

- **Residential density:** No. of residential units per sq.km. in a 500 meter buffer

- **Job density:** No. of jobs per sq.km. in a 500 meter buffer
- **Share of a particular land use:** Weighted average of land use area in a buffer of one kilometer. It is worth mentioning here that we considered seven land use categories in this study: *Residential*, *Commercial*, *Industrial*, *Infrastructure*, *Office*, *Institutional*, and *Undeveloped*.
- **Generalized Land Use Diversity Index (GLUDI):** We define the GLUDI in a 1km buffer as follows.

$$GLUDI = 1 - \left[\frac{\sum_{j=1}^{N} \left| \frac{A_j}{T} - \frac{1}{N} \right|}{2\left(1 - \frac{1}{N}\right)} \right] \tag{3.1}$$

  where there are N different types of land uses, A_j represents the area occupied by the jth land use type, and $T = \sum_{j=1}^{N} A_j$ represents the total land area. We attempt to capture the mix of land uses relative to a perfectly equal distribution of uses through this index. When the land in the area has a single use, the index achieves a value of zero. On the other hand, a value of one indicates perfectly equal mixing among the N different land uses.

- **Shannon Entropy for Land Use (SELU)**: We define the SELU in a 1km buffer as follows.

$$SELU = -\sum_{j=1}^{N} p_j \cdot \log p_j \tag{3.2}$$

  where there are N different types of land uses, and p_j is the proportion of land area belonging to the jth land use type.

3.2.3 Travel skims

Since the trip information in the survey pertains to only a single observed day, we used travel time skims to extract further information about commuting trips undertaken by individuals in the sample. These skims provide values of travel time and cost for an OD pair differentiated by mode.

- **Car:** Distance, Travel time, ERP toll
- **Public transport:** In-vehicle travel time, Walking time, Waiting time, Number of transfers, Cost

Table 3.2: Results of Principal Component Analysis of commute travel variables

Variable	**Factor 1**	**Factor 2**
Distance to nearest bus stop from job location	0.33	-0.61
Distance to nearest MRT station from job location	0.43	-0.48
Road network distance between residential and job locations	0.46	0.44
Difference in total travel time between car and public transport for a trip from residential to job location	0.55	0.01
Number of transfers using public transport for a trip from residential to job location	0.43	0.45
Eigenvalue	2.57	1.18
Proportion of variance explained	0.514	0.236

Such detailed information allows us to examine the perceived gain from choosing a particular mode over the other. Therefore, we conducted a Principal Component Analysis (PCA) for every individual with a job using commute time variables related to the ease of using public transport as a commute mode relative to a private car. The results of the PCA are shown in Table 3.2. The optimal number of factors was found to be five, among which only two had eigenvalues greater than one. The first component explains about 51% of the total variance, and has positive signs for all variable effects. Thus, we can interpret this factor as a measure of generalized travel impedance for distant job locations with low public transport accessibility.

The second factor is relatively weaker as it can explain only about 24% of the variance. The interpretation is also more complex due to mixed effects. The second factor seems to represent the generalized travel impedance for a scenario where the job location is distant but has good access to public transport. While there is no significant time differential between the two modes, transfers would be required if public transport were to be used. Perhaps this is indicative of a trunk line effect, where distant job locations are not as inconvenient to get to if they are on subway lines or close to feeder lines. When both factors were included in our behavioral models, the second factor was always found to be statistically insignificant. Therefore, we dropped it from consideration and re-estimated the models using only the first factor as a generalized travel impedance measure.

THIS PAGE INTENTIONALLY LEFT BLANK

Chapter 4

Using a hybrid approach to predict impacts of new mobility

4.1 Introduction

The past decade has witnessed the emergence of a plethora of new mobility services, such as on-demand mobility, ride-sharing, bike-sharing, electric scooters, and autonomous vehicles (AVs). These emerging transportation technologies are attracting increasing attention from policy-makers as the transportation sector continues its growth to encompass mobility in all forms. Several cities have adopted the Mobility-as-a-Service (MaaS) scheme, where consumers can avail of a package comprised of multiple types of services with a weekly or monthly pass. Such new technologies and business models necessitate revisiting traditional forecasting models to evaluate their appropriateness for the swiftly changing paradigm of personal mobility.

Econometric or discrete choice models have been traditionally used to forecast market shares by modeling the underlying behavioral mechanisms associated with people's observed choices. While they excel at interpretability, their predictive power is quite modest. With the advent of big data and an increase in computing capabilities over the past couple of decades, machine learning has gained popularity in a variety of domains, including but not limited to transportation. Machine learning revolves around the problem of prediction, while many economic applications are concerned about parameter estimation (Mullainathan and Spiess, 2017). Therefore, combining the interpretable nature of econometric models with

the excellent predictive performance of ML models is the research question of the hour.

This is certainly a challenging endeavour as each framework has certain associated shortcomings. Econometric models require data curation and variable transformations, which require significant time and effort on the modeler's part (especially for high-dimensional feature spaces), to tease out economic interpretations and causal effects. On the other hand, machine learning models are criticized for being black-boxes that inhibit modelers from getting a sense of the inner workings of the model that happen under the hood of open-source packages. This is dangerous as ignorant use of these models for policy-making can have unintended but severe consequences.

In this chapter, we propose a framework that tests different econometric and ML models with various types of variable specifications to arrive at a model that maintains high interpretability and predictive power. Our framework also addresses the issue of low-sample alternatives, which is frequently observed in real-world data sets. This is an important contribution because these "abnormal" or rare patterns are the examples that are most interesting to examine. We then apply our framework to predict the impact of new mobility options, using the off-peak car in Singapore as an example. We find that our hybrid framework achieves a significant improvement in predictive performance of minority classes over standard econometric models, while maintaining the same level of transparent interpretability.

The remainder of this chapter is structured as follows. Relevant themes in the literature are reviewed and discussed in Section 4.2. We detail the elements of our hybrid framework in Section 4.3. Methodological details for the models and algorithms considered in this study are provided in Section 4.4. The empirical results of the framework application are discussed in Section 4.5. Finally, we provide concluding remarks and suggest areas for future research in Section 4.6.

4.2 Literature Review

We position our literature review according to three important themes relevant to this chapter and discuss related works in the following sub-sections. The first section identifies different categories of machine learning applications in the transportation domain. The second section discusses ongoing efforts to use machine learning techniques in choice models to

arrive at a hybrid approach. Finally, we discuss broad works that highlight the increasingly ubiquitous nature of obtaining real-world data with imbalanced classes, and state-of-the-art techniques to address this phenomenon.

4.2.1 Machine learning applications in transportation

The advent of big data in the transportation domain has greatly enabled the use of machine learning in several transportation applications. *The first thread of inquiry that we identify from the literature is inference of travel mode from smartphone sensors or GPS trajectories.* Jahangiri and Rakha (2015) obtained data from different smartphone sensors, such as accelerometer, gyroscope, and rotation vector sensors and compared the performance of multiclass classifiers such as k-nearest neighbor (kNN), support vector machine (SVM), and random forest (RF). Their findings indicate that the RF and SVM models had the best performance in identifying the transportation mode. Ashqar et al. (2018) improved upon this study by extracting new frequency domain features from time domain features, which helped in improving the accuracy quite significantly. They too found that RF and SVM classifiers had the best performance.

Using similar data collection techniques, Fang et al. (2017) applied a deep learning approach which was found to outperform traditional machine learning methods. Hoping to circumvent the challenges associated with collecting data via smartphones, Dabiri and Heaslip (2018) leveraged GPS trajectories to infer commuting modes through a Convolutional Neural Network (CNN) approach. Their model structure was found to be superior to traditional ML algorithms, while maintaining interpretability by designing the layout of the input layer to represent fundamental motion characteristics of a moving object (such as speed, acceleration and jerk).

Another frequently researched application is the use of neural networks for traffic predictions. This has been fueled by the explosion of large-scale traffic data availability and the increased attention that intelligent transportation systems (ITS) have been receiving from policy-makers. Ma et al. (2015) combined a deep Restricted Boltzmann Machine with Recurrent Neural Network (RNN) architecture to predict evolution of traffic congestion based on GPS data from taxis. The use of a deep architecture model was enhanced by the representation of traffic flow features through autoencoders by Lv et al. (2015). A hybrid framework using Queueing Theory and ML was used to address outliers in an application

of travel time prediction by Gal et al. (2017).

It is worth noting that the aforementioned studies did not account for the spatial dimension, which has started receiving attention very recently due to advances in computing power. Yu et al. (2017) used a spatiotemporal recurrent convolutional network (SRCN), which inherits the advantages of deep CNNs and long short-term memory (LSTM) NNs for traffic forecasting. Incorporating the spatial domain resulted in better performance compared to traditional deep architectures for both short- and long-term forecasts. While model architectures grow in complexity and achieve better prediction performance, the issue of interpretability is often ignored by most studies. In an attempt to maintain model rationality and interpretability, Wang et al. (2019) proposed a path-based deep learning framework where each critical path of the road network is modeled through a bi-directional LSTM NN. Along with providing better network-level traffic speed predictions, they illustrated the model's interpretability by explaining the physical meaning of the features from the hidden-layer output.

New mobility services like bike-sharing systems (BSS) and on-demand ride-sharing are also being examined with the help of machine learning approaches. Bacciu et al. (2017) analyzed the feasibility of BSS by using ML methodologies to predict when bikes would be returned to empty stations with an eye towards improving user satisfaction. With a similar motivation, Gao and Lee (2019) tried to forecast BSS rental demand through a hybrid model that combined a genetic algorithm (GA) with a back propagation network (BPN) in an unsupervised classification framework. An ensemble learning approach was used by Chen et al. (2017) to better understand ride-splitting behavior of passengers of ride-sourcing companies that provide pre-arranged and on-demand transportation services.

Machine learning is also useful for analyzing opinions related to public policies and new services. Reactions to a novel transportation policy (the Odd-Even Policy in Delhi, India) were examined by Basu et al. (2017) through sentiment analysis of Twitter data. Similarly, Rahim Taleqani et al. (2019) also used Twitter posts to assess public opinion on dockless BSS. These studies highlight the potential of analyzing big data from microblogging media platforms with ML algorithms for planning and decision-making.

4.2.2 Econometrics and Machine Learning: Two peas in a pod

While the statistics community has accepted the ML revolution by and large, adoption of ML methods has been much slower in econometrics. However, a growing body of intersectional work, both empirical and methodological, has begun to emerge in recent times. Computational Social Science (CSS) has become a mainstream approach in the empirical study of issues related to policy analytics in various domains (Kauffman et al., 2017). The slow rate of adoption in the economic community may be due to the community's culture towards establishing causal effects. Now, recent advances in ML methods allow modelers to adjust for differences between treated and control units in high-dimensional settings, and identify and estimate heterogeneous treatment effects (Athey and Imbens, 2017). Athey (2018) provides an excellent review of applications of ML to policy problems. Newly developed intersectional methods that are relevant to applications involving causal inference for average treatment effects, optimal policy estimation, and estimation of the counterfactual effect of price changes in consumer choice models are highlighted in Athey and Imbens (2019).

There is a large body of work that provides empirical comparisons between econometric and machine learning model performance across domains; see Bajari et al. (2015) for microeconomic demand estimation and Hagenauer and Helbich (2017) for travel mode choice modeling. The only paper using an application of household vehicle ownership (to be the best of our knowledge) reports that ML models like RF and SVM outperform the traditional multinomial logit (MNL) model (Paredes et al., 2017). They also find that the predictive performance of MNL increases when features are engineered to be more appropriate for discrete choice contexts. This study motivates us to consider multiple variable specifications in our effort to enhance the predictive power of econometric models. In lieu of discussing empirical work from other domains, we will next highlight a few studies that discuss methodological details of combining these two approaches. These studies are drawn from the discrete choice literature and focus on applications in transportation. More importantly, they seek to propose hybrid models that build on the predictive performance of ML models and interpretability of econometric models.

Brathwaite et al. (2017) highlight how decision trees represent a non-compensatory decision protocol, which generalizes many of the non-compensatory rules used in the discrete

choice literature. In addition to obtaining much better forecasts compared to the MNL model, their bayesian model tree is found to be over 1,000 times more likely to be closer to the true data-generating process than MNL. Sifringer et al. (2018) proposed that the predictive accuracy of discrete choice models (DCMs) can be augmented by introducing an additional error term in the utility equations that is estimated exogenously using DNNs. However, the economic interpretation of this term and its coefficient remain elusive. In an attempt to provide economic meaning to DNN outputs, Wang and Zhao (2018) argued that the output of the penultimate DNN layer can be thought of as utility values that serve as inputs to the logistic function for choice probability prediction. We provide a counter-argument to Wang and Zhao (2018) that including non-linear transformations in the hidden layers, from which DNNs derive their high predictive power, distorts the utility formulation and makes it inconsistent with the traditional economic definition of utility. Therefore, treating penultimate layer outputs as utilities would be fallacious.

4.2.3 The class-imbalance phenomenon

Rare events and unusual patterns are extremely interesting to observe and interpret, but are just as difficult to detect. Examples include natural disasters, fraudulent credit card transactions, cancer gene expressions, and adoption of new technologies. Event detection is primarily a prediction problem, but the infrequent occurrence of rare events makes this prediction problem even more challenging. If one or more classes have a significantly higher number of samples than the remaining classes, then that data set is termed as *imbalanced*. The most prevalent class is called the *majority class*, while the low-sample alternatives are called *minority classes*.

Standard ML classifiers like LR, SVM and RF have been found to provide sub-optimal classification results for imbalanced data sets (Lane et al., 2012). While the majority samples are covered well, predictions for the minority sample are distorted. The learning process is guided by aggregate performance metrics such as prediction accuracy, which induces a bias towards the majority class (Loyola-González et al., 2016). Even if the model produces a high overall precision, the rare events remain unknown. Samples from the minority class are treated as noise by the model. The contrary can also be true, where noise is misclassified as minority class examples as both are rare patterns in the feature space (Beyan and Fisher, 2015). Skewed sample distributions may be easier to learn when the classes are easily

separable. However, minority samples usually invade into other regions, where the prior probabilities of both classes are almost equal (Díez-Pastor et al., 2015). Moreover, small sample size with high dimensionality in the feature space further compounds the difficulty of detecting rare patterns (Wasikowski and Chen, 2010).

This research area has picked up significantly over the past decade, with a variety of algorithms being proposed to address problems related to imbalanced data classification. There are two classical pre-processing state-of-the-art techniques for dealing with imbalanced learning. While we provide a brief overview of these techniques in the following sub-sections, interested readers should refer to a thorough review by Haixiang et al. (2017) for details on methods and applications of imbalanced classification.

Resampling

Resampling techniques are used to rebalance the feature space for an imbalanced data set and realign the skewed class distribution. Since they are independent of the classifier, they are more versatile across applications (Lane et al., 2012). There are three categories of resampling techniques:

1. **Over-sampling:** New minority class samples are created through synthetic sample generation algorithms or by randomly duplicating existing minority samples.

2. **Under-sampling:** Majority class samples are discarded to address the skewed class distribution. The simplest method is Random Under-Sampling (RUS), where majority class samples are randomly dropped from the data set (Tahir et al., 2009).

3. **Hybrid:** These methods are a combination of over-sampling and under-sampling techniques.

Feature selection and extraction

While these techniques are not as popular as resampling, removing irrelevant features from the feature space reduces the risk of misclassifying low-sample classes as noise (Lima and Pereira, 2015). Feature selection entails the selection of a subset of features from the feature space that optimizes the performance of a classifier. Feature selection can be divided into filters, wrappers, and embedded methods (Guyon and Elisseeff, 2003).

Feature extraction is related to dimensionality reduction, whereby the high-dimensional feature space can be transformed to a low-dimensional space. There are a variety of techniques for feature extraction, such as Principal Component Analysis (PCA), Singular Value Decomposition (SVD), and f Non-negative Matrix Factorization (NMF). It should be noted that feature extraction methods tend to be employed more frequently for unstructured data such as images, text and speech. The reader is directed to Allahyari et al. (2017) and Alías et al. (2016) for reviews of feature extraction techniques in different domains.

4.2.4 Key takeaways

The key takeaways from our review of the literature can be summarized as follows:

- Machine learning has been used in transportation applications for quite some time now, but new technologies like on-demand mobility and bike-sharing systems have started to be analyzed with ML techniques in recent times. The computational capacity for data mining and big data analytics have also enabled analysis of public opinions related to these technologies and policies.

- Despite the reluctance of economists in adopting ML models compared to statisticians, recent advances are enabling more intersectional research. Several studies have also conducted comparative experiments between econometric and machine learning models. While there have been a few papers trying to propose interpretable hybrid models in the past couple of years, their conclusions are debatable and this is still an exciting open field of inquiry.

- Real-world data sets contain rare events and unusual patterns, which are quite difficult to detect using standard ML classifiers. These events of interest are misclassified as noise due to their low sample size. Several techniques exist to address the imbalanced classification problem, the most heavily researched ones being (a) resampling, and (b) feature selection and extraction.

4.3 Framework

Our attempt to combine the interpretability of econometric models with the high predictive power of machine learning models is illustrated through a three-stage framework. We provide

relevant details about each stage in the following sub-sections.

4.3.1 Model construction

The first stage of the framework is related to model construction. The goal of the first stage is to identify the best-performing econometric model that is at par with machine learning models considering both overall predictive accuracy and alternative-specific market share predictions. We define two types of scenarios: (a) *estimation*, and (b) *prediction*. This is motivated from contextual knowledge that econometric models are better for estimation purposes, i.e., when the underlying behavioral mechanisms need to be understood. In contrast, machine learning models perform much better at prediction but their black-box nature remains a major caveat.

We used the Household Interview Travel Survey (HITS) from Singapore in this study to illustrate the application of our framework to the prediction of household vehicle availability. The travel survey data is augmented with data about land use, the built environment, and local accessibility to create about 70 features. In the estimation scenario, the HITS 2008 data is split into training data (80% of HITS 2008) and test data (20% of HITS 2008). The prediction scenario is constructed keeping real-world prediction applications in mind, where we used HITS 2008 for training and HITS 2012 for testing. A summary of this procedure can be found in Table 4.1.

Table 4.1: Model scenario description

Scenario	**Training data**	**Test data**
Estimation	80% of HITS 2008	20% of HITS 2008
Prediction	HITS 2008	HITS 2012

Two types of econometric models were selected and compared with six machine learning models using a standard variable specification. We progressively moved towards more advanced variable specifications that improved the predictive power of econometric models. While we focused on both model scenarios (estimation and prediction) for the standard variable specification, we dropped the estimation scenario from consideration for more advanced variable specifications as our ultimate goal was to illustrate the application of this framework for predictive purposes. At the end of the first stage, we were able to identify the

best-performing econometric model enabled by the best-performing variable specification.

4.3.2 Handling low-sample alternatives

The data sets selected for this study, like most real-world data sets, suffer from the phenomenon of class-imbalance. This means that the distribution of samples among the different classes is skewed. We considered six alternatives of household vehicle availability[1], which are constructed in a progressively utilitarian manner that implies a comparatively higher utility value as we move up the mobility scale. The true market shares of the six alternatives for both scenarios are presented in Table 4.2.

Table 4.2: True market shares of vehicle availability alternatives

Category	Description	Train (%)	Test (%)	Δ (%)
	Estimation Scenario [a]			
0	No vehicles	55.49	54.86	-0.63
1	One motorcycle only	5.48	5.37	-0.10
2	One off-peak car w/wo motorcycles	1.44	1.67	+0.24
3	One normal car only	32.02	32.28	+0.26
4	One normal car with other vehicles	1.27	1.44	+0.18
5	Two normal cars w/wo other vehicles	4.32	4.38	+0.06
	Prediction Scenario [b]			
0	No vehicles	55.36	54.28	-1.08
1	One motorcycle only	5.46	4.39	-1.07
2	One off-peak car w/wo motorcycles	1.48	2.03	+0.54
3	One normal car only	32.07	33.96	+1.89
4	One normal car with other vehicles	1.30	1.00	-0.30
5	Two normal cars w/wo other vehicles	4.33	4.34	+0.02

[a] Using HITS 2008 data with 80/20 train/test split

[b] Using HITS 2008 data to predict HITS 2012 shares

We found that differences between the market shares for the training and test data in the estimation scenario are fairly low. This is because a single data set was randomly split

[1]The Off-Peak Car (OPC) is a government scheme in Singapore that aims to curb rush hour traffic by allowing the registration of a car that can only be used during weekday off-peak hours (7PM - 7AM) and weekends. Financial incentives for OPC users include a COE rebate up to 17,000 SGD for a total of 10 years, and an annual road tax rebate up to 500 SGD.

into training and test data using a 80-20 split for estimation. Low difference values imply that the split was fairly random and did not contribute to further class-imbalance. However, for the prediction scenario, the training data set is from HITS 2008 while the test data set is from HITS 2012. The differences for this scenario have meaningful interpretation, wherein they represent the change in the true market share of each alternative over four years from 2008 to 2012. It is worth highlighting that there was a decrease in the no-vehicle market share, while more households bought off-peak cars and normal cars. This would imply a rise in the total vehicle stock in the country, despite the strict regulatory government policies in this domain and the completion of the fourth Mass Rapid Transit line (i.e., the Circle Line) in 2011.

We also noticed the phenomenon of class-imbalance in Table 4.2, where the *no-vehicle* and *one normal car* alternatives are majority classes. All the other classes have market shares ranging from 1% to 5%. We are interested in forecasting market shares of new mobility options, which are going to have zero or low market shares when they are introduced for the first time. We do not believe that autonomous vehicles (AVs), for example, will completely replace cars in a single fell swoop (as most literature would have us believe); rather, this replacement will take place in a gradually incremental fashion. Therefore, we wanted to account for class-imbalance and consider how to handle low-sample alternatives accurately. To that effect, we considered six synthetic sample generation algorithms, and compared the model performance before and after the addition of the newly-generated synthetic samples. The second stage of the framework identifies the synthetic sample generation algorithm that improves the performance of the model selected from the first stage to the greatest degree.

4.3.3 Forecasting market shares of new mobility

Stage three of the framework uses the results from the previous two stages and sets up a process to forecast the market shares of low-sample new mobility options. Recall that we are equipped with the best econometric model (M_{econ}), best variable specification (X), and the best synthetic sample generation algorithm (λ_{SS}) from the previous stages. The steps of the forecasting procedure are detailed below.

1. Since we are concerned with forecasting, the prediction scenario is the appropriate choice here.

2. We term the best variable specification for the training data, i.e., HITS 2008, as X_{08}.

3. We used λ_{ss} to generate synthetic samples for the minority classes in X_{08}, thereby obtaining an augmented training data set $X_{08}^{s} = X_{08} + X_{\lambda_{ss}}$. This data set has balanced classes and, thereby, represents a synthetic population.

4. In order to maintain consistency with the original sample X_{08}, we randomly drew samples from the synthetic population X_{08}^{s} maintaining the overall sample size and the market shares of individual categories.

$$X_{08}^{r} = f(X_{08}^{s}) \quad \text{s.t.} \quad N(X_{08}^{r}(j)) = N(X_{08}(j)) \; \forall j \tag{4.1}$$

   This randomly drawn sample X_{08}^{r} is a representative sample of the synthetic population along all considered dimensions. We hypothesize that X_{08}^{r} will perform better than X_{08} because X_{08} may be sparse along certain dimensions (especially in the case of high-dimensional feature matrices) and, thereby, fail to produce enough useful information for prediction. On the other hand, being a representative and random sample along all dimensions, X_{08}^{r} can successfully address the issue of sparseness while remaining similar in composition to X_{08} with regard to the sample size and existing market shares of the outcome variable.

5. In a real-world forecasting scenario, new mobility options will not exist in the training data, which makes forecasting their market shares all the more challenging. To mimic this challenge, we considered the off-peak car (i.e., category 2) as a proxy for new mobility and removed all samples pertaining to this class from the random sample to be used as the training data. Our choice is motivated by the sustainable nature of the off-peak car, making it closer in spirit to the new mobility options we are witnessing today compared to traditional alternatives like one normal car or multiple vehicle holdings. Thus, we obtained a reduced training data set $X_{08}^{'} = X_{08}^{r} - X_{08}^{opc}$.

6. The model M_{econ} was trained on this data set $X_{08}^{'}$ to obtain coefficient estimates β_M.

7. Consistent with choice modeling literature, we constructed the utility equation for the omitted alternative (OPC in this application) manually based on modelers' hypotheses. This practice is quite commonly used for obtaining mode choice shares of AVs, wherein

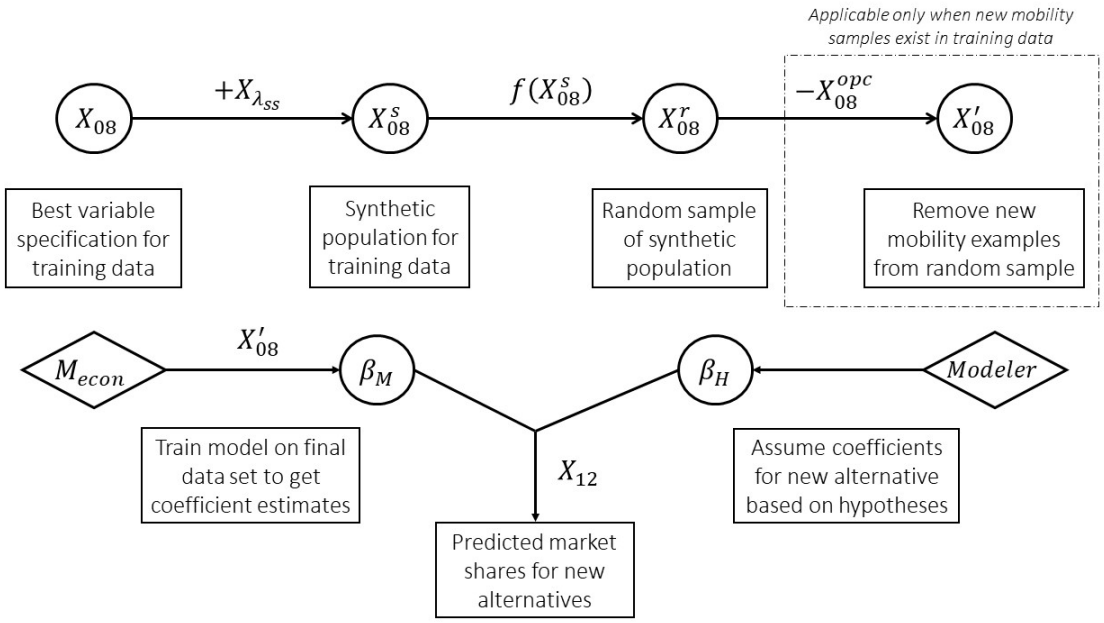


Figure 4-1: Proposed framework to forecast impacts of new mobility

certain time and cost modifications to observed values for cars or taxis are made for AVs. The coefficients that we assume from our hypotheses are termed as β_H.

8. We used both β_M and β_H to construct utility equations for all alternatives in the future scenario. These utility equations were then used to forecast market shares for all alternatives using the full test data set X_{12}, i.e., HITS 2012.

A summary of the forecasting procedure is shown in Figure 4-1.

4.4 Methodology

We provide the mathematical formulations and methodological details about our implementation of the aforementioned framework in the following sub-sections.

4.4.1 Econometric models

We considered the multinomial logit model as a representative of the unordered response structure and the ordinal logit model as a representative of the ordered response structure. The choice set of vehicle ownership was constructed in an ordered fashion on purpose to allow for examining whether considering the incremental ordered response structure yields a

better understanding of underlying preferences and consumer behavior. The mathematical framework of both these models are provided in the following sub-sections.

Multinomial logit (MNL) model

Random utility theory is a structural component of behavioral theory (Manski, 1977) and is a well-explored concept in the field of economics. Expressed briefly, it states that individual n selects alternative i which has the highest utility U_{in} among those in their choice set C_n. The utility U_{in} is composed of a systematic component that can be expressed as a linear-in-parameters function of explanatory variables (V_{in}) and a random component (ϵ_{in}).

$$U_{in} = V_{in} + \epsilon_{in} = \boldsymbol{\beta}^T \cdot \boldsymbol{X_{in}} + \epsilon_{in} \tag{4.2}$$

Therefore, the probability of individual n selecting alternative i from choice set C_n can be expressed as follows:

$$\begin{aligned} P(i \mid C_n) &= P(U_{in} \geq U_{jn}, j \in C_n) = P(U_{in} - U_{jn} \geq 0, j \in C_n) \\ &= P(U_{in} = \max_j U_{jn}, j \in C_n) \end{aligned} \tag{4.3}$$

Consider the case of a binary choice as an example to obtain a tractable expression.

$$\begin{aligned} P_n(1) &= P(U_{1n} \geq U_{2n}) = P(U_{1n} - U_{2n} \geq 0) \\ &= P(\epsilon_{2n} - \epsilon_{1n} \leq V_{1n} - V_{2n}) = F_{\epsilon_2 - \epsilon_1}(V_{1n} - V_{2n}) \end{aligned} \tag{4.4}$$

As can be seen from the expression above, this is the univariate cumulative distribution function (CDF) of $(\epsilon_2 - \epsilon_1)$. Similarly, an extension to three alternatives in the choice set would result in the bivariate CDF of $(\epsilon_2 - \epsilon_1)$ and $(\epsilon_3 - \epsilon_1)$. Different assumptions are made on the joint distribution of the error terms $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, ..., \epsilon_J)^T$ leading to different model structures. There are two major assumptions on the joint distribution of $\boldsymbol{\epsilon}$ that lead to the logit model. Note that a logit model can be binary (i.e. there are only two alternatives in the choice set) or multinomial (i.e. there are multiple — three or more — alternatives in the choice set. First, ϵ_{jn} is independently and identically distributed (i.i.d.), which leads to the following simplification:

$$f(\epsilon_1, \epsilon_2, ..., \epsilon_J) = \prod_{j=1}^{J} f(\epsilon_j) \tag{4.5}$$

Second, ϵ_{jn} follows an extreme value (EV) distribution with the shape parameter as zero and the scale parameter as μ. Thus, we can express $\epsilon_{jn} \sim EV(0, \mu), j \in C_n$. The CDF and probability density function (PDF) expressions for such a distribution are as follows:

$$\begin{aligned} F(\boldsymbol{\epsilon}) &= \exp\left(-e^{-\mu\boldsymbol{\epsilon}}\right), \mu > 0 \\ f(\boldsymbol{\epsilon}) &= \mu e^{-\mu\boldsymbol{\epsilon}} \exp\left(-e^{-\mu\boldsymbol{\epsilon}}\right) \end{aligned} \tag{4.6}$$

Based on these assumptions, we can arrive at a tractable expression for the choice probability of each alternative.

$$P(i \mid C_n) = \frac{\exp\left(\mu V_{in}\right)}{\sum\limits_{j \in C_n} \exp\left(\mu V_{jn}\right)} \tag{4.7}$$

From Equation 4.7, an important property of the logit model — independence from irrelevant alternatives (IIA) — is evident, which can be expressed through the odds ratio as follows:

$$\frac{P(i \mid C_{1n})}{P(j \mid C_{1n})} = \frac{P(i \mid C_{2n})}{P(j \mid C_{2n})} \tag{4.8}$$

where $i, j \in C_{1n}$; $i, j \in C_{2n}$; $C_{1n} \subseteq C_n$; $C_{2n} \subseteq C_n$. However, this property is quite restrictive and, thus, the logit model is only appropriate when the alternatives are uncorrelated.

Ordinal logit (OL) model

As opposed to the MNL model which establishes a relationship between the covariates and the set of probabilities of the alternatives, the ordinal logit (OL) model is used to obtain expressions for the cumulative probabilities. An OL model for an ordinal response Y_i with K classes is defined by a set of $(K-1)$ equations as the last cumulative probability is necessarily equal to one.

$$logit(p_{ki}) = \log\left(\frac{p_{ki}}{1 - p_{ki}}\right) = \psi_k - \boldsymbol{\beta}^T \boldsymbol{X_{ki}} \tag{4.9}$$

where $k = 1, 2, ..., (K-1)$ and $p_{ki} = P(Y_i \leq y_k \mid X_i)$ is the cumulative probability. The parameters ψ_k are called thresholds or cut-offs, and are in increasing order ($\psi_1 < \psi_2 < ... < \psi_{K-1}$). An identification problem arises in the simultaneous estimation of the overall intercept β_0 (which is a part of the vector $\boldsymbol{\beta}$) and all the $(K-1)$ thresholds, which can be solved by either omitting the overall constant from the linear predictor ($\beta_0 = 0$) or fixing

the first threshold to zero ($\psi_1 = 0$). The former approach is used in the implementation for this study.

It should be noted that the vector of slopes in the linear predictor $\boldsymbol{\beta}$ is not indexed by the class index k, which indicates that the effects of the covariates are constant across response categories. This is known as the parallel regression assumption, which yields $(K-1)$ parallel lines while plotting $logit(p_{ki})$ against a covariate. We purposely introduced the negative sign before $\boldsymbol{\beta}$ in the model specification shown in Equation 4.9 so that the interpretation is according to intuition. With this model specification, we can imply that increasing a covariate with a positive slope would be associated with a rise in the probabilities of the higher classes. The cumulative probability for class k can be expressed in the following manner.

$$p_{ki} = \frac{\exp\left(\psi_k - \boldsymbol{\beta}^T \boldsymbol{X_{ki}}\right)}{1 + \exp\left(\psi_k - \boldsymbol{\beta}^T \boldsymbol{X_{ki}}\right)} \tag{4.10}$$

The OL model is also known as the proportional odds model because the parallel regression assumption implies that the ratio of odds for two classes is constant across response categories. This can be expressed as the following:

$$\frac{odds_{ki}}{odds_{kj}} = \exp\left[\boldsymbol{\beta}^T(\boldsymbol{X_j} - \boldsymbol{X_i})\right] \tag{4.11}$$

where odds for a class represent the proportionality of the odds of not exceeding that class, i.e. $odds_{ki} = p_{ki}/(1 - p_{ki})$. Readers interested in further details about the modeling of ordinal outcomes in the setting of choice theory should refer to Greene and Hensher (2010).

4.4.2 Machine learning models

Machine learning explores the study and construction of algorithms that can learn from and make predictions on data. There are multiple algorithms, known as classifiers, which use a mathematical function to map input data to a category through a concrete implementation. While we have explored several types of classifiers in the course of this study, we report four of the best-performing classifiers in this chapter for brevity. As the objective of vehicle ownership prediction is a supervised learning problem, only classification (and not clustering) algorithms are of interest (Kotsiantis et al., 2007).

Random Forest (RF)

Decision Tree (DT) refers to a non-parametric supervised learning method that aims to predict the value of a target variable by learning a set of simple if-then-else decision rules inferred from the features. The deeper the tree, the more complex the decision rules and the fitter the model. DT models are simple to understand and interpret, and can be visualized easily. They require little data preparation, can handle both numerical and categorical data, and perform well even if the assumptions are violated by the true model from which the data is drawn. However, they tend to easily overfit by creating over-complex trees that do not generalize well. They can also be unstable as small variations in the data can result in a completely different tree being generated. Moreover, locally optimal decisions made at each node using greedy algorithms cannot guarantee to return the globally optimal model (Anyanwu and Shiva, 2009).

Many of the drawbacks associated with DT models can be overcome by using them within an ensemble learner like the random forest (RF) algorithm. The goal of an ensemble method is to combine the predictions of multiple base estimators built with a given learning algorithm in order to improve the generalizability and robustness over a single estimator. In random forests, each tree in the ensemble is built from a sample drawn with replacement (i.e., a bootstrap sample) from the training set. Note that, when splitting a node during the construction of the tree, the split that is chosen is no longer the best split among all features. In contrast, the split that is picked is the best split among a random subset of the features. As a result of this randomness, the bias of the forest usually slightly increases (with respect to the bias of a single non-random tree). However, due to averaging, its variance also decreases, usually more than compensating for the increase in bias, hence yielding an overall better model. Our implementation combines classifiers by averaging their probabilistic predictions, instead of aggregating to a single class vote for each classifier as shown in Breiman (2001).

Neural Network (NN)

We use the Multi-layer Perceptron (MLP) algorithm that learns a function $f(.) : \mathbb{R}^m \rightarrow \mathbb{R}^o$ by training on a data set, where m is the number of input dimensions and o is the number of output dimensions (Haykin et al., 2009). There can be one or more non-linear layers, called hidden layers, between the input and output layer. Each neuron in the hidden layer

transforms the values from the previous layer with a weighted linear summation, followed by a non-linear activation function $g(.) : \mathbb{R} \rightarrow \mathbb{R}$. The output layer receives the values from the last hidden layer and transforms them into the output values using the final activation function. The MLP classifier minimizes the cross-entropy loss function and trains using gradient descent where the gradients are calculated using backpropagation. While the MLP classifier performs well with non-linear models and on-line learning, it is sensitive to feature scaling and requires tuning several hyperparameters such as the number of hidden neurons, layers, and iterations. In our implementation, we used two hidden layers (100 neurons and 6 neurons respectively) in a feed-forward neural network architecture with rectified linear unit (ReLU) as the activation function for the hidden layers and the softmax activation function for the output layer.

Support Vector Machine (SVM)

A support vector machine (SVM) constructs a set of hyperplanes in a high-dimensional space which can be used for classification. A good separation is achieved by the hyperplane that has the largest distance to the nearest training data points of any class, known as the margin, since a larger margin generally implies lower generalization error of the classifier. The objective function that the SVM tries to minimize can be expressed as follows:

$$J(\boldsymbol{\theta}, \boldsymbol{\theta_0}) = \frac{1}{n} \sum_{i=1}^{n} L_h \left[y^{(i)} (\boldsymbol{\theta}^T \cdot \boldsymbol{\phi(x)} + \boldsymbol{\theta_0}) \right] + \lambda \|\boldsymbol{\theta}\|^2 \tag{4.12}$$

where $\boldsymbol{\phi(x)}$ is the kernel or basis function of original feature vector $\boldsymbol{x}$; $\boldsymbol{y}$ is the label vector; λ is the regularization parameter; $L_h(.)$ is the hinge loss defined as $L_h(\boldsymbol{x}) = max(0, \boldsymbol{x})$; $\boldsymbol{\theta}$ is the weight vector; $\boldsymbol{\theta_0}$ is the bias vector. The first term of $J(\boldsymbol{\theta}, \boldsymbol{\theta_0})$ measures the loss which is a measure of the difference between the true and predicted labels, while the second term is the regularization term that prevents the classifier from overfitting the data and increases its generalizability. While other loss functions can be used, the most commonly used one is the hinge loss. Instead of using a linear kernel, we used a radial polynomial basis (rbf) kernel to extend generalizability and capture possible non-linear relationships in the data. We used the Pegasos algorithm to solve the optimization problem cast by the SVM, which has been found to extend well to non-linear kernels (Shalev-Shwartz et al., 2011).

Logistic Regression (LR)

Logistic Regression (LR) is a linear classification algorithm that models the probabilities describing the possible outcomes of a single trial using a logistic function. The objective function that the LR algorithm tries to minimize can be expressed as follows:

$$J(\boldsymbol{\theta}, \boldsymbol{\theta_0}) = -\frac{1}{n}\sum_{i=1}^{n}\left[y^{(i)}\log(h_\theta(\boldsymbol{x}^{(i)})) + (1-y^{(i)})\log(1-h_\theta(\boldsymbol{x}^{(i)}))\right] + \lambda\|\boldsymbol{\theta}\|^2 \tag{4.13}$$

This is different from the SVM objective function in the sense that the loss function is the cross-entropy loss, also known as the log loss, instead of the hinge loss. These smooth functions make it easy to calculate the gradient and minimize loss. There are two techniques to implement this algorithm. The first is the more common one-vs-rest approach where separate binary classifiers are trained for all classes. The other is the true multinomial logistic regression approach which provides better calibrated probability estimates (Caruana and Niculescu-Mizil, 2006). We implemented the latter for this research using the stochastic average gradient descent solver for faster convergence.

Need for cross-validation

From the algorithm descriptions above, it is evident that the SVM and LR models involve a regularization parameter. Additional analysis is required to obtain the optimal value of the hyperparameter λ (which influences regularization) as it depends both on the training data set and the learning algorithm. Therefore, 10-fold cross-validation was used to randomly create train and validation data sets from the overall training data set (with replacement and shuffling) and specify a grid of values for λ. After extensive testing with different grid sizes, we decided to use $\lambda \in [0.01, 10]$ with a step size of 0.01.

For each value of λ within this specified grid and each train data set (90% of training samples), the learning algorithm was used to obtain a prediction accuracy score for the validation data set (10% of training samples). Consequently, we obtained two 1000 x 10 matrices containing train accuracy scores and validation accuracy scores for each of these runs. By averaging the score over the 10 cross-validation runs, a 1000 x 1 column vector of average scores was obtained. Since the validation score needs to be maximized, λ is said to be optimal for the highest score in the column vector of validation scores. Thus, we obtained an optimal hyperparameter λ^* for the training data set with which the learning algorithm

was used on the entire training data set to obtain optimal weights $\boldsymbol{\theta}^*$. Finally, we employed the optimal λ^* and $\boldsymbol{\theta}^*$ to provide predictions for the test data set.

4.4.3 Variable specifications

In addition to understanding the variations between different modeling techniques, it is necessary to recognize the importance of appropriate *variable specification* (as the econometricians call it), or *feature selection* (more frequently used in the machine learning community). It is common practice to use features in a standard manner (after appropriate cleaning and scaling, of course) in ML models without much thought given towards interpretability. However, econometric model specifications are developed through an iterative process of adding different variables to the constants-only model and removing variables that turned out to be statistically insignificant. Moreover, we combined variables when we found that their effects on the model were not statistically different. In general, such a model construction exercise is guided by intuitive consideration on the part of the modeler, and parsimony in the representation of covariate effects. We start with the standard specification, and build on that sequentially to enhance the specification through econometrics and machine learning techniques in the following sub-sections.

Standard specification

The raw data is cleaned and scaled appropriately, following which features are created in a standard manner without considering the economic or non-linear effects. An example of such a feature is the percentage of individuals holding a car license in a household. Consideration of a feature in this form for an econometric model implies that doubling the percentage (i.e., if twice as many individuals were to hold car licenses) would double the probability of that household owning a car. This implication is flawed as we would not expect a linear relationship between such a feature and the outcome variable (i.e., car ownership). However, ML models are able to successfully discover and account for non-linearities in feature space automatically, which is why modelers rarely go beyond the standard specification. On the contrary, econometricians need to be more innovative with feature construction.

Econometric specification

We build on the standard specification by considering variables that are of economic importance for this application. Such features should be represented by the modeler in the form that makes most sense for interpretation, rather than in the aforementioned standard form. Continuing with the example of car licenses, it should be clear to the reader that the feature has a marginal effect on the probability of household car ownership. Therefore, the feature should be transformed into a log-type function that captures the decreasing marginal utility of the effect, rather than a percentage function. We hypothesize that the use of econometric variable specification will enhance the model fit and improve prediction accuracy of econometric models.

ML-enhanced specification

While the aforementioned specifications are employed in current practice, studies in the literature do not go beyond them. We argue that there is further improvement to be made. Recall that econometric models are formulated as linear-in-parameters functions, which implies a linear relationship between the covariates and the outcome. This is not always true, as significant non-linearities may exist in the data. It is important to draw a distinction between this specification and the previous specification here. The econometric specification accounts for non-linearities in the variable interpretation through economic hypotheses and is not dependent on the data. However, the ML-enhanced specification builds on that and tries to also account for non-linearities due to correlations among independent variables.

In high-dimensional feature space, it is challenging to detect non-linearities manually due to the large size of the feature matrix, especially as x-y scatter plots are not feasible in multi-class classification problems. Therefore, we propose a two-stage framework to create an ML-enhanced specification. The first stage conducts feature selection of the most valuable features (in terms of information captured and of use to the model) using permutation feature importance. The necessity of analyzing variable importance for effective modeling was highlighted by Hagenauer and Helbich (2017) through a travel behavior application. While a technique to calculate feature importance was first proposed by Breiman (2001) for random forests, Fisher et al. (2018) proposed a model-agnostic version of the feature importance and termed it as model reliance. We used the following algorithm proposed by

the latter to calculate model reliance for this study. Note that f is the trained model, X is the feature matrix, y is the target vector, and $L(y, f)$ is the error measure.

1. Estimate the original model error $e_i = L(y, f(X))$
2. For each feature $j = 1, 2, ..., K$:
   - Generate feature matrix X_j by permuting feature j in the data X. This breaks the association between feature j and the true outcome y.
   - Estimate error $e_j = L(y, f(X_j))$ based on the predictions of the permuted data.
   - Calculate permutation feature importance $\tau_j = e_j - e_o$.
3. Sort features by descending τ_j.

The second stage is to uncover the non-linearities in the feature space corresponding only to the most important features (which were identified from the first stage) through the use of Accumulated Local Effects (ALE) plots. For applications with low-dimensional feature matrices, modelers can use ALE plots for all features instead of a few selected ones that we were constrained to consider due to high-dimensionality in this application. Accumulated local effects describe how features influence the prediction of a model on average (Apley, 2016). The uncentered effect is estimated first.

$$\widehat{\tilde{f}}_{j,ALE}(x) = \sum_{k=1}^{k_j(x)} \frac{1}{n_j(k)} \sum_{i:x_j^{(i)} \in N_j(k)} \left[f(z_{k,j}, x_j^{(i)}) - f(z_{k-1,j}, x_j^{(i)}) \right] \tag{4.14}$$

The differences in predictions are calculated first, wherein the feature of interest is replaced with grid values z. The difference in prediction is the effect the feature has for an individual instance in a certain interval. The summation on the right adds up the effects of all instances within an interval which appears in the formula as neighborhood $N_j(k)$. We divide this sum by the number of instances in this interval to obtain the average difference of the predictions for this interval. The summation symbol on the left means that we accumulate the average effects across all intervals. The (uncentered) ALE of a feature value that lies, for example, in the third interval is the sum of the effects of the first, second and third intervals. Next, we center the effect so that the mean is zero.

$$\widehat{f}_{j,ALE}(x) = \widehat{\tilde{f}}_{j,ALE}(x) - \frac{1}{n} \sum_{i=1}^{n} \widehat{\tilde{f}}_{j,ALE}(x_j^{(i)}) \tag{4.15}$$

The value of the ALE can be interpreted as the main effect of the feature at a certain value compared to the average prediction of the data. For example, an ALE estimate of -0.5 at $x_j = 1$ means that when the jth feature has a value equal to one, then the prediction is lower by 0.5 compared to the average prediction. One of the major advantages of ALE plots, especially over partial dependence plots (PDPs), is that they are unbiased and account for correlations between features. This makes ALEs much more applicable to real-world datasets compared to PDPs as features are usually correlated to some extent.

Using this two-stage framework, we can successfully (a) identify the most important features in high-dimensional feature space, and (b) uncover the intervals along which these selected features have non-linear effects on the outcome. We can then use linear transformations to model these non-linearities. While there are several methods to do that, we employ a piecewise-linear transformation of such variables, also known as a linear spline transformation, in this study. For example, if the ALE of income on car ownership is seen to have three distinct intervals, we can create three separately defined linear variables that represent the non-linear effect of income in a combined manner.

4.4.4 Model assessment metrics

We consider a variety of metrics to assess model validity and compare performance across models. These metrics belong to three umbrella categories. First, *goodness-of-fit* metrics determine how well the model can explain the variations in the training data. Second, metrics that evaluate *prediction accuracy* indicate how well the model performs in out-of-sample prediction when used on a test data set that it was not trained on. Finally, we also consider the *predicted market shares* as a consideration of the model over- or under-predicting certain alternatives. More details about these metrics are provided in the following sub-sections.

It is also worth noting here that consideration of a baseline model is necessary to compare the performance of different modeling approaches. We used the zero-information, also known as the zero-R, method to construct the baseline. This method assumes a prediction of the class with the maximum frequency, i.e. the majority class, for all data samples.

Goodness-of-fit

A generic measure of goodness-of-fit for any model is to examine its performance on the training data set. We consider the *average training accuracy* of the model as such a met-

ric. The average training accuracy of a model is defined as the percentage of samples in the training data set that the model is able to classify correctly. While a higher value is better, it also implies a risk of the model overfitting the training data, which motivates our consideration of other metrics.

$$\text{Avg. train accuracy} = \frac{N_{train}(y = \widehat{y})}{N_{train}} \tag{4.16}$$

Another measure of goodness-of-fit is *McFadden's pseudo R-squared*; however, it is relevant only to econometric models (McFadden, 1973). It compares a model with predictors (M_{Full}) to a model without predictors ($M_{Intercept}$), i.e. using only the intercepts as explanatory variables, and uses the log-likelihood values of these two models. The ratio of the likelihoods suggests the level of improvement over the intercept model offered by the full model. Thus, a small ratio of log-likelihoods indicates that the full model is a far better fit than the intercept model. If comparing two models on the same data, McFadden's pseudo R-squared would be higher for the model with the greater likelihood.

$$\rho^2 = 1 - \frac{\ln \widehat{L}(M_{Full})}{\ln \widehat{L}(M_{Intercept})} \tag{4.17}$$

However, overly complex models tend to overfit the data, which is why a penalty must be included for the number of parameters. If the predictors in the model are effective, then the penalty will be small relative to the added information of the predictors. However, if a model contains predictors that do not add sufficiently to the model, then the penalty becomes noticeable and the adjusted R-squared can decrease with the addition of a predictor, even if the R-squared increases slightly. Accordingly, *McFadden's adjusted psuedo R-squared* is defined.

$$\bar{\rho}^2 = 1 - \frac{\ln \widehat{L}(M_{Full}) - K}{\ln \widehat{L}(M_{Intercept})} \tag{4.18}$$

These metrics are similar to the R^2 and *adjusted* R^2 metrics for OLS models. While higher values are better, they are not as high as OLS metrics. McFadden (1977) reports that values between 0.2 and 0.4 indicate "excellent model fit".

Prediction accuracy

As predictions are made on out-of-sample data, it is worth keeping in mind that the following metrics apply only when the model is used on the test data set. We define the *average test accuracy* in a fashion similar to the average train accuracy, with the exception of the data set.

$$\text{Avg. test accuracy} = \frac{N_{test}(y = \widehat{y})}{N_{test}} \tag{4.19}$$

While accuracy is a good overall measure, it can be critiqued because it assumes equal costs for both types of errors (false positives and false negatives). Therefore, we should consider other measures such as precision, recall and F-measure. *Recall* is defined as the ratio of the total number of correctly classified positive examples and the total number of positive examples. High recall indicates the class is correctly recognized (small number of false negatives).

$$\text{Recall} = \frac{N_{true(+)}}{N_{true(+)} + N_{false(-)}} \tag{4.20}$$

Precision is the ratio of the total number of correctly classified positive examples and the total number of predicted positive examples. High precision indicates an example labeled as positive is indeed positive (small number of false positives).

$$\text{Precision} = \frac{N_{true(+)}}{N_{true(+)} + N_{false(+)}} \tag{4.21}$$

Since we now have two separate measures (precision and recall), it helps to have a measurement that represents both of them. We calculate an *F-measure* which uses harmonic mean in place of arithmetic mean as the harmonic mean punishes the extreme values more than the arithmetic mean.

$$\text{F-measure} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4.22}$$

Predicted market shares

Most studies in the literature draw the line at the above two metrics and fail to consider predicted market shares. We posit that examination of alternative-specific market shares are just as important, especially in applications with multiple alternatives where over- or under-prediction of certain categories can have serious implications despite an overall high

model accuracy. We consider two types of errors as performance metrics to compare the predicted market shares with the actual market shares. The first is the *mean absolute error (MAE)*, which measures the average magnitude of the errors in a set of predictions without considering their direction.

$$MAE = \frac{\sum_j |y_j - \widehat{y_j}|}{J} \tag{4.23}$$

The second type of error is the *root mean squared error (RMSE)*, which is the square root of the average of squared differences between the predictions and actual observations. Note that the RMSE is particularly useful when large errors are particularly undesirable (such as skewed predicted market shares in our application) because large errors are given a relatively higher weight.

$$RMSE = \sqrt{\frac{\sum_j (y_j - \widehat{y_j})^2}{J}} \tag{4.24}$$

4.4.5 Synthetic sample generation

As discussed earlier, the observed data needs to be augmented with synthetic samples for minority classes to account for imbalanced classes. While we tested several sample generation algorithms, we report six of the best-performing algorithms in this chapter for brevity. Details of these algorithms are provided below.

Adaptive Synthetic (ADASYN) sampling

ADASYN uses a weighted distribution for different minority class samples according to their level of difficulty in learning. This allows for more synthetic data to be generated for minority class samples that are harder to learn compared to those minority class samples that are easier to learn. The ADASYN approach helps reduce the bias introduced by the class imbalance, and adaptively shifts the classification decision boundary toward the difficult examples (He et al., 2008).

Synthetic Minority Over-sampling Technique (SMOTE)

SMOTE builds on previous work of under-sampling the majority class, which achieved a reasonable improvement in classifier performance, by combining over-sampling minority classes and under-sampling the majority class. Its main idea is to form new minority class examples by interpolating between several minority class examples that lie together. Chawla et al.

(2002) found that SMOTE significantly improves classifier performance (in ROC space) compared to loss ratio variation in Ripper and class prior modification in Naïve Bayes.

SMOTE with Tomek Link cleaning

Batista et al. (2003) extended SMOTE by identifying that class clusters are often not well-defined since some majority class examples might be invading the minority class space; the opposite can also be true but is less frequently observed. Therefore, they identified *Tomek links* (Tomek, 1976) and removed them from the augmented data set to create better-defined class clusters. Consider two examples x and y belonging to separate classes, where $d(x, y)$ is the distance between x and y. A (x, y) pair can be called a Tomek link if no case z exists where $d(x, z) < d(x, y)$ or $d(y, z) < d(y, x)$. The elimination of noise and borderline examples through Tomek links allows for creation of well-defined class clusters.

SMOTE-IPF

Sáez et al. (2015) further extended SMOTE through an iterative ensemble-based noise filter called Iterative-Partitioning Filter (IPF). This extension, known as SMOTE-IPF, can overcome the problems produced by noisy and borderline examples in imbalanced data sets. The method was tested on a set of synthetic data sets with different levels of noise and shapes of borderline examples as well as real-world data sets, and found to perform better than other SMOTE generalizations for all the considered scenarios.

Proximity Weighted Synthetic (ProWSyn) Over-sampling Technique

As discussed earlier, some synthetic sampling approaches create weight values for original minority samples based on their importance and distribute the synthetic samples according to weight values. However, Barua et al. (2013) critique most of the existing algorithms by arguing that the generated weight values are inappropriate, which result in a poor distribution of the generated synthetic samples. They propose a new synthetic oversampling algorithm, Proximity Weighted Synthetic Oversampling Technique (ProWSyn). This algorithm generates effective weight values for the minority data samples based on the sample's proximity information, i.e., distance from the boundary which results in a proper distribution of generated synthetic samples across the minority data set.

Polynomial Fitting

When the trained classifier accurately classifies the majority class and marginalizes the minority classes, the True Negatives rate (TNr) will be very high while the True Positives rate (TPr) will be low. Gazzah and Amara (2008) propose the use of polynomial fitting functions that can improve TPr without much sacrifice in TNr. While they showcased four approaches, only the polynomial curve topology approach has been implemented in this study. In this approach, for each feature of the minority class matrix, we pass a curve through a set of minority instances in such a way that the estimated curve shows the best trend in these instances. To define the "trend curve", we compute the coefficients of a polynomial $p(x)$ of degree n that fits the features using the least squares method.

$$p(x) = p_1 \cdot x^n + p_2 \cdot x^{n-1} + p_3 \cdot x^{n-2} + ... + p_n \cdot x + p_{n+1} \tag{4.25}$$

4.5 Results & Discussion

First, we describe the insights obtained from the model selection framework by considering the results of the three variable specifications in a sequential manner. Second, the impact of handling low-sample alternatives through synthetic sample generation techniques is examined through the lens of model performance. Finally, we discuss how our proposed framework improves upon current methods and is successfully effective in forecasting market shares of new mobility options that did not exist in the training data set.

4.5.1 Model selection

Insights from comparing econometric models such as multinomial logit (MNL) and ordinal logit (OL) with machine learning models such as random forest (RF), neural network (NN), support vector machine (SVM) and logistic regression (LR) are described below. We discuss both model scenarios (i.e., estimation and prediction) for the standard specification. As we move on to more advanced specifications, we focus specifically on the prediction scenario as this study is motivated by the prediction of new mobility market shares. Results for the estimation scenario with advanced specifications can be found in Appendix B.

Standard variable specification

Results of the model goodness-of-fit and prediction accuracy with the standard variable specification are summarized in Table 4.3. We find that all the models considered in this study perform better than the baseline for both estimation and prediction scenarios. While this might seem trivial, this result validates the use of these models. We find that MNL performs at par with RF and NN in estimation, but ML models perform much better in prediction. Moreover, MNL has a much better model fit and outperforms OL in prediction accuracy for both scenarios. This corroborates similar findings of unordered response structures performing better than their ordered counterparts by Bhat and Pulugurta (1998) across three US data sets and one Dutch data set of household car ownership.

Table 4.3: Gof and prediction accuracy with standard specification

Metric	Baseline	MNL	OL	RF	NN	SVM	LR
Estimation Scenario [a]							
Avg. train accuracy	0.55	0.74	0.71	1.00	0.76	0.73	0.70
Avg. test accuracy	0.55	0.73	0.71	0.73	0.73	0.71	0.70
Execution time (sec)	-	16.0	21.0	5.77	14.61	13.69	14.87
Avg. precision	0.30	0.71	0.64	0.72	0.69	0.62	0.60
Avg. recall	0.55	0.73	0.71	0.73	0.73	0.71	0.70
Avg. F-measure	0.39	0.71	0.67	0.71	0.70	0.66	0.64
McFadden's $\bar{\rho}^2$	-	0.632	0.489	-	-	-	-
Prediction Scenario [b]							
Avg. train accuracy	0.55	0.74	0.71	1.00	0.76	0.73	0.70
Avg. test accuracy	0.54	0.72	0.71	0.74	0.74	0.72	0.71
Execution time (sec)	-	29.0	18.0	0.72	2.57	11.88	7.12
Avg. precision	0.29	0.70	0.65	0.74	0.72	0.63	0.62
Avg. recall	0.54	0.72	0.71	0.74	0.74	0.72	0.71
Avg. F-measure	0.38	0.70	0.67	0.71	0.72	0.67	0.66
McFadden's $\bar{\rho}^2$	-	0.629	0.486	-	-	-	-

[a] Using HITS 2008 data with 80/20 train/test split

[b] Using HITS 2008 data to predict HITS 2012 shares

Our hypothesis that econometric models are better used for estimation and ML models are better used for prediction is confirmed when we examine the predicted market shares

in Table 4.4. While all models have a lower aggregate RMSE than the baseline, MNL has the lowest RMSE for both estimation and prediction scenarios. Since MNL was at par with ML models for estimation, its low RMSE makes it the best-performing model for estimation. Despite their relatively higher RMSE for prediction, ML models had higher prediction accuracy. Therefore, there is room for improvement in MNL, which forms the basis for considering more advanced variable specifications. An additional point of note here is related to the predicted market shares for the low-sample alternatives like categories 2 and 4. We find that only MNL is able to predict non-zero, albeit abysmally low, market shares for these alternatives. All other models, including OL, fail to produce meaningful predictions. While we will discuss this in further detail in a later section, it is worth calling attention to at this stage.

Table 4.4: Predicted market shares with standard specification

Category	**Actual**	**Baseline**	**MNL**	**OL**	**RF**	**NN**	**SVM**	**LR**
Estimation Scenario [a]								
0	54.86	100	59.15	61.99	56.34	52.46	69.50	67.17
1	5.37	0	6.07	0	2.97	6.35	0	0
2	1.67	0	0.05	0	0	0	0	0
3	32.28	0	33.73	36.34	40.13	41.19	30.5	32.83
4	1.44	0	0.12	0	0	0	0	0
5	4.38	0	0.88	1.68	0.56	0	0	0
MAE	-	15.05	2.15	3.73	3.11	3.30	4.88	4.29
RMSE	-	22.85	2.50	4.25	3.85	4.28	6.71	5.84
Prediction Scenario [b]								
0	54.28	100	55.95	58.30	51.37	53.13	66.76	63.61
1	4.39	0	3.58	0	1.71	5.58	0	0
2	2.03	0	0.16	0	0	0	0	0
3	33.96	0	39.11	39.93	46.86	41.22	33.24	36.39
4	1.00	0	0.34	0	0	0	0	0
5	4.34	0	0.86	1.77	0.07	0.07	0	0
MAE	-	15.24	2.27	3.33	4.30	2.82	4.16	3.92
RMSE	-	23.41	2.77	3.71	5.85	3.62	5.77	4.76

[a] Using HITS 2008 data with 80/20 train/test split

[b] Using HITS 2008 data to predict HITS 2012 shares

As mentioned earlier, we will consider only the prediction scenario as we move on to more advanced specifications. Moreover, we select the best-performing models from these results to reduce the comparative dimensionality that makes these tables difficult to read. The best econometric model is clearly MNL, while it is difficult to ascertain a single best machine learning model. Therefore, due to comparatively similar performance, we select both RF and NN for further consideration.

Econometric variable specification

We build on the standard variable specification by creating new features that have economic interpretation based on modelers' hypotheses to create the econometric variable specification. In this particular application of modeling household vehicle availability, we selected three types of effects in our feature construction process. First, based on the seminal paper by Lerman (1976), we assume that children, teenagers, and seniors in the household have an effect on the probability of the household purchasing a vehicle that is similar to the law of diminishing marginal utility. While it is expected that having a child in the household increases the probability of car purchase, the marginal effect assumption implies that the probability does not increase linearly with the increase in the number of children. Rather, the rate of increase starts to decrease and finally reaches zero after a certain threshold.

In addition to these individuals with diverse mobility needs, we assume that individuals holding a vehicle license also have a marginally diminishing effect. While we had modeled these features as percentages in the household for the standard variable specification $f(x)$, we use a log-function to model the marginally diminishing effect, as shown by $g(x)$ below. For the license-related features, N_x represents the number of adults in the household. However, for features related to children, teenagers and seniors, N_x represents the total household size.

$$f(x) = \frac{x}{N_{hh}} \quad \rightarrow \quad g(x) = \begin{cases} 1 + \frac{\log(x)}{N_x} & x \geq 1 \\ 0 & x = 0 \end{cases} \tag{4.26}$$

Second, we assume that children have a marginally diminishing effect on the household income. While we used total household income in the standard specification, we calculate a per-capita income measure accounting for this marginal effect.

$$f(x) = I_{hh} \quad \rightarrow \quad g(x) = \frac{I_{hh}}{N_{adult} + \log(1 + N_{child})} \tag{4.27}$$

Finally, we create binary variables to tease out impacts of different levels of access to public transit and primary schools. We had used the distances to the nearest bus stop, MRT station, and primary school directly for the standard specification. However, we opt for a more fine-grained approach for the econometric specification. The Singaporean government ensures that public linkways are provided to public amenities and public transport up to a radius of 400 meters from public housing through the Walk2Ride program (LTA, 2013). Therefore, we consider two buffers with radii of 200 and 400 meters respectively for bus stops.

Since MRT stations are lower in number and occur less frequently, we select typical walking distances of 400 meters and 800 meters as buffer radii from the literature. Similar station catchment area buffers have been used by Zhao et al. (2013) in China, Lee et al. (2005) in Korea, Advani and Tiwari (2005) in India, and Hess and Almeida (2007) in the United States. While considering access to primary schools, we selected one and two kilometers as buffer radii based on similar radii being used as preferential criteria by the Ministry of Education while allocating vacant spots in Singaporean primary schools (MoE, 2019).

The goodness-of-fit and prediction accuracy of the best-performing models in the standard specification, i.e., MNL, RF and NN, are reported in Table 4.5. For ease of comparison, we report metrics for both the standard and econometric specifications in the same table. Our hypothesis that the econometric specification would greatly benefit econometric models is proven true. We find that both the average test accuracy and McFadden's $\bar{\rho}^2$ show significant improvement for MNL. However, the predictive power of ML models is not affected by this new specification, which is to be expected.

The predicted market shares are reported in a similar comparative fashion in Table 4.6. We find that the increased predictive power of MNL has come at the price of skewed disaggregate market shares. This is probably because of non-linearities in the data that are not being appropriately modeled in the econometric specification, which motivates our consideration of the ML-enhanced specification. While RF has not been affected in this aspect, NN shows a fair bit of improvement in aggregate RMSE. Although this seems promising, it is worth keeping in mind that neural networks are challenging to interpret and, therefore,

Table 4.5: Gof and prediction accuracy with econometric specification

Metric	Baseline	MNL[a]	RF[a]	NN[a]	MNL[b]	RF[b]	NN[b]
			Prediction Scenario [c]				
Avg. train accuracy	0.55	0.74	1.00	0.76	0.75	1.00	0.76
Avg. test accuracy	0.54	0.72	0.74	0.74	0.74	0.74	0.74
Execution time (sec)	-	29.0	0.72	2.57	29.0	0.76	1.46
Avg. precision	0.29	0.70	0.74	0.72	0.72	0.73	0.69
Avg. recall	0.54	0.72	0.74	0.74	0.74	0.74	0.74
Avg. F-measure	0.38	0.70	0.71	0.72	0.72	0.71	0.71
McFadden's $\bar{\rho}^2$	-	0.629	-	-	0.640	-	-

[a] Standard Specification; [b] Econometric Specification

[c] Using HITS 2008 data to predict HITS 2012 shares

have limited value in prediction exercises where the alternative of interest is missing from the training data.

Table 4.6: Predicted market shares with econometric specification

Category	Actual	Baseline	MNL[a]	RF[a]	NN[a]	MNL[b]	RF[b]	NN[b]
				Prediction Scenario [c]				
0	54.28	100	55.95	51.37	53.13	52.23	50.97	56.31
1	4.39	0	3.58	1.71	5.58	4.98	1.95	4.94
2	2.03	0	0.16	0	0	0	0	0
3	33.96	0	39.11	46.86	41.22	41.98	47.01	38.75
4	1.00	0	0.34	0	0	0.55	0	0
5	4.34	0	0.86	0.07	0.07	0.25	0.08	0
MAE	-	15.24	2.27	4.30	2.82	2.87	4.35	2.46
RMSE	-	23.41	2.77	5.85	3.62	3.87	5.92	2.92

[a] Standard Specification; [b] Econometric Specification

[c] Using HITS 2008 data to predict HITS 2012 shares

ML-enhanced variable specification

The ML-enhanced specification builds on the econometric specification by identifying non-linearities in the feature space that have not been adequately modeled through the linear-

in-parameters form of econometric models. Due to the feature space spanning over 70 dimensions, we calculated model reliance for all features in order to select the most important features for further consideration. The model reliance values of the top eight features are shown in Figure 4-2. We observe that only two features — *marginal effect of car licenses* and *marginal effect of per-capita household income* — have model reliance values greater than 0.10. All other features have model reliance values less than 0.05. Therefore, we consider only these two features for extraction of non-linearities.

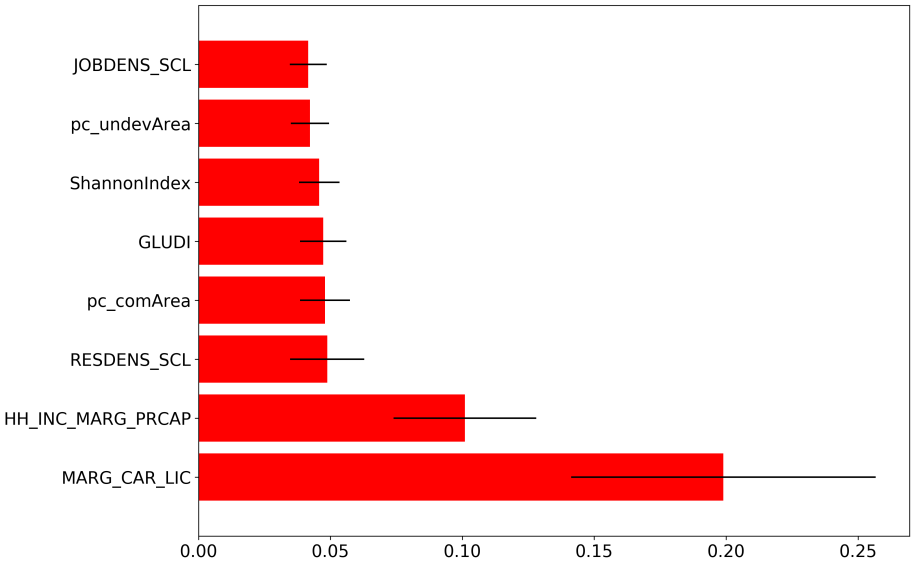
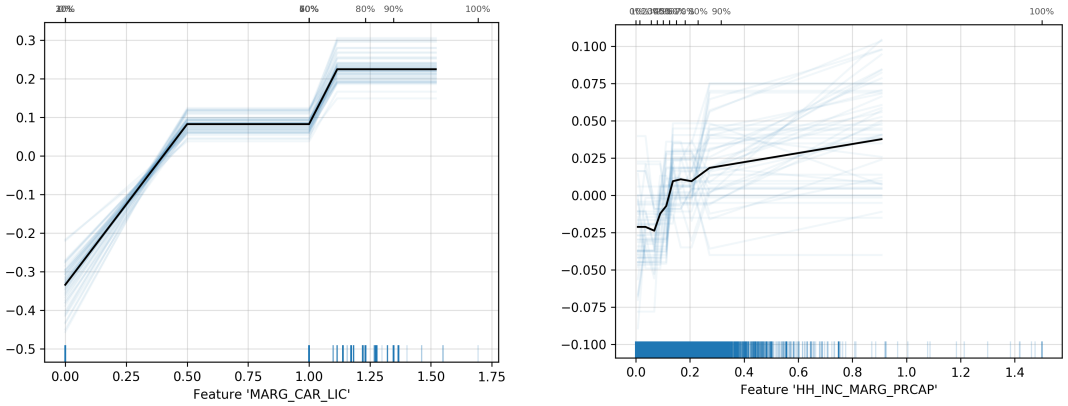


Figure 4-2: Model reliance for top eight features in econometric specification

We plot centered ALEs for these two features using 50 Monte-Carlo simulations. Figure 4-3(a) shows the ALE plot for the marginal effect of car licenses. We find that the relationship is piece-wise linear with four sub-regimes clearly visible in the ALE plot. Therefore, we define three new features that capture the linear effect of three sub-regimes, maintaining the fourth one as a reference. The ALE plot for the marginal effect of per-capita income is shown in Figure 4-3(b), where the relationship is evidently non-linear. However, in an attempt to convert it to a piece-wise linear function, we identify six sub-regimes and, consequently, define five new features for these sub-regimes keeping one of them as a reference.

We compare the econometric and ML-enhanced specifications based on model goodness-of-fit and prediction accuracy metrics in Table 4.7. We find that enforcing linear relationships between the covariates and the outcome results in a decrease in all metrics for MNL, while



(a) Car licenses

(b) Per-capita income

Figure 4-3: Centered Accumulated Local Effects (ALEs) for top two features

those for ML models remain the same or increase slightly. This is counter-intuitive as we would expect the ML-enhanced specification to further enhance MNL by treating non-linearities adequately. The decrease in model fit, as evidenced by McFadden's $\bar{\rho}^2$, can be attributed to creating linear sub-regimes that are too fine. Having five piece-wise linear functions for per-capita income is perhaps an overkill, especially when low-income households are not expected to exhibit significant variations in behavior.

Table 4.7: Gof and prediction accuracy with ML-enhanced specification

Metric	**Baseline**	**MNL**^{a}	**RF**^{a}	**NN**^{a}	**MNL**^{b}	**RF**^{b}	**NN**^{b}
Prediction Scenario ^{c}							
Avg. train accuracy	0.55	0.75	1.00	0.76	0.74	1.00	0.76
Avg. test accuracy	0.54	0.74	0.74	0.74	0.73	0.74	0.75
Execution time (sec)	-	29.0	0.76	1.46	21.0	0.63	2.16
Avg. precision	0.29	0.72	0.73	0.69	0.7	0.73	0.73
Avg. recall	0.54	0.74	0.74	0.74	0.73	0.74	0.75
Avg. F-measure	0.38	0.72	0.71	0.71	0.71	0.72	0.72
McFadden's $\bar{\rho}^2$	-	0.640	-	-	0.614	-	-

^{a} Econometric specification; ^{b} ML-enhanced specification

^{c} Using HITS 2008 data to predict HITS 2012 shares

The predicted market shares are examined in a similar comparative fashion in Table 4.8. In contrast to the model goodness-of-fit, MNL is the best-performing model with regard to aggregate RMSE. While it has the lowest RMSE among all considered models

with the ML-enhanced specification, its performance also improves in comparison to the econometric specification. Thus, we have shown that the two advanced specifications were necessary to improve the overall model performance of MNL and make it at par with the machine learning models. At the conclusion of the model selection exercise, we find that the MNL is the best-performing econometric model and the ML-enhanced variable specification is the best specification that enables MNL to predict market shares with the least error, while maintaining an overall prediction accuracy that is almost as good as that of the best-performing machine learning models such as RF and NN.

Table 4.8: Predicted market shares with ML-enhanced specification

Category	**Actual**	**Baseline**	**MNL**[a]	**RF**[a]	**NN**[a]	**MNL**[b]	**RF**[b]	**NN**[b]
				Prediction Scenario [c]				
0	54.28	100	52.23	50.97	56.31	53.71	49.89	48.95
1	4.39	0	4.98	1.95	4.94	5.30	2.34	3.64
2	2.03	0	0	0	0	0	0	0.01
3	33.96	0	41.98	47.01	38.75	40.66	47.69	47.26
4	1.00	0	0.55	0	0	0.15	0	0
5	4.34	0	0.25	0.08	0	0.19	0.07	0.14
MAE	-	15.24	2.87	4.35	2.46	2.54	4.58	4.43
RMSE	-	23.41	3.87	5.92	2.92	3.37	6.26	6.17

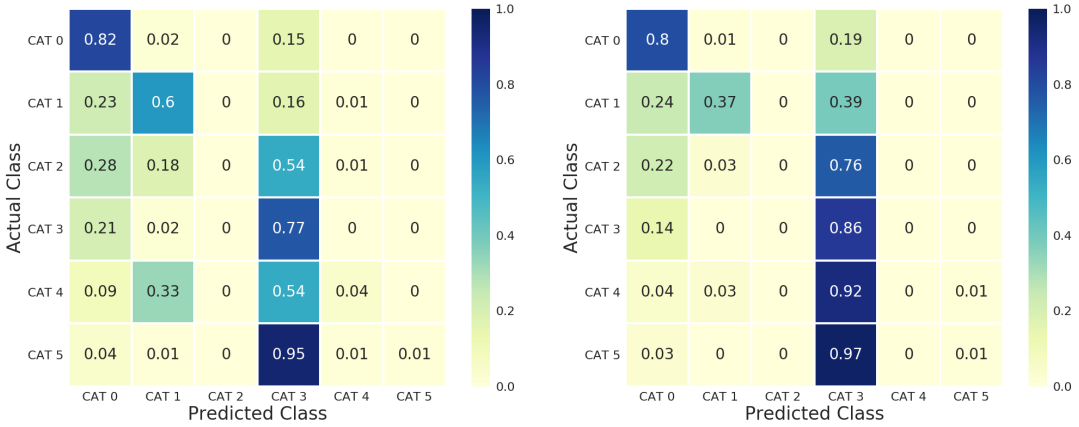
[a] Econometric specification; [b] ML-enhanced specification

[c] Using HITS 2008 data to predict HITS 2012 shares

4.5.2 Handling low-sample alternatives

While we were able to enhance the predictive performance of MNL using advanced specifications in the previous section, the disaggregate predictions of low-sample alternatives such as categories 2 and 4 remained abysmally low (and even zero for some cases). The confusion matrices for MNL and RF are shown in Figure 4-4. We observe that the major diagonal values, which represent the percentage of samples correctly classified for each alternative, are high only for the majority classes, i.e., categories 0 and 3. More importantly, we find that almost all minority class samples are being classified as category 3, which reinforces our earlier discussion about minority classes being perceived as noise by the model. To

address this imbalanced classification phenomenon, we used synthetic sample generation algorithms to enhance the training data set with artificially generated samples for minority classes. These six algorithms are ADASYN (abbreviated as ADA), SMOTE (SM), SMOTE with Tomek Link cleaning (SMT), SMOTE-IPF (IPF), ProWSyn (PRO), and Polynomial Fitting (PLF).



(a) Multinomial Logit (b) Random Forest

Figure 4-4: Confusion matrices for prediction using ML-enhanced specification

Table 4.9 compares the predictive accuracy of a random forest model with the ML-enhanced specification without any sample adjustment (w/o SA) with the same model fitted on an augmented training data set. We select the random forest model to illustrate the performance of these algorithms because RF had the highest RMSE for predicted market shares in the previous sub-section. Therefore, this is akin to using a worst-case scenario. We find that generating synthetic samples does not adversely affect overall prediction accuracy of the model. While some algorithms exhibit better F-measures, the test accuracy remains constant across all algorithms.

Looking at the predicted market shares in Table 4.10, we observe that the aggregate RMSE decreases for all algorithms by a significant margin (values range from 33% to 51%). Therefore, we can conclude that generating synthetic samples to augment minority classes in the training data set is beneficial to produce more accurate disaggregate predicted market shares without sacrificing overall accuracy. In addition to validating our hypothesis, this exercise provides the best synthetic sample generation algorithm for our application, *ProWSyn*, which had a reduction of 51% in RMSE.

Table 4.9: Gof and prediction accuracy for RF with sample adjustment

Metric	Baseline	w/o SA	ADA	SM	SMT	IPF	PRO	PLF
				Prediction Scenario [a]				
Avg. train accuracy	0.55	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Avg. test accuracy	0.54	0.74	0.75	0.74	0.74	0.74	0.74	0.74
Execution time (sec)	-	0.63	5.26	4.20	22.09	102.64	21.18	12.88
Avg. precision	0.29	0.73	0.74	0.74	0.74	0.74	0.74	0.73
Avg. recall	0.54	0.74	0.75	0.74	0.74	0.74	0.74	0.74
Avg. F-measure	0.38	0.72	0.73	0.73	0.73	0.73	0.74	0.72

[a] Using HITS 2008 data to predict HITS 2012 shares

Table 4.10: Predicted market shares for RF with sample adjustment

Category	Actual	w/o SA	ADA	SM	SMT	IPF	PRO	PLF
				Prediction Scenario [a]				
0	54.28	49.89	50.10	49.72	50.08	50.95	51.07	52.86
1	4.39	2.34	6.40	6.63	6.29	6.35	6.13	3.86
2	2.03	0	0.11	0.15	0.20	0.21	0.29	0.02
3	33.96	47.69	41.92	41.94	41.60	40.73	39.69	42.89
4	1.00	0	0.52	0.71	0.84	0.55	1.22	0.27
5	4.34	0.07	0.94	0.86	0.98	1.22	1.59	0.11
MAE	-	4.58	3.33	3.41	3.18	2.91	2.57	2.98
RMSE	-	6.26	4.09	4.19	3.96	3.51	3.08	4.17
ΔRMSE [b]	-	-	-35%	-33%	-37%	-44%	-51%	-33%

[a] Using HITS 2008 data to predict HITS 2012 shares

[b] The percentage change in RMSE is calculated using a baseline w/o sample adjustment.

4.5.3 Forecasting new mobility market shares

As a recap, the best performances were shown by:

- **Econometric model:** MNL
- **Variable specification:** ML-enhanced
- **Synthetic sample generation algorithm:** ProWSyn

Therefore, we selected the MNL model to predict new mobility impacts. Features were

engineered from the 2008 data set according to the ML-enhanced specification, and were further augmented using ProWsyn-generated synthetic samples for minority classes. Next, a random sample was drawn from this augmented synthetic population while maintaining overall sample size and individual alternative market shares. All samples related to off-peak cars (OPC, i.e., Category 2) were removed from this random sample, as OPCs are treated as a proxy for a new mobility option in this application. The MNL model was then trained on this OPC-removed data set, following which we were able to obtain the estimated coefficients for the utility equations of the observed alternatives that we had termed as β_M in Figure 4-1.

The assumption of appropriate coefficients for the utility equation of the unobserved alternative (i.e., OPC, in this application) depends on the modeler. We assumed that these hypothesized coefficients (β_H) would reflect a reduced but similar effect to those for Category 3 (i.e., one normal car only). However, certain effects would require a reversal of direction, such as the residential location effect as off-peak cars cannot be used for daily commutes owing to their restricted usage. It is likely that living in a low-density suburb has a negative effect on the likelihood of owning an off-peak car, while a positive effect can be observed in the case of a normal car. Therefore, we halved the magnitudes of all estimated coefficients pertaining to Category 3, and reversed the directions for location effects, to obtain the hypothesized coefficients for the utility equation of OPC. Now equipped with both β_M and β_H, we were able to predict market shares for all six alternatives using the 2012 data set.

Since this procedure involves drawing a random sample, we repeated the procedure five time to account for stochasticity and sample biases. Five separate models were estimated on the randomly drawn samples after adjustment, which are reported as SA1, SA2, SA3, SA4, and SA5 in Table 4.11. The predicted market shares of these models are compared to a baseline MNL model, which was estimated on a sample without adjustment for low-sample alternatives. Our findings indicate a significant improvement in aggregate RMSE, to the tune of 60% on average. The low standard deviation across the five models indicate that the procedure is quite stable and is not affected by the stochastic nature of random sampling.

Moreover, upon examining the disaggregate market shares of individual alternatives, we find reasonable values being predicted for off-peak cars. This is in contrast to all previously examined models, where the OPC market share was abysmally low or even zero in some cases. Therefore, we successfully illustrate the application of our framework to predicting

Table 4.11: Predicted market shares for MNL with sample adjustment

Category	**Actual**	**w/o SA**	**SA1**	**SA2**	**SA3**	**SA4**	**SA5**	**Mean**[a]	**Std. Dev.**[a]
				Prediction Scenario [b]					
0	54.86	53.71	52.81	52.95	53.05	52.39	52.3	52.70	0.34
1	5.37	5.3	5.67	4.8	4.98	4.76	4.77	5.00	0.39
2	1.67	0	2.51	3.01	2.57	3.43	3.67	3.04	0.51
3	32.28	40.66	33.64	33.64	33.35	33.89	33.91	33.69	0.23
4	1.44	0.15	2.57	3.15	3.41	2.86	2.73	2.94	0.34
5	4.38	0.19	2.81	2.54	2.64	2.66	2.62	2.65	0.10
MAE	-	2.54	1.11	1.17	1.18	1.21	1.25	1.18	0.05
RMSE	-	3.37	1.22	1.35	1.37	1.41	1.46	1.36	0.09
ΔRMSE [c]	-	-	-64%	-60%	-59%	-58%	-57%	-60%	-

[a] The mean and standard deviation are calculated across the five MNL models with sample adjustment.

[b] Trained on HITS 2008 without OPC to predict HITS 2012 with OPC

[c] The percentage change in RMSE is calculated using a baseline without sample adjustment (w/o SA).

the impacts of new mobility in scenarios where the training data does not contain information about new mobility options. Along with achieving robust predictions, our framework improves upon the state-of-the-art practices currently being used by the choice modeling community. It is worth noting here that the degree of improvement depends on the assumptions leading to the creation of β_H, as the new alternative is not observed in existing data. Further fine-tuning of the assumptions for OPC availability decisions could lead to an even greater improvement, and is left as an avenue for future research.

4.6 Conclusion

We are witnessing a revolution in the transportation industry, which is evidenced by the wide variety of emerging technologies and services such as bike-sharing, ride-sharing, on-demand mobility, electric scooters, and autonomous vehicles. These new mobility options are likely to cause a transition from the traditional vehicle ownership paradigm to a renting and sharing economy on the lines of the Mobility-as-a-Service (MaaS) framework. Therefore, it becomes increasingly important to revisit traditional modeling frameworks to evaluate their appropriateness for forecasting adoption rates of these new mobility services.

While econometric frameworks are able to provide insights into the underlying behavioral decision-making behind observed choices, they suffer from low predictive capabilities. The emergence of big data and high-performance computing capabilities in recent times has greatly accelerated the adoption of machine learning models across a variety of domains. While ML models are much better for predictive purposes, they are notoriously difficult to interpret and provide shrouded effects that modelers are loath to accept as causal. Therefore, we propose a three-stage framework that seeks to use ML techniques to boost the predictive performance of state-of-the-art econometric models, while maintaining transparency and interpretability.

The first stage of our framework compares different modeling techniques based on three types of metrics — goodness-of-fit, predictive accuracy, and alternative-specific predicted market shares. Using an application of household vehicle availability in Singapore, we corroborate that econometric models are better for estimation purposes, while ML models have superior forecasting performance. The standard variable specification enables us to select the best-performing econometric and machine learning models. Selected variables are then transformed into economically sound representations to arrive at an econometric variable specification. We find that this greatly improves the performance of the multinomial logit (MNL) model, while significant changes could not be observed for ML models. Finally, we create the ML-enhanced variable specification, that utilizes model reliance to select the most valuable features and ALE plots to capture non-linearities in high-dimensional feature space. Accordingly, the non-linear regions are represented through piece-wise linear functions. This linear specification allows MNL to be almost at par with ML models in terms of predictive accuracy, while having a much lower RMSE for predicted market shares.

The second stage of our framework considers the possibility of using class-imbalanced data sets. We observe that our data set has skewed distributions of the market shares, and accounts for that by augmenting the data set with newly-generated synthetic samples. We test the performance of six different algorithms for synthetic sample generation and arrive at an optimal algorithm that has the best empirical performance in reducing RMSE. Building on our results from the previous stages of the framework, we select the optimal econometric model (MNL), variable specification (ML-enhanced), and synthetic sample generation algorithm (ProWSyn).

The third stage of our framework uses these optimal selections to predict market shares

for the off-peak car in Singapore, which is considered as a proxy for new mobility services on account of being a sustainable option as well as having a low value of the true market share. We ensure that the training data set does not contain any samples pertaining to the OPC, as new mobility examples are likely to be absent in a forecasting exercise. We find that our framework reduces the RMSE by 60% on average compared to the traditional MNL model, while maintaining the same level of interpretability. This is a significant improvement upon commonly used methods in practice. Future work can focus on testing other choice models beyond the multinomial and ordinal logit models considered in this study, although we expect our findings to remain consistent.

In summary, our model-agnostic framework can be applied to any scenario where choices for a low-sample alternative need to be predicted without having any samples for that alternative in the existing data set used for training. While this can be useful in a wide variety of applications, such a framework can prove to be particularly relevant for predicting market shares of new mobility services, as illustrated through the empirical example of the off-peak car in this study.

Chapter 5

Examining household dynamics in vehicle availability and use decisions

5.1 Introduction

With the emergence of new mobility services such as on-demand shared mobility and autonomous vehicles, the traditional paradigm of vehicle ownership is going to be put to the test. Several cities have already started transitioning to the Mobility-as-a-Service (MaaS) paradigm, where individuals can avail of a variety of mobility services including public transport and on-demand mobility for a fixed weekly or monthly charge using a single smart card. Anecdotes related to millennials being less prone to car ownership have become increasingly popular in the media. We believe that these changing times call for revisiting the traditional vehicle ownership paradigm.

This study seeks to improve upon existing literature in the following ways. First, we seek to shift the conversation away from ownership and rather focus on availability. This is motivated by the MaaS paradigm, where individuals do not own a private vehicle but can avail of multiple mobility services involving private vehicles. Second, most studies focus merely on car ownership as a categorical variable. While very few studies also consider motorcycle ownership, they are considered exogenous to the choice of car ownership. We address this issue by considering a mobility bundle instead of segregated ownership of different types of vehicles. The choice set of households is constructed in an ordinal manner, where the utility gained from a higher-ranked mobility bundle is greater owing to the increased options that

bundle provides for travel.

We are also aware of the possibility of self-reporting biases related to the position of the household head. To circumvent this as well as uncover the existence of gender biases in the household's decision-making process, we consider variables related to both the male and female highest income earners in the household. Finally, we observe that most discussions around vehicle usage are centered around aggregate measures (such as vehicle-miles traveled) that reflect the total usage of a vehicle by the household over a certain period of time. However, this approach does not lend insights into the intra-household interactions that occur during the decision-making process of who gets to use the car. With the inclusion of intra-household interactions, we propose a model that identifies the individual who is most likely to be the major car user in their household, should a car be available.

The Singaporean government has been particularly active with regard to mobility and transportation in the past couple of decades. Singapore's public transport network is one of the most extensive and reliable networks in the world, which results in a PT mode share of close to 65%. There are also efforts to provide at least one bus stop within 400 meters of every public housing project. Other innovative policies include the availability of the option to purchase an off-peak car. While normal cars are prohibitively expensive (due to high market prices and the requirement of a Certificate of Entitlement), the off-peak cars are heavily discounted and offer other benefits in the form of reduced road taxes and tolls. These benefits are provided in return for a restriction on the use of the off-peak car, as it cannot be used during weekday peak hours (7AM - 7PM). In light of the above discussion, we believe that Singapore is an apt environment (both in terms of policy efforts and future trends) to situate our empirical study in.

The remainder of this chapter is structured as follows. Relevant themes in the literature are reviewed and discussed in Section 5.2. Methodological details for the models considered in this study are provided in Section 5.3. The empirical results of our models and direct elasticities for selected variables of interest are discussed in Section 5.4. Finally, we provide concluding remarks and suggest areas for future research in Section 5.5.

5.2 Literature review

Long-term decisions like vehicle ownership and residential location choice are often interwoven with medium-term decisions like activity scheduling and commute mode. Moreover, these long-term decisions are undertaken at the household level while medium-term decisions including job location are undertaken at the individual level. It is worth bearing in mind that these decisions are not taken in isolation; rather, they feed off of each other and should be considered to have impacts across the household.

Most studies have attempted to conduct joint analysis of these decisions. Perhaps the first attempt was made in the seminal paper by Lerman (1976) who grouped residential location, housing type, auto ownership, and commuting mode into a bundle. He estimated the choice of this bundle using a multinomial logit (MNL) model, which unfortunately did not consider the correlations between different alternatives. This was rectified by Pinjari et al. (2011), who used a mixed logit (MXL) model to jointly consider residential location, auto ownership, bicycle ownership, and commute mode choices. Their model was able to successfully incorporate self-selection effects, endogeneity effects, correlated error terms, and unobserved heterogeneity. Paleti et al. (2013) extended the choice continuum to include more dimensions such as work location choice, commuting distance, and number of stops on commute tours. As the dimension of the choice set increased, the modeling methods grew increasingly complex and required large sample sizes for convergence during estimation.

This choice set explosion issue was addressed by reducing the dimensionality of the full choice set through the consideration of aggregate-level choices for each dimension. Guerra (2015) considered only four residential location categories (first, second, third, and fourth urban rings) and three car ownership categories (zero, one, and two or more cars). Similarly, Tran et al. (2016) considered only three categories (urban core, urban fringe, and suburb) for modeling location choices for both residences and jobs. Salon (2009) was able to disaggregate the choices further, but not by a significant margin. She considered three car ownership categories (zero, one, and two or more cars) and randomly selected ten census tracts to jointly model residential location choice, car ownership, and commute mode. In addition to some of the studies mentioned above, Ding et al. (2018) sought to jointly model car ownership and travel mode choice. However, they did not consider any location choices in their framework.

Some studies sought to delve deeper into the dynamics of vehicle ownership, but this attention to detail came at the cost of dropping other choices from consideration. Chiou et al. (2009) modeled car and motorcycle ownership, type, and usage in an integrated fashion for estimating energy consumption and emissions. Quite understandably, such disaggregation is necessary for an energy application because vehicle emissions are dependent on engine type and age. While they used a nested model structure, Liu et al. (2014) used a multinomial probit, a multinomial logit, and a regression to model vehicle ownership, type (class and vintage), and usage decisions respectively.

It should be noted that vehicle ownership and type are discrete choices, while vehicle usage is a continuous variable. To model such applications, Bhat (2005) proposed the Multiple Discrete–Continuous Extreme Value (MDCEV) model. Bhat and Sen (2006) presented an application of the MDCEV model through a simultaneous consideration of holding of multiple vehicle types (passenger car, SUV, pickup truck, minivan and van) and miles of usage of each vehicle type. This model was extended by Bhat et al. (2009) to include a nested structure that analyzed the choice of vehicle type and usage through a MDCEV component in the upper level, while the choice of vehicle model was modeled through a MNL component in the lower level. It is worth noting that these studies required merging of multiple data sources (including proprietary vehicle databases) as detailed vehicle data is usually not reported in traditional travel surveys, or conducting detailed surveys on their own.

While these studies improved upon statistical methods to estimate causal effects, they did not consider the temporal dimension in decision-making. Zhang et al. (2014) pointed out there are significant inter-dependencies between life domains, such as residential location and vehicle ownership, over the life course. Using a web-based life history survey to keep track of mobility biographies of 1,000 households in major Japanese cities, they suggested the necessity of developing biographical interdependence models that account for different time scales. Macfarlane et al. (2015) too found that consideration of past exposure to the built environment makes for a more nuanced understanding of current vehicle ownership patterns. The aforementioned findings from Japan and Atlanta were corroborated by a study in Beijing, which found that life-cycle events such as marriage or the birth of children have the strongest effect on car purchase and use (Zhao and Zhang, 2018).

5.2.1 Key takeaways

The key takeaways from our review of the literature can be summarized as follows:

- Studies have approached the joint consideration of long-term and medium-term decisions using two major methods. The first method uses complex integrated model structures that require large sample sizes for estimation of parameters. The second method addresses the choice set explosion issue by considering highly aggregate categories or randomly sampling a few disaggregate categories from the full choice set.

- None of the studies considering both residential location and vehicle ownership sought to use any monetary data. It should be clear to the reader that the joint choice of a mobility-housing bundle would be subject to the household's budget constraint, especially in an environment like Singapore where private cars are prohibitively expensive. This makes previous studies in this domain seemingly inadequate, as the effects captured are more aligned towards preferences rather than causation. The only study (to the best of our knowledge) that begins to acknowledge the budget constraint is by Manville (2017), who estimates the effect of bundled residential parking on household car ownership.

- Mobility biographies are an effective method to consider the temporal dimension in joint decision-making. Some studies find that inclusion of major life-cycle events and previous residential locations can improve our understanding of current vehicle ownership patterns. While most studies consider these decisions to be made jointly at a single point in time, the ground truth is that there is a direction of causality. Certain households choose their residential location keeping their job locations fixed, such as when families move to a new city. On the other hand, some households choose their job location keeping their residential location fixed, such as when individuals change their employers. Mobility biography data can help us examine the direction of these causal effects in a segregated manner and perhaps even identify latent classes of households.

- Most studies model vehicle use as a measure of the vehicle-miles traveled (VMT) of that vehicle. Since this is an aggregate measure of vehicle usage, there is a gap in the literature related to understanding which member in the household is most likely to use the vehicle. While a naïve method would be to assign the household head as the

major vehicle user, self-reporting of the household head role is often biased. This is particularly true in Asian contexts, where traditional gender roles can bias perceptions of an individual's importance to their household.

5.2.2 Research contributions of this study

Unfortunately, this study is restricted by the availability of data from only traditional travel surveys. Despite our best efforts in adopting the mobility biography approach, our survey is still actively collecting sample responses. Therefore, we will have to make do without this richly detailed data source at this point in time. Notwithstanding this setback, this study seeks to contribute to existing literature on two fronts.

First, we model household vehicle ownership through the consideration of a mobility bundle of multiple vehicles. This is an improvement over the studies discussed above, almost all of which consider three categories of car ownership (zero, one, and two or more). Ownership of multiple vehicle types is treated separately, which can lead to inefficient parameter estimates owing to correlations between choices of different vehicle types not being taken into account. Moreover, consideration of a mobility bundle allows us to isolate alternative-specific effects, helping us understand the driving factors behind the choice of each bundle. Another salient feature of our model is the consideration of two separate household heads (so to speak), i.e., the male and female highest income earners. This allows us to examine gender biases in the household's decision-making process.

Second, we attempt to identify which individual in the household is most likely to be the major user of the car, should the household have one available. This is an important aspect in planning daily travel decisions and scheduling activities. Intra-household interaction effects are taken into consideration through the inclusion of attributes related to the individual's job location and the relative location of any child's school location. These variables are interacted with gender to uncover gender roles in intra-household interactions, which might be missing or suffer from biased reporting in revealed preference surveys.

5.3 Methodology

We used standard econometric approaches to model household vehicle availability and identify the individual who is most likely to be the major user of the car, should the household

own one. The next sub-section entails details about the multinomial logit modeling approach for household vehicle availability. That is followed by a description of the binary logit modeling approach for identifying the major user of the car in the household. Finally, we provide details about the computation of direct elasticities, which allow us to examine the impact of changes in the covariates on the aggregate market shares of alternatives.

We developed the two econometric models in this study through an iterative process of adding different variables to the constants-only model and removing variables that turned out to be statistically insignificant. Moreover, we combined variables when we found that their effects on the model were not statistically different. In general, such a model construction exercise is guided by intuitive consideration on the part of the modeler, and parsimony in the representation of covariate effects.

It is worth noting that decision-making related to vehicle availability and use (like all other long-term household decisions) might involve multiple household members in the decision-making process. There are two ways to reflect the influence of such decision-making at the household level. One is to build a choice model with intra-household interactions, where the household utility function is defined as a function of each member's utility (Zhang et al., 2009). The other is to introduce some household-related attributes into an individual choice model. We adopt the latter approach in this study, by accounting for the influences of residential and work locations on observed choices by including relevant location attributes.

5.3.1 Vehicle availability

Most studies in the literature seek to model vehicle ownership as a categorical variable representing number of cars (or vehicles of a particular type), either in isolation or jointly with other decisions. We argue that considering vehicle ownership segregated by vehicle type is a flawed approach, which can be circumvented by considering a mobility bundle, i.e. joint ownership of multiple vehicles. Since the choice of a mobility bundle is a collective decision, it is likely to be jointly influenced by unobserved factors, which are captured by the error terms of the ownership equations. Ignoring the correlations among the error terms will produce statistically consistent but inefficient estimates.

While one approach to address this issue is to develop mixed-process models, such an approach is likely to suffer from insufficient sample size for certain alternatives. For example, Huang et al. (2017) developed three ordered probit models for auto, bike and e-bike owner-

ship with three categories (0, 1, and 2 or more) each, along with a binary probit model for motorcycle ownership (0, and 1 or more). However, this approach is likely to be inconsistent if automobiles dominate the mobility market, as it does in Singapore. Ownership of other mobility options like bikes, motorcycles and off-peak cars are significantly low in comparison. Therefore, we opt to construct our alternatives in an ordinal fashion as a mobility bundle that encompasses higher utility as we move up the mobility scale. These alternatives are described as follows:

- **CAT 0:** No private vehicles
- **CAT 1:** One or more motorcycles only
- **CAT 2:** One off-peak car w/wo motorcycles
- **CAT 3:** One normal car only
- **CAT 4:** One normal car with other vehicles
- **CAT 5:** Two or more normal cars w/wo other vehicles

We opted to use the multinomial logit (MNL) model here for two reasons. First, MNL was found to perform much better than the ordinal logit (OL) model for both calibration and prediction purposes in the previous chapter. Second, MNL allows for more detailed specification of utility equations that can be constructed in an alternative-specific fashion. For examples, car licenses are certainly expected to affect the choice of alternatives with cars, but perhaps not the alternative which considers only motorcycles (i.e., Category 1). This freedom is not available in the specification of an OL model. The methodological details of the MNL model have been discussed earlier in Section 4.4.1.

5.3.2 Vehicle use

While most studies model vehicle use as the number of vehicle-miles traveled (VMT) over an extended period of time (usually in the range of months to years), this requires more extensive data than that is available to us for the Singaporean context. Moreover, there is a gap in the literature about intra-household vehicle use decision-making, i.e. identifying the individual who is most likely to be the major user of the car. This is a particularly important

exercise, as it allows modelers to gain insights into the characteristics of individuals who drive the household decision-making process to purchase a vehicle.

We use the one-day travel diary from HITS 2012 to identify the individual who was the major user of the car based on total in-vehicle travel time (for households that owned a car). Other adults in the households that these major car users belong to were added to the data set, resulting in a final sample of 3,505 observations. We used a binary logit (BL) model for this decision, accounting for intra-household interaction effects. BL is an apt choice here as the outcome variable is binary, i.e., whether the individual is the major car user in their household or not. The consideration of intra-household interaction effects is particularly important, because failure to do so will result in inefficient parameter estimates as discussed earlier in Section 5.3.1.

5.3.3 Direct elasticities

Direct elasticities allow modelers to anticipate the impact of a change of the value of a certain covariate on the individual-level choice probability of an alternative being explained by that covariate, and subsequently on the market share of that alternative. The disaggregate direct elasticity of the model with respect to the qth variable associated by individual n to alternative i, (i.e., x_{inq}) is given by:

$$E_{x_{inq}}^{P_n(i)} = \frac{\partial P_n(i|x_n, C_n)}{\partial x_{inq}} \cdot \frac{x_{inq}}{P_n(i|x_n, C_n)} \tag{5.1}$$

Consequently, the aggregate direct elasticity of the model with respect to the average value x_{iq} is given by:

$$E_{x_{iq}}^{W_i} = \frac{\partial W_i}{\partial x_{iq}} \cdot \frac{x_{iq}}{W_i} \tag{5.2}$$

Note that the market share of alternative i in the population is:

$$W_i = \frac{1}{N_s} \sum_{n=1}^{N_s} w_n \cdot P_n(i|x_n, C_n) \tag{5.3}$$

where the sample size N_s is equal to the sum of the normalized sampling weights.

$$N_s = \sum_{n=1}^{N_s} w_n \tag{5.4}$$

Substituting Eq. (5.3) in Eq. (5.2), we obtain:

$$E_{x_{iq}}^{W_i} = \frac{1}{N_s} \sum_{n=1}^{N_s} w_n \cdot \frac{\partial P_n(i|x_n, C_n)}{\partial x_{iq}} \cdot \frac{x_{iq}}{W_i} \tag{5.5}$$

Assuming that the infinitesimal change of the variable is the same for every individual in the population, we obtain an expression for the aggregate elasticity.

$$E_{x_{iq}}^{W_i} = \sum_{n=1}^{N_s} E_{x_{inq}}^{P_n(i)} \cdot \frac{w_n \cdot P_n(i|x_n, C_n)}{\sum_{n=1}^{N_s} w_n \cdot P_n(i|x_n, C_n)} \tag{5.6}$$

Thus, we see that the aggregate elasticity is a weighted sum of disaggregate elasticities. However, the weight is not simply w_n as for the market share in Eq. (5.3), but it is a normalized version of $\big[w_n \cdot P_n(i|x_n, C_n)\big]$.

5.4 Results & Discussion

We report the estimation results of the MNL and BL models in this section, along with goodness-of-fit metrics that allow for evaluating the calibration performance of these models. Finally, we also discuss the implications of direct elasticity values of selected variables on the two decisions considered in this study.

5.4.1 Household mobility bundle choice

Since the MNL model specification allows for consideration of different variables to explain each alternative, we report the estimation results for each alternative in a separate table. The discussion of these results are provided below in separate sub-sections. It is worth noting that the no-vehicle alternative, i.e., Category 0, was considered as the base or reference alternative in this study. Moreover, to aid interpretation, we categorize covariate effects into three categories — *household demographics*, *housing*, and *individual jobs*.

One motorcycle only

Estimated parameters pertaining to the utility equation for this alternative are reported in Table 5.1. We find that motorcycle-owning households are primarily from minority ethnic backgrounds, as 80% of households in Singapore are of Chinese ethnicity. The marginal

effect of motorcycle licenses is quite strong, which aligns with our expectations. Owing to private cars being prohibitively expensive in Singapore, it is also expected that motorcycle-owning households are likely to be low-income. The direct income effect is corroborated by an indirect wealth effect through the negative coefficient associated with private housing.

Table 5.1: Estimation results of MNL model for HH mobility bundle (CAT 1) [a,b]

Variable Description	Coefficient	Robust std. error	p-value
Demographic effects			
Chinese ethnicity	-0.30	0.13	0.01***
Marginal effect of motorcycle licenses	4.70	0.15	0.00***
Marginal effect of per-capita HH income	-1.52	0.66	0.02**
Housing effects			
Private housing	-0.78	0.45	0.08*
Nearest bus stop is within 200 meters	-3.99	0.19	0.00***
Nearest bus stop is within 200 – 400 meters	-3.97	0.26	0.00***
Nearest MRT station is within 400 meters	-0.30	0.17	0.08*
Job location effects			
Male highest income earner: Nearest MRT station to job location is within 400 meters	-0.25	0.13	0.05**

[a] Estimated parameters are significant at 90% (*), 95% (**), and 99% (***) confidence levels.

[b] Only parameters corresponding to the utility of CAT 1 (one motorcycle only) are presented here.

Proximity to public transport results in very strong negative effects on the probability of motorcycle ownership, especially in the case of bus stops. This is consistent with the Singaporean landscape, as at least one bus stop is provided within 400 meters of every public housing project by government policy. Finally, we notice that only the male highest income earner has a statistically significant impact on the household's decision to purchase a motorcycle. This finding is consistent with our observation that an overwhelming majority of motorcycle users in Singapore are males.

One off-peak car w/wo motorcycles

We report the estimated parameters pertaining to the utility equation for this alternative in Table 5.2. While we know that off-peak cars enjoy a minuscule market share in the

mobility landscape, the decision to purchase an off-peak car seems to stem from a higher level of consciousness about the environment, rather than the mere availability of a lower-cost car. Households from minority ethnic backgrounds seem more favorable to this alternative. Moreover, households already owning bicycles are more likely to purchase off-peak cars.

Table 5.2: Estimation results of MNL model for HH mobility bundle (CAT 2) [a,b]

Variable Description	**Coefficient**	**Robust std. error**	**p-value**
Alternative specific constant	-7.76	0.87	0.00***
Demographic effects			
Chinese ethnicity	-1.22	0.17	0.00***
Percentage of adult males	-1.35	0.51	0.01***
Marginal effect of car licenses	4.10	0.44	0.00***
Bicycle ownership	0.41	0.18	0.02**
Housing effects			
Public housing in 3-room apartment	-0.76	0.25	0.00***
Land use diversity in residential location	1.70	1.00	0.09*
Nearest primary school is within 1 kilometer	0.38	0.16	0.02**
Job location effects			
Male highest income earner: Nearest bus stop to job location is within 200 meters	0.59	0.30	0.05**
Male highest income earner: Generalized travel impedance	1.65	0.37	0.00***
Female highest income earner: Nearest MRT station to job location is within 400 – 800 meters	0.61	0.19	0.00***

[a] Estimated parameters are significant at 90% (*), 95% (**), and 99% (***) confidence levels.

[b] Only parameters corresponding to the utility of CAT 2 (one off-peak car w/wo motorcycles) are presented here.

Unit type effects are mixed and inconclusive. However, residential location effects reinforce our hypothesis of these households being environment-friendly. They prefer to live in diverse neighborhoods, which have primary schools in close proximity to residences. Since the off-peak car cannot be operated during the day (restricted between 7AM - 7PM), the significantly positive effect of generalized travel impedance for the male highest income earner implies that these individuals are more likely to be working night-shift jobs. We also find

that women have an impact on this household decision, albeit with a smaller magnitude than males.

One normal car only

Estimated parameters pertaining to the utility equation for this alternative are reported in Table 5.3. We notice varied demographic effects guiding the choice of a normal car. Normal car-owning households are more likely to be of Chinese ethnicity, unlike those owning motorcycles and off-peak cars. A male-dominated household (in terms of number of male members) is less likely to own a car, perhaps because cars are primarily purchased by households having individuals with varied mobility needs such as children, teenagers and students.

Having young children in the household have a positive impact on this choice, which is moderated but still remains positive when we consider teenagers. The effect of students in the household is also quite strong. Our hypothesis of diverse mobility needs driving this decision is confirmed by these findings in addition to the negative effect of workers. A high proportion of workers in the household decreases the choice probability of this alternative, as does taxi ownership. Finally, we observe a very strong impact of household income on this decision. In summary, high-income households with children are more likely to purchase normal cars.

We considered different unit types for public housing to differentiate their impacts on this decision. We find a progressively decreasing negative impact, implying that wealthier households who live in more spacious public housing apartments or private housing are more likely to buy normal cars. These households also live in neighborhoods with low residential density, thereby pointing towards the stereotype of the high-income couple living in the suburbs with a kid and a car. This is corroborated by the negative effect of proximity to MRT stations, implying that households favoring the normal car do not have good accessibility to MRT stations from their residential locations, which can be frequently observed in the suburbs.

Finally, we observe that the job location of the male highest income earner in the household has an impact on this decision. The generalized travel impedance from the residential location to this individual's job location has a positive effect, which is tempered by access to bus stops from their job location. The absence of variables related to the female highest

Table 5.3: Estimation results of MNL model for HH mobility bundle (CAT 3) [a,b]

Variable Description	Coefficient	Robust std. error	p-value
Alternative specific constant	-3.03	0.27	0.00***
Demographic effects			
Chinese ethnicity	0.49	0.08	0.00***
Marginal effect of children	0.45	0.06	0.00***
Marginal effect of teens	0.17	0.08	0.03**
Percentage of adult males	-0.79	0.18	0.00***
Percentage of students	0.36	0.18	0.05**
Percentage of workers	-0.89	0.15	0.00***
Percentage of white-collar workers	0.22	0.12	0.07*
Marginal effect of car licenses	3.44	0.11	0.00***
Taxi ownership	-1.40	0.23	0.00***
Marginal effect of per-capita HH income	4.16	0.47	0.00***
Housing effects			
Public housing in 1-room apartment	-1.32	0.45	0.00***
Public housing in 2-room apartment	-1.41	0.32	0.00***
Public housing in 3-room apartment	-1.26	0.13	0.00***
Public housing in 4-room apartment	-0.72	0.11	0.00***
Public housing in 5-room apartment	-0.26	0.11	0.01***
Residential density in residential location	-0.31	0.13	0.02**
Nearest MRT station is within 400 meters	-0.18	0.08	0.03**
Nearest MRT station is within 400 – 800 meters	-0.11	0.07	0.09*
Job location effects			
Male highest income earner: Nearest bus stop to job location is within 200 meters	-0.35	0.18	0.06*
Male highest income earner: Nearest bus stop to job location is within 200 – 400 meters	-0.36	0.20	0.07*
Male highest income earner: Generalized travel impedance	0.45	0.15	0.00***

[a] Estimated parameters are significant at 90% (*), 95% (**), and 99% (***) confidence levels.

[b] Only parameters corresponding to the utility of CAT 3 (one normal car only) are presented here.

income earner indicates a gender gap in the decision-making power within the household.

One normal car with other vehicles

We report the estimated parameters pertaining to the utility equation for this alternative in Table 5.4. The effect of children and workers noticed here are quite similar to those evident for the previous alternative. An interesting addition here is the positive effect of seniors but the negative effect of the percentage of retired individuals in the households. This implies that the normal car might be used by the seniors in multi-generational families, while the younger individuals use other privately owned vehicles (such as off-peak cars and motorcycles) or public transit. Retiree-dominant households are likely to have a different family structure, where the purchase of motorcycles and off-peak cars may not be useful.

We also notice an indirect wealth effect, illustrated through negative coefficients for public housing apartments. Strong negative effects for proximity to bus stops from the residential location are clearly evident. Quite similar to households preferring only one normal car, these households are likely to live in the suburbs or neighborhoods with low residential density.

Proximity to bus stops from the job location of the male highest income earner has a strong negative effect on the probability of this choice. While the effect for proximity to MRT stations is also negative, the magnitude is not as strong. Finally, we notice a positive effect for the generalized travel impedance of the female highest income earner. It is worth noting that this is in contrast to the alternative with only one normal car, where the female highest income earner did not have any impact on the decision. However, availability of multiple vehicles in the household might imply that the male highest income earner is more likely to use public transport or motorcycles (if available) to commute to work, while the female highest income earner is more likely to use the normal car.

Two normal cars w/wo other vehicles

Estimated parameters pertaining to the utility equation for this alternative are reported in Table 5.5. We see that households owning two normal cars are more likely to be Chinese. Moreover, positive effects due to children and seniors are noticed. White-collar workers also have a positive impact on this decision, along with a very strong effect evident for household income. Meanwhile, the probability of owning two normal cars is seen to decrease as the proportion of retired individuals and adult males increase. By considering all these effects

Table 5.4: Estimation results of MNL model for HH mobility bundle (CAT 4) [a,b]

Variable Description	Coefficient	Robust std. error	p-value
Demographic effects			
Marginal effect of children	0.46	0.22	0.03**
Marginal effect of seniors	0.49	0.26	0.06*
Percentage of retired individuals	-2.16	0.79	0.01***
Percentage of workers	-1.47	0.49	0.00***
Marginal effect of motorcycle licenses	3.77	0.28	0.00***
Marginal effect of car licenses	4.26	0.50	0.00***
Marginal effect of per-capita HH income	0.81	1.03	0.43
Housing effects			
Public housing in 3-room apartment	-1.00	0.39	0.01***
Public housing in 4-room apartment	-0.62	0.25	0.01***
Nearest bus stop is within 200 meters	-5.75	0.76	0.00***
Nearest bus stop is within 200 – 400 meters	-5.95	0.78	0.00***
Residential density in residential location	-1.30	0.43	0.00***
Job location effects			
Male highest income earner: Nearest bus stop to job location is within 200 meters	-1.65	0.39	0.00***
Male highest income earner: Nearest bus stop to job location is within 200 – 400 meters	-2.28	0.54	0.00***
Male highest income earner: Nearest MRT station is within 400 meters	-0.58	0.26	0.02**
Female highest income earner: Generalized travel impedance	0.90	0.53	0.09*

[a] Estimated parameters are significant at 90% (*), 95% (**), and 99% (***) confidence levels.

[b] Only parameters corresponding to the utility of CAT 4 (one normal car with other vehicles) are presented here.

in unison, we can conclude that high-income families with children, and multi-generational families with young professionals and white-collar workers are highly likely to purchase two normal cars.

We notice very strong negative effects for public housing apartments. Similar effects are noticed for proximity to bus stops and MRT stations, albeit the effect is more muted in

magnitude for the latter. Finally, residential density also has a fairly strong negative effect. These housing effects lead us to conclude that these households live in private housing located in low-density neighborhoods with relatively poor access to public transit. Our findings are quite similar to those for alternatives involving a normal car, i.e., categories 3 and 4.

Both the male and female highest income earners have a say in this decision, as illustrated by the presence of statistically significant variables pertaining to their job locations in the final model. It is interesting to observe that only variables considering the lowest buffer radii for calculating proximity to transit are significant. This implies that the households that are more likely to purchase two normal cars might only reconsider their decision if their household heads' job locations are extremely close to transit stops, all else held equal. Even relatively good proximity, i.e., at the second buffer radii (800 meters) level, is not sufficient to make an impact on their decision.

Model goodness-of-fit

Goodness-of-fit metrics for the MNL model are reported in Table 5.6. 78 parameters were estimated in this model, indicating the complexity of considering a mobility bundle as opposed to vehicle ownership for individual vehicle types. While the freedom to include selected covariates in alternative-specific utility equations enables better understanding of the underlying behavioral mechanisms, it also tends to increase model complexity. However, the inclusion of a relatively large number of parameters (compared to the literature) does not seem to have adversely affected the model's explanatory power as these parameters included important information for the model. This is evidenced by McFadden's pseudo R-squared (ρ^2) being quite high, along with the adjusted measure $\bar{\rho}^2$ still remaining high even when the number of parameters is taken into consideration. It is worth mentioning here that McFadden (1977) reported an "excellent model fit" being indicated by the value of this measure being in the range of 0.2 to 0.4. Using this statement as a yardstick, we can conclude that our MNL model for the household choice of a mobility bundle has an excellent fit.

Direct elasticities

Direct elasticities for a few selected variables on the choice probability of mobility bundle alternatives are reported in Table 5.7. The effect of children on mobility bundles containing one normal car is quite evident, with an increase in magnitude observed as we move up

Table 5.5: Estimation results of MNL model for HH mobility bundle (CAT 5) [a,b]

Variable Description	Coefficient	Robust std. error	p-value
Alternative specific constant	-7.24	1.05	0.00***
Demographic effects			
Chinese ethnicity	0.99	0.20	0.00***
Marginal effect of children	0.59	0.14	0.00***
Marginal effect of seniors	0.77	0.15	0.00***
Percentage of adult males	-1.34	0.43	0.00***
Percentage of retired individuals	-2.18	0.47	0.00***
Percentage of workers	-1.97	0.36	0.00***
Percentage of white-collar workers	0.30	0.27	0.27
Percentage of young professionals	0.92	0.24	0.00***
Marginal effect of car licenses	8.72	0.56	0.00***
Marginal effect of per-capita HH income	5.03	0.62	0.00***
Housing effects			
Public housing in 3-room apartment	-3.08	0.42	0.00***
Public housing in 4-room apartment	-2.18	0.22	0.00***
Public housing in 5-room apartment	-1.06	0.19	0.00***
Nearest bus station is within 200 meters	-3.00	0.87	0.00***
Nearest bus station is within 200 – 400 meters	-2.68	0.88	0.00***
Nearest MRT station is within 400 meters	-0.34	0.20	0.08*
Nearest MRT station is within 400 – 800 meters	-0.25	0.15	0.09*
Residential density in residential location	-0.82	0.25	0.00***
Job location effects			
Male highest income earner: Nearest bus stop to job location is within 200 meters	-0.39	0.19	0.04**
Female highest income earner: Nearest MRT station to job location is within 400 meters	-0.24	0.14	0.09*

[a] Estimated parameters are significant at 90% (*), 95% (**), and 99% (***) confidence levels.

[b] Only parameters corresponding to the utility of CAT 5 (two normal cars w/wo other vehicles) are presented here.

the mobility scale. White-collar workers also have a positive effect on normal car-inclusive alternatives, but with a much smaller magnitude. It is interesting to note that their elasticity

Table 5.6: Summary of MNL model for HH mobility bundle

Metric	Value
No. of observations (N)	9,222
No. of estimated parameters (K)	78
Likelihood of constants-only model $\left[\mathcal{L}(\beta_0)\right]$	-16,523.61
Likelihood of full model $\left[\mathcal{L}(\hat{\beta})\right]$	-5,933.17
Likelihood ratio $\left(-2[\mathcal{L}(\beta_0) - \mathcal{L}(\hat{\beta})]\right)$	21,180.87
McFadden's pseudo R-squared (ρ^2)	0.641
McFadden's adjusted pseudo R-squared ($\bar{\rho}^2$)	0.636

for Category 4, i.e., one normal car with other vehicles, is zero. This may be because white-collar workers are unlikely to purchase motorcycles, while off-peak cars would not be useful for them to commute to work. Therefore, white-collar workers would be more inclined to prefer one normal car only (Category 3) or two normal cars w/wo other vehicles (Category 5).

We notice strong income effects on the probability of purchasing one or two normal cars. However, that is not the case for the alternative pertaining to one normal car with other vehicles. This may be because the purchase of this mobility bundle is necessitated by other variables such as family structure and diverse mobility needs. Land use diversity in the residential location has a very strong effect on the choice of the relatively sustainable bundle containing an off-peak car. If the land use diversity is doubled, i.e., increased by 100%, then the probability of a household purchasing an off-peak car is likely to increase by 85%. This is certainly a strong incentive for the government to encourage development of mixed-use neighborhoods. Similar policy implications surface from the elasticities of residential density. Doubling the residential density might result in a reduction of the probability of purchasing multiple vehicles (including at least one normal car) by 76% to 122%. It appears that improved public transport and walking accessibility, through mixed use and high density, makes households less reliant on cars and increases the likelihood of use of alternative mobility options. Moreover, such scenarios incentivize the limited use of a car, especially because households perceive the utility from an off-peak car to be worth the cost savings. Therefore, the policy recommendation from this analysis is to encourage high-density mixed-use development in Singapore.

Table 5.7: Direct elasticities for HH mobility bundle

Variable	Elasticity
Marginal effect of children in HH	
CAT 3: One normal car only	0.3065
CAT 4: One normal car with other vehicles	0.4902
CAT 5: Two normal cars w/wo other vehicles	0.6109
Percentage of white-collar workers in HH	
CAT 3: One normal car only	0.0729
CAT 5: Two normal cars w/wo other vehicles	0.1515
Marginal effect of per-capita HH income	
CAT 1: One motorcycle only	-0.2677
CAT 3: One normal car only	0.4523
CAT 4: One normal car with other vehicles	0.1461
CAT 5: Two normal cars w/wo other vehicles	0.8620
Land use diversity in HH's residential location	
CAT 2: One off-peak car w/wo motorcycles	0.8538
Residential density in HH's residential location	
CAT 3: One normal car only	-0.2129
CAT 4: One normal car with other vehicles	-1.2233
CAT 5: Two normal cars w/wo other vehicles	-0.7608

5.4.2 Major car user in household

The parameters estimated for the binary logit model to identify the major car user in the household are reported in Table 5.8. Along with individual demographic effects and the impact of their job location, we also examine intra-household interaction effects to account for the fact that vehicle use decisions are made jointly by multiple members of the household. We find that the strongest effect towards an individual being a major car user is that of gender, wherein males are highly likely to become major car users in their household. When we examine the effect of age, we find that mid-aged and mature adults are more likely to be major car users compared to young adults and seniors. This may be because they are more likely to be employed full-time, and probably bring in the lion's share of the household income. This effect is corroborated when the highest earning member of the household is considered. Conversely, students are highly unlikely to become major users of the household

car due to this very reason.

Table 5.8: Estimation results of BL model for major car user in HH [a]

Variable Description	Coeff.	Robust std. error	p-value
Alternative Specific Constant (ASC)	-1.85	0.24	0.00***
Demographic effects			
Male	2.09	0.23	0.00***
Student	-2.36	0.28	0.00***
Blue-collar worker	0.19	0.14	0.18
Young adult (*20 - 34 years old*)	0.35	0.21	0.10*
Mid-aged adult (*35 – 44 years old*)	0.83	0.22	0.00***
Mature adult (*45 – 59 years old*)	0.79	0.22	0.00***
Highest earning member of HH	1.17	0.11	0.00***
Job location effects			
Generalized travel impedance	1.40	0.46	0.00***
Employment density at job location	-0.58	0.15	0.00***
Intra-household interaction effects			
Car travel time between residential and job locations (*interacted with Male*)	1.09	0.55	0.05**
Car travel time between residential and child's school locations (*interacted with Male & Presence of children*)	-1.54	0.48	0.00***
Car travel time between job and child's school locations (*interacted with Presence of children*)	-0.27	0.22	0.23

[a] Estimated parameters are significant at 90% (*), 95% (**), and 99% (***) confidence levels.

Job location effects for adults with jobs are found to be significant. The generalized travel impedance has a strong effect, implying that the individual who faces the greatest inconvenience in commuting to work is most likely to get the car. Additionally, employment density at the job location has a negative effect. This finding aligns with our intuition that locations with high employment density, such as the CBD and other employment zones, are likely to be better connected by public transit, thereby reducing the likelihood of the individual working in such a location to use the car.

Finally, significant intra-household interaction effects are noticeable, thereby validating the inclusion of these additional variables in lieu of a more complex model structure. We

interact gender with some of the effects discussed earlier in order to uncover gender biases in household decision-making. In a scenario where two individuals of different genders have high commuting times to their jobs, the male individual is significantly more likely to get the car. However, if the household has children and their school location is close to the residence, it is more likely for a female member to get the car. This finding is reminiscent of the soccer mom in the United States, where the matriarch of the household uses the car to satisfy the mobility needs of the children, while the patriarch uses public transport to commute. If the child's school is located close to a particular individual's job location, they are more likely to get the car so that they can drop-off and/or pick-up the child from school. However, this effect is statistically insignificant as such cases are rare in Singapore owing to the prevalence of families with children living near the schools of their children.

Model goodness-of-fit

Goodness-of-fit metrics for the BL model are reported in Table 5.9. Compared to the large number of parameters estimated earlier for the MNL model, this model has only 13 parameters. Despite being a relatively simpler model, we still observe an excellent model fit, as illustrated by McFadden's adjusted pseudo R-squared ($\bar{\rho}^2$) being well above the suggested range of 0.2 - 0.4. Moreover, the insignificant drop in the measure while accounting for the number of parameters indicates that these parameters hold valuable information relevant to the model.

Table 5.9: Summary of BL model for major car user in HH

Metric	**Value**
No. of observations (N)	3,505
No. of estimated parameters (K)	13
Likelihood of constant-only model $\left[\mathcal{L}(\beta_0)\ \right]$	-2,429.481
Likelihood of full model $\left[\mathcal{L}(\hat{\beta})\ \right]$	-1,211.754
Likelihood ratio $\left(-2[\mathcal{L}(\beta_0) - \mathcal{L}(\hat{\beta})]\right)$	2,435.454
McFadden's pseudo R-squared (ρ^2)	0.501
McFadden's adjusted pseudo R-squared ($\bar{\rho}^2$)	0.496

Direct elasticities

We report the direct elasticities of selected variables for the BL model that identifies the individual who is most likely to be the major car user for households owning at least one normal car in Table 5.10. The impacts of job location effects on the likelihood of being a major car user are moderate, implying that the demographic effects tend to dominate this decision. Doubling the generalized travel impedance between an individual's residential and job locations increases their probability of getting the car by only about 13%. The intra-household interaction elasticities are even lower in magnitude. Therefore, we can conclude that the most dominant factors guiding the allocation of the major car user in the household are the gender of the individual, and their proportional share of the total household income.

Table 5.10: Direct elasticities for major car user

Variable	**Elasticity**
Job location effects	
Generalized travel impedance between residential and job locations	0.1270
Employment density at job location	-0.1177
Intra-household interaction effects	
Car travel time between residential and job locations (*interacted with Male*)	0.0506
Car travel time between residential and child's school locations (*interacted with Male & Presence of children*)	-0.0533
Car travel time between job and child's school locations (*interacted with Presence of children*)	-0.0428

5.5 Conclusion

Emerging transportation technologies that focus on sharing and on-demand services necessitate modelers to revisit the traditional vehicle ownership paradigm. One approach to prepare for the shared mobility paradigm is to focus our attention on vehicle availability rather than ownership. Moreover, it would be more useful to focus on a bundle of different mobility services instead of considering ownership of different types of vehicles separately. Combining these approaches would enable the analysis to become quite similar to examining

Mobility-as-a-Service (Maas), which has already started to become prevalent in several cities across the globe.

The first part of this study models vehicle availability as an ordinal construction of mobility bundles. We use travel survey data from Singapore to create mobility bundles that differentiate between motorcycles, off-peak cars, and normal cars. The multinomial logit (MNL) model is employed to model the underlying behavioral mechanisms driving a household's choice of a particular mobility bundle. While it is likely that the household head might have more decision-making power for such decisions, self-reporting of the household head role can be flawed due to biased perceptions of traditional gender roles. We seek to circumvent this by considering attributes of both the male and female highest income earners.

Our findings indicate that the male highest income earner dominates decision-making for motorcycles and normal cars. However, when it comes to purchasing a normal car with other vehicles (such as motorcycles and off-peak cars), the female highest income earner has a statistically significant effect on the decision. High-income households with complex and diverse mobility needs, which arise with the presence of children and seniors in the household, are more likely to prefer mobility bundles involving one normal car. Moreover, these households are located in the suburbs where residences do not enjoy good access to public transport. Households preferring the off-peak car bundle, however, seem to be more environment-friendly in their outlook. Along with owning bikes, they also tend to live in more diverse neighborhoods with mixed land uses. Our examination of direct elasticities motivates us to infer that creating more high-density mixed-use neighborhoods might induce significant decreases in household preferences for private cars. This is an important policy recommendation for the Singaporean government to follow up on.

The decision to purchase a car is often influenced by the variety of mobility needs that car is perceived to fulfill for the household. Therefore, in the second part of this study, we construct a model to identify the individual who is most likely to become the major user of the car in their household. Our model considers several intra-household interaction effects, and finds them to be statistically significant. We find that the gender effect is the most dominant, as a male is more likely to become the major car user (all else held equal). The effect is similarly strong if the individual is the highest income earner of the household.

Job location effects are also found to play a role in this decision, as the individual who

faces the greatest inconvenience in commuting to work has a higher likelihood of getting the car. However, when we interact this effect with gender, we find that males tend to skew the decision yet again. For households with children, a female member (most likely the matriarch) is likely to get the car if the child's school location is close to home. In the rare cases of the school being located close to an individual's work, they might have the upper hand in getting to use the car. In summary, we find that the use of cars in Singapore is strongly mediated by intra-household effects. Since private cars are prohibitively expensive, they are often used for multi-purpose tours that involve trip-chaining.

A key limitation of this study (and other studies in the literature) is the failure to consider wealth. While we do consider direct effects of income, we can only capture the wealth effect in an indirect manner through the consider of housing type. While we should bear in mind that wealth is often not reported or under-reported in surveys, it is possible to arrive at proxy measures through the use of neighborhood-level attributes such as median housing price. It is for this very reason that we choose not to model vehicle availability jointly with other long-term decisions in this study, as is observed in the literature. Without the consideration of the household budget, the joint choice of a housing-mobility bundle is likely to be flawed as the choice mechanism will lean towards observed preferences rather than causal effects.

We also choose not to consider job location choice in our framework, as most individuals are unlikely to consider only the location of the job but not job type while searching for employment opportunities. We believe that joint decision-making models in the literature treat the job-related dimension inadequately. It is likely that there are latent classes in the population that have different directions of causality, i.e., some households hold residential location constant while searching for jobs, and vice-versa.

An avenue for future research that we are currently exploring is the use of mobility biographies. By observing long-term choices of households over an extended period of time, mobility biographies allow us to incorporate the temporal dimension into our modeling framework. We are currently in the midst of collecting data from 2016-2018 in Singapore through a mobility biography survey that seeks to address many of the limitations outlined above. The data will provide information about the direction of causality in addition to more detailed vehicle information such as make/model, cost and usage. We hope to augment this data with hedonic price models calibrated on housing unit transaction data, which can

provide market rates for residential units and enable us to incorporate household budget constraints.

In summary, this chapter discussed a useful study to examine household choice behavior in Singapore related to vehicle availability and use. We were able to identify significant household-interaction effects, which reinforce the necessity of using mobility biographies to track life-cycle events and vehicle transactions of households over an extended period of time.

Chapter 6

Evaluating the impact of car-lite policies on housing-mobility choices

6.1 Introduction

Emerging transportation technologies like on-demand shared mobility and autonomous vehicles (AVs) are motivating discussions on the future of cities. AVs offer several benefits to the transportation system such as reduced day parking locations and per-kilometer commute costs in addition to enhanced consumer experience through lower prices and higher comfort of traveling. While examining city-level impacts, Zakharenko (2016) found that increased AV availability can increase worker welfare, travel distances and city size. Meyer et al. (2017) corroborated this by stating that AVs could cause a quantum leap in accessibilities. However, the net effects of AVs are uncertain, and recent investments by car manufacturers (such as Ford and General Motors) and transportation network companies (like Uber and Lyft) in the AV market have attracted attention from policy-makers with regard to regulation strategies and equity considerations.

While the benefits of AVs are heavily advertised in media outlets, experts hold more nuanced opinions of the impacts of AVs on cities. Using a survey of domain experts, Milakis et al. (2018) reported that there are three major viewpoints about AVs.

- **Uncertain benefits:** The accessibility benefits due to AVs will be highly uncertain, because the induced travel demand will cancel out the travel time and cost savings in the long-term.

- **Changing urban form:** AVs will cause the city center to become denser while the peripheries will expand leading to urban sprawl.
- **Only for the rich:** The benefits of AVs will be enjoyed only by those who can afford them, thereby leading to negative implications for social equity.

A recent review of modeling studies related to the impacts of AVs by Soteropoulos et al. (2019) found that the literature can be broadly segmented into two categories: (a) *impacts on travel behavior*, such as trip generation rates, mode choice, and vehicle-kilometers traveled (VKT); and (b) *impacts on land use*, such as location choices and reduction of parking spaces. Long-term effects of AVs on transport-land use interactions are undeniably very complex and are co-determined by exogenous factors like housing supply, area attractiveness and land use policy. Their review indicates a gap in the literature with regard to an integrated treatment of long-term impacts of AVs on urban regions.

Integrated land use and transportation (ILUT) models can be critical tools in addressing this research gap. Hawkins and Nurul Habib (2019) review several ILUT models from the literature, and highlight both their capabilities and shortcomings in undertaking such an exercise. They mention that existing model applications have focused on the discrete addition of new infrastructure or policy at a fixed point in time. However, we posit that AV adoption will occur incrementally through time as opposed to one fell swoop of private vehicles being replaced by AVs, as most literature would have us believe. Therefore, we consider the case of an AV-related policy being implemented as a pilot, and examine its impacts using our ILUT modeling platform, *SimMobility* (Adnan et al., 2016).

The remainder of this chapter is structured as follows. Relevant themes in the literature are reviewed and discussed in Section 6.2. We detail the elements of our simulation platform in Section 6.3. Methodological details for this study are provided in Section 6.4. The empirical results of the simulations are discussed in Section 6.5. Finally, we provide concluding remarks and suggest areas for future research in Section 6.6.

6.2 Literature Review

The two long-term impacts of of AVs that are of primary interest to policy-makers are related to (a) *private vehicle ownership*, and (b) *residential relocation*. Therefore, we examine recent literature on these fronts in the following sub-sections.

6.2.1 Impact of AVs on private vehicle ownership

Three major threads of inquiry emerge from the literature. *The first technique assumes complete replacement of private vehicles by AVs.* Two types of methodologies are commonly used in such studies: (a) *agent-based models*, and (b) *activity-based models*. Different AV fleet sizes are explored with marginal constraints implemented for matching trip generation rates and total travel demand, while VKT (as a proxy for emissions) and total travel times (as a proxy for congestion) are considered as efficiency indicators. Using an agent-based microsimulator *MATSim*, Hörl et al. (2016) found that private ownership would not be attractive in a future dominated by 24/7 on-demand mobility. Through the analysis of travel surveys and trip profiles from the Atlanta Metropolitan Area, Zhang et al. (2018) found that current travel patterns can be maintained with a 9.5% reduction in private vehicles in the study region. While more empirical applications can be found in Soteropoulos et al. (2019), there is a general consensus that the exogenous assumptions (about time or cost savings in particular) and the complete replacement assumption (which is quite unrealistic) limit the use of such studies for policy-making.

The second technique uses stated preference surveys to elicit user preferences regarding AVs. Different business models (shared or private AVs), cost structures, travel experience metrics (travel time or cost savings for AVs), and market penetration rates are explored to construct the choices in the stated preference interviews. Menon et al. (2019) examined the impact of shared AVs on the likelihood of relinquishing private vehicles using a sample of university personnel and American Automobile Association (AAA) members. One of their findings highlighted that households are more likely to give up a private vehicle if they own multiple cars, compared to their only car. An online survey conducted in Germany by Pakusch et al. (2018) compared current and future travel modes in a pairwise fashion. They too found that private cars would remain the preferred travel mode, thereby challenging the popular hypothesis that the ownership model will become outdated due to availability of AVs. Haboucha et al. (2017) used a stated preference questionnaire to examine user preferences for currently owned private cars, shared AVs, and privately-owned AVs. They uncovered significant socio-demographic and cultural differences in their combined sample of Israeli and North American residents. Their most substantial finding points to the reluctance in AV adoption, as only 75% of their sample were willing to use shared AVs even if the service

was made completely free.

The third technique employs a cost-based approach by comparing the costs of ownership over the long-term. Bösch et al. (2018) used a cost-benefit analysis to conclude that private vehicle ownership is the cheaper alternative to shared AVs. A similar study in the UK among different income groups reported that high-income households would benefit more from AVs due to their higher perceived value of time (Wadud, 2017). Jiang et al. (2018) utilized a cost-based approach to construct an online stated preference survey for car users in Japan. They found that respondents in their sample were willing to pay an additional JPY 402, 233 – 793,611 (USD 3,632 – 7,166) to purchase an AV in the future. This has serious implications for social equity, as most households will not be able to afford AVs. However, shared AV fleets offering mobility-on-demand (MoD) services can be more affordable than privately owned vehicles. A financial analysis of AMoD services in Singapore found that shared autonomous vehicles can save around 15,100 USD/year in total mobility costs compared to a privately owned human-driven car (Spieser et al., 2014).

The key takeaway from our review of existing literature is the challenge associated with influencing people to relinquish currently owned private vehicles. This is rooted in behavioral theory as the *endowment effect*, which posits that people value ownership of certain commodities more than their willingness-to-pay (WTP) for them (Kahneman et al., 1990). This implies that a shift to AVs will require a boost in utility greater than that expected solely based on the attributes of the alternative modes.

6.2.2 Impact of AVs on residential relocation

Most studies concerning impacts of AVs on location choices of people show an increase in accessibility, and an expansion of population size in well-connected outer suburbs and rural regions (Soteropoulos et al., 2019). However, it is worth bearing in mind that these findings of urban sprawl are associated with assumptions of reductions in the value of time and travel time along with an increase in the roadway capacity. Assuming a reduction of value of time by 50% for private AVs, Thakur et al. (2016) modeled travel behavior and residential location choices for Melbourne in 2046. Their findings indicate an out-migration rate of 3% from the inner city to the suburbs. However, their results are inconclusive due to mixed effects when shared AVs are considered. Zhang and Guhathakurta (2018) used an agent-based simulation approach to model changes in residential location choice in a scenario where shared AVs are

considered a popular travel mode in the Atlanta Metropolitan Area. Their results indicated that older people moved closer to the inner-city core while younger people moved out to the suburbs. Additionally, Meyer et al. (2017) found that shared AVs could curb urban sprawl.

6.2.3 Research contributions of this study

We outline our research contributions based on the gaps that emerged from our literature review. First, almost all ILUT models fail to merge agent-based simulation dynamics with behavioral economics in at least one aspect of the land use-transport interactions. We use our ILUT model *SimMobility* that uses embedded econometric models in an agent-based microsimulation framework to simulate both long-term and medium-term decisions.

Second, existing literature treats private vehicle ownership and residential relocation as separate decisions while examining the impact of AVs. While some literature exists on jointly considering these decisions, those studies deal with empirical modeling in a past or current setting without involvement of AVs (Zhang et al., 2014). We address this issue by considering them jointly in a sequential simulation framework.

Third, instead of considering complete replacement of private vehicles by AVs, we maintain a more realistic outlook where the planning agency introduces a car-lite policy in a particular neighborhood as a pilot. This policy enhances accessibility for residents by introducing AVs and AMoD services inside the study region. We use our sequential simulation framework to examine the impacts of this policy on household vehicle availability decisions. Different market reactions to the policy are modeled as different scenarios and compared to a baseline where the policy is never implemented. This allows for a more robust examination of private vehicle substitution patterns due to availability of AVs.

6.3 Framework

We provide methodological details about our microsimulation platform *SimMobility* and our proposed framework for policy analysis in this section.

6.3.1 SimMobility: An overview

SimMobility is a multi-scale agent-based microsimulation platform that incorporates time-scale dependent behavioral modeling through activity-based frameworks (Adnan et al.,

2016). Through the consideration of interactions between transportation and land use, *SimMobility* can be used for a variety of applications ranging from implementation of intelligent transportation systems to evaluation of alternative future scenarios. *SimMobility* encompasses three major components:

- **Long-Term (LT):** This detailed land use-transport simulator involves the creation of a synthetic population of individuals and households (Zhu and Ferreira Jr, 2014), and firms and establishments (Le et al., 2016). This is followed by household-level residential location and vehicle ownership choices, and individual-level job location choices. The temporal scale of this component ranges from days to years.

- **Medium-Term (MT):** This component contains a mesoscopic supply simulator coupled with a microscopic demand (daily activity) simulator (Lu et al., 2015; Basu et al., 2018). Daily travel decisions like mode choice, route choice, activity-travel patterns, and incident-sensitive (re)scheduling are considered at the temporal scale of minutes to hours, up to a single day.

- **Short-Term (ST):** This microscopic traffic simulator involves lane-changing, gap acceptance, route choice, and acceleration-braking behavior at the temporal scale of seconds to minutes (Azevedo et al., 2017).

The LT and MT components are connected through activity-based accessibility (ABA) measures that are disaggregate utility-based measures of the value of alternative daily activity patterns. Simulation of alternative scenarios in MT enables measurement of individual-level ABAs that would result from participation in those scenarios through logsums. For example, consider an individual living in the ith Traffic Analysis Zone (TAZ) and working in the jth TAZ. It is crucial to evaluate hypothetical scenarios such as (a) *fixed home-variable work*, where the individual can choose from all possible TAZs for their work location while keeping the residential location (TAZ i) fixed, and (b) *variable home-fixed work*, where the work location is fixed but the residential location is allowed to vary. For every possible combination of home and work TAZs, MT provides a logsum value for each individual in the synthetic population. These ABAs are used as explanatory independent variables in LT models such as residential location choice and job location choice. As the focus of this study is solely on the LT component, readers are directed to He et al. (2019) for more details on ABA formulation, estimation and use in LT models.

6.3.2 Behavioral models in SimMobility-Long Term

The LT simulator constitutes a housing market module that simulates daily dynamics in the residential housing market. The synthetic population is constructed prior to the simulation, and assumed to be the state of the system for "*day-0*", or t=0. Households are assigned particular residential units in specific buildings. While unoccupied units available for sale constitute the housing market supply, we also allow for resales, new sales and advance purchases. Asking prices are determined by sellers through a *hedonic price model*, while *willingness-to-pay* (WTP) of buyers is evaluated through an econometric framework based on expected utility maximization. These models are specified as a linear-in-parameters function of the following variables:

- **Neighborhood characteristics**, which represent spatial location
- **Housing unit attributes**, which represent heterogeneity in the housing supply
- **Household demographics and socio-economics**
- **Accessibility measures**, which are captured through ABA values

We represent a generalized functional form for the ith unit below.

$$P_i = \beta_0 + \boldsymbol{\beta_1} \cdot \boldsymbol{X^i_{sp}} + \boldsymbol{\beta_2} \cdot \boldsymbol{X^i_{unit}} + \boldsymbol{\beta_3} \cdot \boldsymbol{X^i_{dem}} + \boldsymbol{\beta_4} \cdot \boldsymbol{X^i_{econ}} + \boldsymbol{\beta_5} \cdot \boldsymbol{X^i_{ABA}} \tag{6.1}$$

Housing market transactions are modeled as a daily bidding process among active buyers and sellers. This is a salient feature of *SimMobility*, which (to the best of our knowledge) is the only LUT microsimulator that uses a daily bid-auction housing model at the metropolitan area scale using econometric frameworks. Potential buyers become active on the market based on explicit probabilistic models for awakening. The *awakening model* determines the likelihood that a currently inactive household will become active and begin searching for alternative residential locations on any given day.

A similar approach is adopted for *screening* of residential units based on housing and neighborhood type, which results in construction of finite-sample and economically plausible choice sets. This is motivated by the typical search process wherein households tend to narrow their search toward particular neighborhoods and unit types before investing more time and energy to visit and compare individual housing units. Active buyers evaluate each

residential unit in their choice set and select a unit to bid on, based on maximization of expected consumer surplus. If a household has a successful bid, they move into the new unit, and then reassess the job and school assignments of household members along with their vehicle availability. It should be noted that these moving rates vary by household demographics and tenure.

Household vehicle availability is modeled through a two-stage framework. The first stage involves estimating a *taxi availability model* that predicts whether the household owns a taxi. We use the Household Interview Travel Survey (HITS) data from 2012 for model estimation. It is worth noting that the HITS underestimates taxis by about 7,500. We impute these 7,500 taxis among the weighted HITS sample using socio-demographic criteria to select eligible households, and an iterative allocation procedure to assign additional taxis to new households. The taxi availability model is estimated using the binary logit framework with household socio-demographics as explanatory variables. The model is then used to predict taxi availability as a binary value (taxi owner or not) for households in the synthetic population. A generalized functional form for the ith household is expressed below.

$$V^i_{taxi} = \beta_0 + \boldsymbol{\beta_1} \cdot \boldsymbol{X^i_{dem}} \tag{6.2}$$

$$P^i_{taxi} = \frac{1}{1 + \exp\left(V^i_{taxi}\right)} \tag{6.3}$$

The second stage involves estimating a *vehicle availability model* that utilizes the results of the first-stage taxi availability model. We combine the HITS data and the taxi availability predictions with local accessibility measures (such as distance to nearest MRT station from residential location) and individual-level ABA values as explanatory variables. Our model considers vehicle availability as a mobility bundle rather than segregated ownership of specific vehicle types, as is the norm in most literature. Six vehicle availability categories are considered, which follow an increasing scale with regard to utilities derived from mobility.

- **CAT 0:** No private vehicles
- **CAT 1:** One or more motorcycles only
- **CAT 2:** One off-peak car w/wo motorcycles
- **CAT 3:** One normal car only

- **CAT 4:** One normal car with other vehicles
- **CAT 5:** Two or more normal cars w/wo other vehicles

Six logsum values are calculated for each individual keeping their residential and job locations fixed (i.e. one for each of the six categories). We aggregate the individual logsums to the household level by considering the highest income-earner as the head of the household. Ties are broken by considering the individual with the highest logsum for the '*one normal car*' category (CAT 3). Due to the possibility of six choices, the multinomial logit framework is used to estimate this model. Finally, we use the estimated model to predict vehicle availability as a categorical variable (with six categories) for households in the synthetic population. We show a generalized functional form for the ith household below.

$$V_i = \beta_0 + \boldsymbol{\beta_1} \cdot \boldsymbol{X^i_{dem}} + \boldsymbol{\beta_2} \cdot \boldsymbol{X^i_{taxi}} + \boldsymbol{\beta_3} \cdot \boldsymbol{X^i_{access}} + \boldsymbol{\beta_4} \cdot \boldsymbol{X^i_{ABA}} \tag{6.4}$$

$$P_i = \frac{\exp\left(V_i\right)}{\sum_j \exp\left(V_j\right)} \tag{6.5}$$

We are forced to limit ourselves to a high-level discussion of the framework in this chapter for brevity. The reader should refer to Zhu et al. (2018) for further details on *SimMobility Long-Term* and estimation results of these models.

6.4 Methodology

First, we provide details about the car-lite policy we seek to examine. This is followed by the selection of our study areas. In subsequent sub-sections, we define various scenarios and outline the assumed changes in behavior due to the policy.

6.4.1 Car-lite Policy

We assume that the municipal government (or the planning agency) introduces a car-lite policy in a certain region, which we call the *study area*, as a pilot. This policy entails the restriction of private vehicles and use of AVs only in the study area, and nowhere else in the metropolitan area (termed as the *spatial setting*). Inside the study area, AVs can be used for both ride-hailing (single passenger trips like UberX or JustGrab) and ride-sharing (multi-passenger shared trips like UberPool or GrabShare) services.

We anticipate that the introduction of AVs will lead to a positive perception of the study area in two aspects. First, households living outside the study area may want to move in and enjoy the benefits associated with AV availability. Second, households living inside may want to reconsider private vehicle ownership due to widespread availability of AVs for trip making. These anticipated effects are particularly relevant to a region like Singapore, where about 80% of households live in public housing provided by the Housing & Development Board (HDB) and private vehicle ownership is prohibitively expensive. The reader should refer to Basu and Ferreira (2019) for further details on vehicle ownership in the Singaporean context like the requirement of a Certificate of Entitlement (COE) for private vehicle registration. However, it is worth highlighting here that our proposed framework is not limited to Singapore; we merely note the relevance of such a car-lite policy to the Singaporean context.

6.4.2 Study Areas

We consider Singapore as our spatial setting for this paper. Singapore has around 1.15 million households[1] and 1.24 million residential units, leading to an island-wide vacancy rate of 7.10%. There are 55 planning districts in Singapore, among which we choose four districts grouped into three separate study areas where the car-lite policy could conceivably be introduced (see Figure 6-1). Key description metrics for these three study areas are reported in Table 6.1, and compared to those for Singapore as a whole.

Our choice of Toa Payoh is motivated by considering a plausible albeit somewhat extreme case, i.e. a study area neither in the central business district nor in the periphery, and which is already more vehicle-free than Singapore on average. This planning district has 43,841 households (3.82% of total households) and 45,706 residential units (3.70% of total housing stock), implying a vacancy rate of 4.08%. It should be noted that Toa Payoh has a lower vacancy rate than the country as a whole.

Moreover, the average vehicle ownership in Toa Payoh is already relatively low with a no-vehicle share of 66.55% vs. 54.28% in Singapore. Therefore, we expect a muted effect of the policy on residential mobility (due to lower vacancy rate) and private vehicle ownership (due to higher no-vehicle share right at the outset). This is beneficial in establishing a

[1]We do not include construction workers, work permit holders, and other foreigners in our analysis because they do not have full access to the housing market.

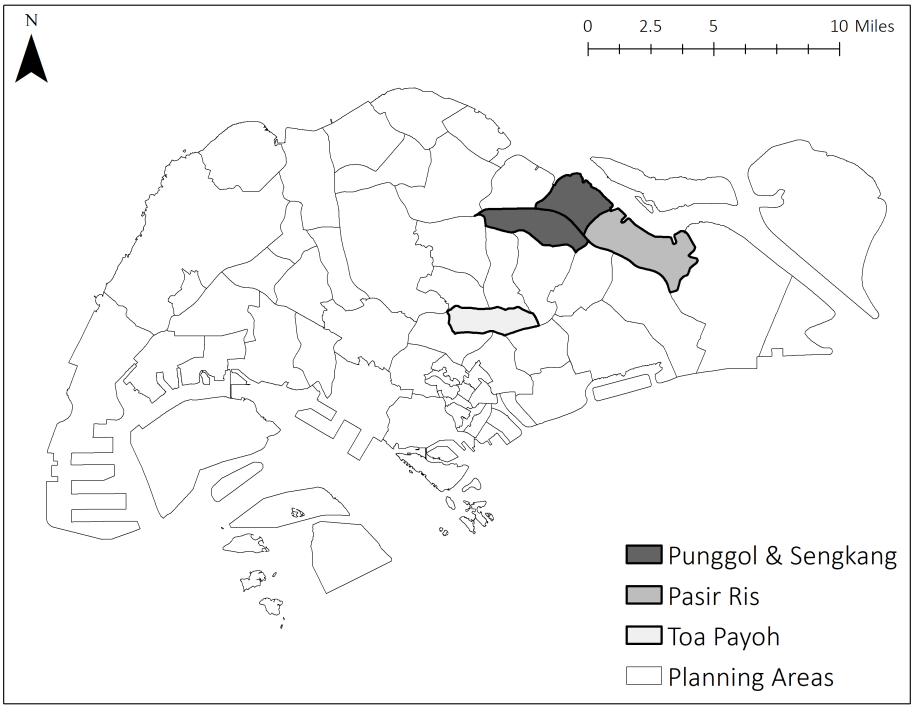


Figure 6-1: Location of study areas in the spatial setting of Singapore

conservative estimate of the magnitude of policy impacts, as other planning districts with less vehicle-free initial settings will enjoy higher benefits from the car-lite policy.

Table 6.1: Description of study areas in Singapore

Planning Area	**HHs**	**Units**	**Vacant units**	**Vacancy rate (%)**	**No-vehicle market share (%)**
Toa Payoh	43,841	45,706	1,865	4.08	66.55
Pasir Ris	38,138	40,989	2,851	6.96	49.01
Punggol & Sengkang	68,733	78,817	10,084	12.79	51.62
Singapore	1,148,066	1,235,833	87,767	7.10	54.28

We next consider Pasir Ris as our second study area. While having a similar number of households and housing units as Toa Payoh, Pasir Ris has a higher vacancy rate in addition to a higher vehicle availability rate (as evidenced by a significantly lower no-vehicle market share). It is worth highlighting here that Pasir Ris has a greater number of vehicles than the whole of Singapore on average, but a similar vacancy rate. Finally, we combine the districts of Punggol and Sengkang to create our third study area. The population here is almost double that of the other two study areas, and so is the vacancy rate. However, the

no-vehicle market share is much closer to the national average. In summary, readers can interpret the study area selection as follows:

- **Toa Payoh:** Lower vacancy rate and lower vehicle availability rate (compared to the national average) represents a tight market, and will produce a lower bound on the estimated policy impacts.

- **Pasir Ris:** National vacancy rate and higher vehicle availability rate allow for examining the effect of initial vehicle availability in isolation.

- **Punggol & Sengkang:** Higher vacancy rate and national vehicle availability rate allow for examining the effect of initial vacancy rate in isolation.

6.4.3 Scenario Design

To examine the impact of this policy on household vehicle availability, we adopt a scenario-based approach:

- **Baseline:** We conduct a baseline run that simulates the long-term evolution of Singapore, assuming that the car-lite policy was never introduced. All parameters in the behavioral models are held constant and equal to the estimated values.

- **Scenario I (Minimal effect):** The car-lite policy is now introduced in the study area. This results in (a) an increase in the likelihood that housing units in the study area are included in a household's choice set, and (b) an increase in local accessibilities to different amenities in the study area. It is worth highlighting that there are no other market effects in this scenario, apart from an awareness about the policy being introduced in the study area.

- **Scenario II (Buyer valuation increases):** In addition to the changes in the previous scenario, we hypothesize that the policy leads to an increase in demand for housing in the study area. The increased demand can be represented through parameter modifications in the WTP model, which reflects the valuation of housing from the buyer's perspective. This scenario can be thought of as a *short-run market reaction*, where only consumers have reacted to the policy.

- **Scenario III (Both buyer & seller valuations increase):** We construct this scenario as a representation of the *longer-run market reaction*, where both consumers and suppliers have reacted to the policy. The market has had enough time to respond to the increased demand through an increase in the asking price of units inside the study area, which is captured through parameter modifications in the hedonic price model that reflects the valuation of housing from the seller's perspective. Modifications to the local accessibilities and WTP parameters are made in a fashion similar to the previous scenario.

The aforementioned model parameter modifications are represented through the following behavioral assumptions:

- **Good marketing:** We assume that the screening probability doubles for all residential units inside the study area. This implies that such units are twice as likely to appear in the choice set of any household searching for a new home.

- **Higher buyer valuation:** Recall that ABA is used as an explanatory variable in the WTP model. Due to the car-lite policy implying increased accessibility, we increase the ABA by 0.25 times the standard deviation across all TAZs if the household lives *or* works in the study area. For households living *and* working in the study area, the ABA is increased by 0.5 times the standard deviation across all TAZs.

- **Higher seller valuation:** The hedonic price model includes local accessibility measures (i.e., distances to amenities such as shopping malls, bus stops, MRT stations, schools, etc.) in addition to ABA as explanatory variables. In addition to increasing ABA using the aforementioned technique, we boost the local accessibilities of postcodes inside the study area by halving the perceived distances to the four amenities exemplified above. These modifications cause an increase in the asking price of units, which is a result of higher seller valuation.

- **Improved no-vehicle accessibility:** Recall that we estimate ABA for the *fixed home-fixed work-varying vehicle availability* setting, i.e., we calculate six logsum values pertaining to the six vehicle availability categories for each household assuming fixed residential and job locations. This assumption leads us to consider that the public transit accessibility (or no-vehicle accessibility, i.e., CAT 0 logsum) is at least as good

as that of the next category (motorcycles only, i.e., CAT 1).

6.4.4 Sequential simulation framework for car-lite policy analysis

In light of the above discussion, it is clear that the car-lite policy will affect both residential mobility and private vehicle ownership decisions. Therefore, we consider the housing-mobility bundle, which combines residential location choices with vehicle availability choices, for joint evaluation. A summary of our proposed framework is illustrated in Figure 6-2.

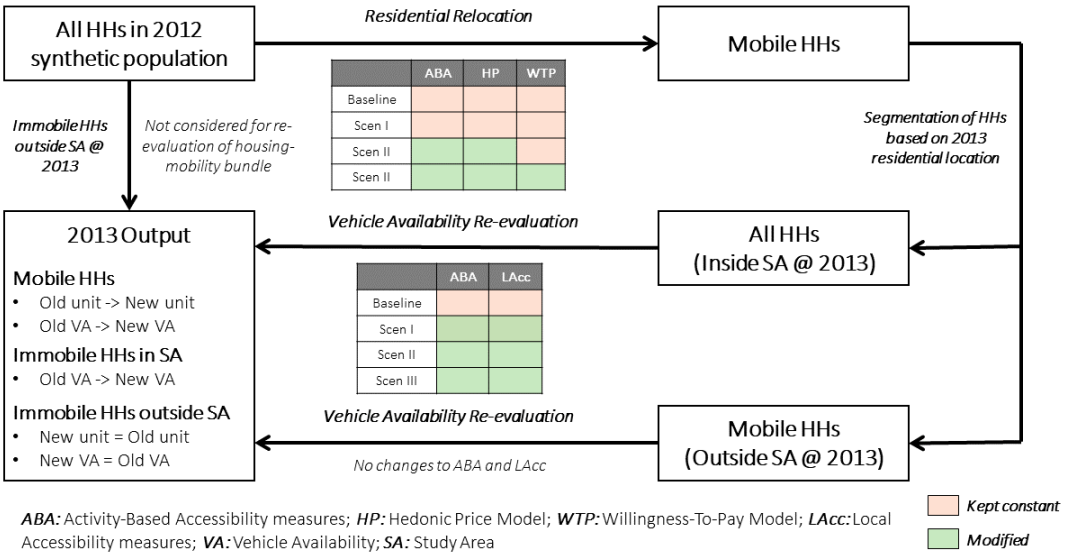


Figure 6-2: Sequential simulation framework for the housing-mobility bundle

All the households in the synthetic population in the base year (2012) are considered to be eligible for the daily bidding process. We use the bid-auction housing model described earlier to simulate housing market transactions for the simulation period (considered as one year in this study). Recall that the baseline scenario does not require any model parameter modifications. Scenario I involves modifying vehicle availability model parameters, but no changes are made to parameters for the housing models. The ABA and WTP parameters are modified in Scenario II with additional modifications made to the hedonic model for Scenario III. At the end of the simulation period, we are able to identify households who had successful bids along with their new residential locations. Households that changed units, i.e. moved during the simulation period, are termed as *mobile households*. Households whose residential locations remained the same in 2013 as in 2012 are called *immobile households*. It should

be noted that immobile households include those that never bid on a new unit along with those that had unsuccessful bids.

All mobile households are considered to re-evaluate their vehicle availability based on their new residential location in addition to households inside the study area at the end of the simulation period in 2013 (irrespective of whether they moved or not). New vehicle availability logsums are calculated for mobile households using their new residential location. Moreover, all households inside the study area in 2013 experience modifications in ABA values due to the improved no-vehicle accessibility assumption and the augmentation of perceived local accessibility values. However, these local accessibility modifications are not applicable to mobile households outside the study area, for whom the only update is the calculation of new vehicle availability logsums, as they cannot enjoy the increased accessibility inside the study area.

The new vehicle availability categories are calculated for both mobile households and households in the study area based on the above modifications to the vehicle availability model. Therefore, we obtain a new population after the simulation period in 2013 comprising (a) mobile households with new residential locations and new vehicle availabilities, (b) immobile households inside the study area with new vehicle availabilities, and (c) immobile households outside the study area with no change to either residential location or vehicle availability.

Recall from our literature review that residential relocation and vehicle ownership are often considered as simultaneous decisions. However, we consider them sequential decisions in our simulation framework because of computational practicality [2]. While considering residential relocation, the choice set of each household comprises 30 alternatives. If we were to consider a simultaneous framework, we would need to calculate at most (1.1 million households * 30 residential choices * 6 vehicle choices =) 198 million logsum values in total. Note that all of this has to be done on the fly because each household's choice set is constructed in a stochastic manner during the simulation due to which these logsum values cannot be pre-computed and stored in a database. Our simultaneous framework allows us to scale down the computation time significantly as at most (1.1 million households * 30

[2]There is an additional reason to favor the sequential framework. As our WTP function is not sensitive to the individual household's logsum (due to the use of logsums averaged across each TAZ), sequential consideration will lead to the same result. Given this setting, there is no reason to add computational burden by considering the simultaneous framework. That being said, we are exploring heuristics to improve performance for a future scenario when we have a more sensitive WTP model.

residential choices =) 33 million logsum values are required in the first phase and (1.1 million households * 6 vehicle choices =) 6.6 million logsum values are required in the second phase, leading to a total of 39.6 million logsum values. Thus, we obtain computational savings to the tune of (198 - 39.6 / 198 =) 80% by using a sequential framework.

6.5 Results & Discussion

We conducted five simulation runs for a one-year period, i.e. from 2012 to 2013, related to each of the four scenarios described in the previous section. The results from these simulations are discussed in the following sub-sections. Note that we report the average values across the five runs, with standard deviation reported when deemed appropriate and significant.

6.5.1 Stochasticity of simulated market shares

Recall the structure of the multinomial logit model from Eq. (6.4), which provides deterministic probabilities for each alternative in every household's choice set. However, it is standard practice to use multiple realizations of the maximum likelihood estimator during prediction. As maximum likelihood estimates are asymptotically normally distributed, we draw from a multivariate normal distribution $N(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\Sigma}})$ where $\hat{\boldsymbol{\beta}}$ is the vector of estimated parameters and $\hat{\boldsymbol{\Sigma}}$ is the estimated variance-covariance matrix. Using ten random draws, we calculate individual choice probabilities for each realization.

Consequently, aggregate market shares are calculated for each simulation and corresponding summary statistics are reported in Table 6.2. We find that the average simulated market shares are reasonably close to the true (i.e., HITS 2012) market shares[3]. The standard deviation is low enough to be termed as insignificant. Therefore, we can be confident about the stability of our simulation procedure to predict a vehicle availability category for each household.

[3]An older version of the VA model is currently implemented in SimMobility, leading to different market shares compared to those reported in previous chapters. This earlier model did not include the techniques we showed earlier to address misrepresentation of low-sample alternatives. However, this is not as relevant to this chapter, as our intent is to compare results across different scenarios.

Table 6.2: Stochasticity of simulated vehicle availability market shares [a]

Category	Description	HITS 2012 market share (%)	Mean simulated market share (%)	Std. Dev.
0	No vehicles	54.28	49.48	0.04
1	One motorcycle only	4.39	4.12	0.01
2	One off-peak car w/wo motorcycles	2.03	2.23	0.01
3	One normal car only	33.96	35.81	0.05
4	One normal car with other vehicles	1.00	1.18	0.01
5	Two normal cars w/wo other vehicles	4.34	7.19	0.01

[a] Sample mean and standard deviation (of sample mean) across ten simulation runs are reported.

6.5.2 Computational performance

We used a 64-bit machine with an Intel Xeon E3-1505M v5 (2.80GHz) CPU and 32GB RAM. Every simulation was run using eight threads on an Ubuntu 18.04.1 LTS (Bionic Beaver) OS. The average simulation run time for every combination of study area and scenario is reported in Table 6.3, along with the standard deviations. We find that the results are consistent with our expectations.

The run time for Scenario I is almost equal to that of the baseline because the only additional burden in Scenario I is related to the modifications to the vehicle availability model. Due to changes in the residential relocation process in Scenario II, more movers are simulation as a result of increased demand. This requires computation of a considerable amount of logsums on the fly, thereby increasing the run time compared to Scenario I. However, the increased price setting in Scenario III manages to reduce the number of movers, and consequently the number of on-the-fly computations, which results in a decrease in run time. Recall that Punggol & Sengkang had a significantly higher population and vacancy rate compared to the other study areas. This manifests itself in higher run times across all three scenarios.

Table 6.3: Simulation run times (in minutes) [a]

Study Area	Scenario	Mean	Std. Dev.
-	Baseline	230	14
Toa Payoh	Minimal effect	231	11
	Demand increases	239	18
	Demand and price increase	227	10
Pasir Ris	Minimal effect	232	14
	Demand increases	253	10
	Demand and price increase	239	18
Punggol & Sengkang	Minimal effect	243	15
	Demand increases	255	23
	Demand and price increase	244	17

[a] Mean and standard deviation values across five runs are reported.

6.5.3 Temporal variation in the study areas

The car-lite policy certainly increases the attractiveness of the study area, as evidenced by more households moving into the study area than moving out. This in-migration leads to an increase in the total population of the study area. However, that does not always translate to more car-lite behavior at the aggregate population level. Additionally, despite the in-migration, aggregate socio-demographics of the study area population (such as household size, income, presence of children and seniors, etc.) do not undergo statistically significant change in any scenario.

Table 6.4 summarizes the aggregate no-vehicle market shares in the study areas for the four scenarios, while comparing them to the original market shares in the study areas in 2012. Note that only mean values are reported. Standard deviations for the rows related to original market shares are obviously zero, while those for rows related to the scenarios are insignificantly low (*min:* 0.01 and *max:* 0.08). Recall that Toa Payoh was more vehicle-free than Singapore in 2012, where 66.55% of households did not own any private vehicle compared to 54.28% island-wide. We see that the no-vehicle share decreases in Toa Payoh even in the baseline scenario when the car-lite policy was not introduced. With the introduction of the policy in Scenario I, we witness a comparative increase in the no-vehicle market share, which is a positive indication towards the effectiveness of the policy. However, as

market effects start to creep in, the no-vehicle share decreases significantly and is eventually lower than even the baseline share, which is counter-productive to the motivation behind the policy.

Table 6.4: Changes in no-vehicle market shares in the study areas over time [a,b]

Study Area	Scenario	Full SA population (%)	Non-movers inside SA (%)
Toa Payoh	Original (2012)	**66.55**	-
	Baseline	-0.46 (0.09)	+0.33 (0.07)
	Minimal effect	-0.18 (0.06)	+0.37 (0.05)
	Demand increases	-1.01 (0.18)	+0.31 (0.04)
	Demand and price increase	-0.68 (0.10)	+0.34 (0.04)
Pasir Ris	Original (2012)	**49.01**	-
	Baseline	+0.77 (0.15)	+0.20 (0.03)
	Minimal effect	+0.96 (0.08)	+0.20 (0.06)
	Demand increases	-0.01 (0.08)	+0.29 (0.90)
	Demand and price increase	+0.25 (0.12)	+0.23 (0.03)
Punggol & Sengkang	Original (2012)	**51.62**	-
	Baseline	+0.38 (0.17)	+0.11 (0.07)
	Minimal effect	+0.59 (0.08)	+0.10 (0.06)
	Demand increases	+0.15 (0.05)	+0.15 (0.04)
	Demand and price increase	+0.29 (0.08)	+0.09 (0.07)

[a] Changes are reported in a comparison with the original study area market share in 2012.

[b] Mean values are reported with standard deviations shown inside parentheses.

This can be interpreted as a reluctance to give up private vehicles despite enjoying the benefits of the car-lite policy, thereby reinforcing our hypothesis of the endowment effect. The households that moved in were more likely to own a car in 2012, compared to those that moved out. Considering the overall in-migration, our results indicate that the number of households that gave up their private vehicles after moving into the study area was not high enough to influence Toa Payoh positively. Thus, the in-movers were less vehicle-free to begin with and remained so even after moving in, compared to the highly vehicle-free initial population in the study area.

The trends observed for Toa Payoh are consistent for the other two study areas as well,

i.e. the policy is effective in the beginning but proves to become counter-productive as market effects set in. However, we find that the policy tends to provide an overall benefit to these study areas, unlike Toa Payoh. The highest benefit considering market effects is enjoyed by Punggol & Sengkang, while Pasir Ris enjoys the highest benefit if market effects are ignored. We would also like to highlight that non-movers inside the study area are more likely to forgo their private vehicles due to the accessibility benefits associated with the car-lite policy. This is evidenced by the positive values in the last column of Table 6.4, which are consistently positive for all combinations of study areas and scenarios. The difference in magnitudes across study areas can be attributed to the variations in the initial settings in the study areas, which would create a difference in the amount of relative accessibility gain due to the policy. It is worth keeping in mind that vehicle availability choice transitions are driven by relative, *not absolute*, utility gains; this is the very basis of a random utility model.

6.5.4 Behavioral variations across movers

Although the study area does not change significantly over the simulation period in terms of aggregate statistics, we expect behavioral variations across movers. Therefore, we categorize two types of movers:

- **In-movers:** These households experienced a residential relocation during the simulation period that resulted in them relocating to a unit inside the study area. They could have moved into the study area from outside ($OUT \rightarrow IN$ transition), or could have moved but always remained inside the study area ($IN \rightarrow IN$ transition).
- **Out-movers:** These households were inside the study area in 2012 but moved out during the simulation, and live outside the study area in 2013 ($IN \rightarrow OUT$ transition).

While an $OUT \rightarrow OUT$ transition representing mobile households that were always outside the study area throughout the simulation period is also possible, we are interested only in the aforementioned categories for the purposes of this case study. The cohort sizes are reported in Table 6.5 along with the original study area population for comparison. We find that Toa Payoh experiences a net in-migration of around 1,000 households in the baseline scenario. This increases by an additional 200 households as the policy is introduced in Scenario I. We notice a further increase as demand for the study area is augmented in

Scenario II. However, when price increases along with demand (Scenario III), fewer households are able to move in because of affordability constraints leading to a decrease in net in-migration. It is worth noting that there is always a positive net in-migration effect for every combination of study area and scenario.

Table 6.5: Residential mobility and migration in the study areas [a,b,c]

Study Area	Scenario	SA Pop. (2012)	In-movers	Out-movers	Net effect
Toa Payoh	Baseline	43,841	3,914 (715)	2,897 (121)	**+1,017**
	I		4,123 (53)	2,923 (45)	+18%
	II		4,223 (51)	2,886 (30)	+31%
	III		4,035 (36)	2,902 (29)	+11%
Pasir Ris	Baseline	38,138	4,542 (52)	2,961 (42)	**+1,581**
	I		4,482 (38)	2,985 (61)	-5%
	II		4,649 (67)	2,915 (47)	+10%
	III		4,421 (53)	2,972 (49)	-8%
Punggol & Sengkang	Baseline	68,733	8,520 (592)	6,249 (154)	**+2,271**
	I		8,508 (130)	6,319 (43)	-4%
	II		9,061 (143)	6,107 (64)	+30%
	III		8,831 (104)	6,255 (54)	+13%

[a] Scenario-specific migration effects are reported in comparison with the baseline migration effect.

[b] Mean values are reported with standard deviations shown inside parentheses.

[c] Scenario I: *Minimal effect*; Scenario II: *Buyer valuation increases*; Scenario III: *Both buyer and seller valuation increase*

Similar findings are observed for Pasir Ris and Punggol & Sengkang, albeit with different magnitudes. Recall that Toa Payoh had a low vacancy rate, while the rate for Pasir Ris was higher and close to the national average. The additional number of vacant units causes more market transactions, which is reflected in the net in-migration baseline effect for Pasir Ris. Punggol & Sengkang had an even larger vacancy rate, and consequently have a comparatively larger baseline in-migration effect. For both these study areas, the in-migration effect in Scenario I is similar to the baseline, unlike Toa Payoh where an immediate increase was noticeable. However, there is a significant comparative increase as demand increases in Scenario II, which is moderated by the consequent price increase in Scenario III. While the

effect in Scenario III is worse than the baseline for Toa Payoh and Pasir Ris, the converse is true for Punggol & Sengkang. This points towards the considerably high vacancy rate of this study area overpowering the effect of increased prices in Scenario III.

We also explored socio-demographic variables such as household size, presence of seniors, presence of children, and ethnicity. There are no statistically significant differences among in-movers and out-movers with regard to these socio-demographics. However, significant differences in household income are noticeable across the four scenarios. Movers are richer than the overall population on average in tight real estate markets like Toa Payoh, but there are mixed effects for more open markets. The additional market effects causing changes in the relative attractiveness of the study areas result in further income gaps among movers. In-movers are usually richer than out-movers for most scenarios in Tao Payoh and Pasir Ris, as evidenced in Table 6.6. However, the converse is true for Punggol & Sengkang, where the high vacancy rate would have resulted in lower unit prices as the real estate market was relatively open.

Out-movers are closer to the original study area population in terms of average income for all study areas. We also notice that market effects cause the unit prices to shoot up, thereby causing relatively higher-income households to move into the study areas. It is also worth noting that average incomes are lower in Scenario III compared to Scenario II despite rising prices. This may be attributed to extremely higher-income households not perceiving a worthwhile gain in accessibility for the higher market price. Therefore, they choose to bid elsewhere, while the middle- and upper-middle-income groups still bid on units inside the study area as the relative accessibility gains are worth the higher price. While this does not have any negative consequences for relatively open markets like Pasir Ris and Punggol & Sengkang, a tight market like Toa Payoh witnesses an average rise in household income between 16% to 24% for in-movers. This would lead to a long-term gentrification effect, which might be undesirable for several reasons. However, we limit ourselves to examining the implications of gentrification with regard to only vehicle availability for this study, as higher-income households are more likely to own private vehicles.

6.5.5 Vehicle availability transitions

We focus on mobile households to examine the transitions between vehicle availability categories over the simulation period. Figure 6-3 represents the inter-categorical transitions

Table 6.6: Average monthly household income (in SGD) in the study areas [a,b,c]

Study Area	Scenario	SA Pop. (2012)	In-movers	Out-movers
Toa Payoh	Baseline	$6,557	+9%	+8%
	I		+3%	+9%
	II		+24%	+6%
	III		+16%	+8%
Pasir Ris	Baseline	$9,123	-13%	+3%
	I		-12%	+3%
	II		+3%	+4%
	III		+0%	+3%
Punggol & Sengkang	Baseline	$8,409	-16%	-1%
	I		-15%	-2%
	II		-6%	-2%
	III		-4%	-2%

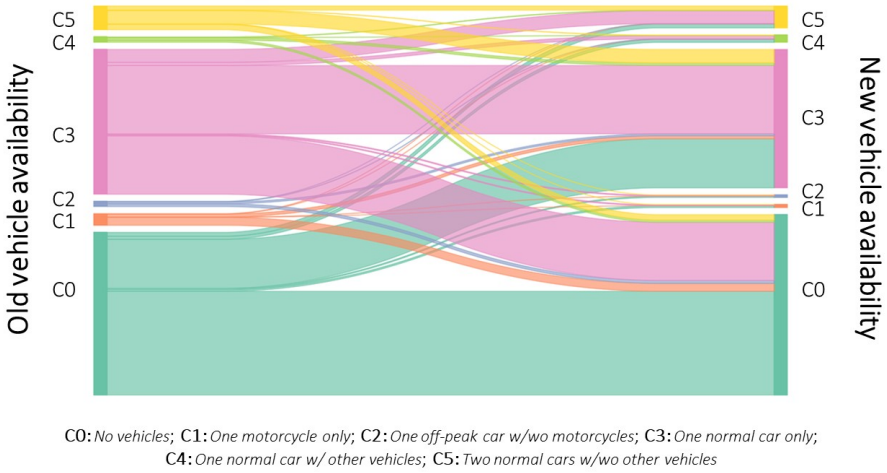
[a] Cohort-specific income changes are reported in comparison with the original study area population income.

[b] Only mean values are reported due to insignificantly low standard deviations.

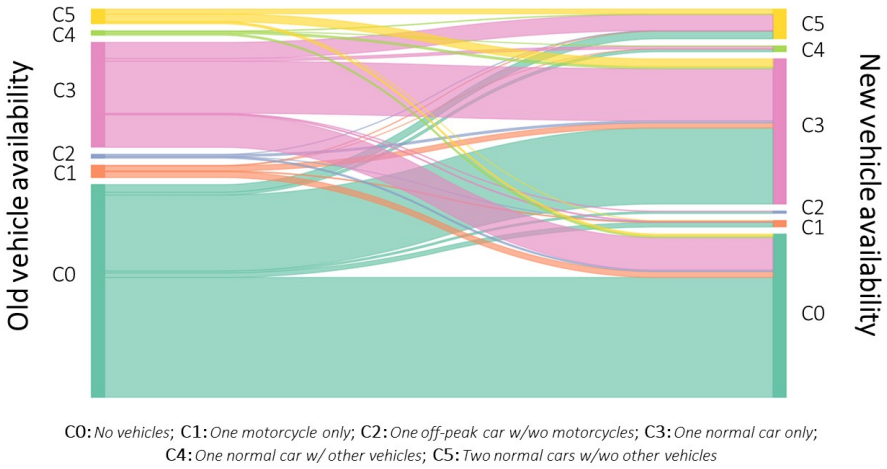
[c] Scenario I: *Minimal effect*; Scenario II: *Buyer valuation increases*; Scenario III: *Both buyer and seller valuation increase*

for in-movers and out-movers in Toa Payoh for Scenario II. While similar plots can be constructed for every combination of study area and scenario, we focus on this particular combination only as it has the highest gentrification effect. Out-movers were more likely to have no vehicles initially, as evidenced by the original no-vehicle share of 60% compared to 46% for in-movers. This is a result of Toa Payoh being more vehicle-free than Singapore on average. However, the car-lite policy succeeds in reversing the behavior. At the end of the simulation in 2013, 51% of in-movers are vehicle-free compared to 46% of out-movers. More importantly, 40% of in-movers with a private vehicle became vehicle-free compared to only 31% of private vehicle owning out-movers. While 56% of vehicle-free out-movers remained devoid of private vehicles, 64% of vehicle-free in-movers chose to do the same. Similar trends were observed for Scenario III, which is why we omit reporting those results for brevity. It should be noted that these trends are consistent across study areas, albeit with different

magnitudes.



(a) In-movers



(b) Out-movers

Figure 6-3: Transitions for mobile households in Toa Payoh (Scenario II)

In addition to the transitional behavior discussed above, we examine the aggregate no-vehicle market shares among mobile households across study areas and scenarios. As noted earlier, the initial population in Toa Payoh owns fewer vehicles than Singapore on average and in-movers are more likely to own vehicles compared to out-movers. We see from Table 6.7 that out-movers are almost always more vehicle-free than in-movers, except for Scenario I where the car-lite policy is able to positively influence in-movers to give up their vehicles. However, when market effects are considered, in-movers are significantly less vehicle-free as

they refuse to let go of their private vehicles. As alluded to earlier, this is a result of higher demand causing richer households to move to the study area, who are reluctant to give up their private vehicles despite enjoying benefits from the car-lite policy. There is a decrease of almost 5% in the no-vehicle share among in-movers when Scenario II is in effect. However, the rate of decrease in Scenario III is much lower (about 2%), and is comparatively more palatable.

Table 6.7: No-vehicle market shares for study area movers [a,b,c]

Study Area	**Scenario**	**SA Pop. (2012)**	**In-movers (%)**	**Out-movers (%)**
Toa Payoh	Baseline	66.55	**46.87**	**48.22**
	I		+2.44 (0.75)	-0.92 (0.65)
	II		-4.49 (0.98)	+0.24 (1.40)
	III		-1.68 (0.36)	-0.67 (0.84)
Pasir Ris	Baseline	49.01	**37.91**	**34.96**
	I		+0.40 (0.95)	-0.01 (0.70)
	II		-6.09 (0.89)	-0.74 (1.32)
	III		-4.94 (1.30)	-0.68 (0.86)
Punggol & Sengkang	Baseline	51.62	**39.82**	**37.44**
	I		+1.29 (0.75)	+2.90 (0.93)
	II		+1.08 (0.43)	+2.78 (0.43)
	III		-2.64 (0.48)	+2.08 (0.50)

[a] Scenario-specific changes in the no-vehicle market share are reported in comparison with the baseline market share.

[b] Mean values are reported with standard deviations shown inside parentheses.

[c] Scenario I: *Minimal effect*; Scenario II: *Buyer valuation increases*; Scenario III: *Both buyer and seller valuation increase*

Examining results for all study areas, we find that in-movers are less prone to vehicle-free behavioral shifts that are evidenced by their decreasing no-vehicle shares across scenarios. The initial study area settings tend to matter significantly, as Pasir Ris (which had the lowest initial no-vehicle market share) experiences the highest decrease in that market share. Recall that Pasir Ris had the highest average income as well. This reinforces our hypothesis that higher-income households with initial vehicle holdings do not perceive the policy to provide

sufficient accessibility gains to warrant a shift to no vehicles. Similar behavior is observed in Toa Payoh, albeit with a smaller magnitude. Punggol & Sengkang, whose no-vehicle market share was close to the national average, witnesses an increased share for in-movers even with market effects in Scenario II. Unfortunately, that swings to a decrease in Scenario III when the increased price effect sets in.

On the other hand, the no-vehicle share for out-movers does not change significantly across scenarios. This is perhaps because of the stochasticity in their final residential location, which drives their vehicle re-evaluation decisions. While there are marginal decreases in no-vehicle market shares for out-movers from Toa Payoh and Pasir Ris, we observe an increase in the case of Punggol & Sengkang. Despite having a share close to the national average, the study area has poor overall accessibility. Therefore, when households relocate to greener pastures outside the study area, their accessibility gains are adequately high to warrant a decrease in private vehicle ownership. It is also worth noting that the in-movers never achieve the no-vehicle share of the original study area population in the presence of market effects, although they do become more vehicle-free when the policy exists without market effects in all study areas. This can be attributed to the higher incomes of in-movers, who are consequently more likely to own private vehicles and bring those with them to their new residential locations inside the study area.

6.6 Conclusion

Recent diversification of mobility options has motivated policy-makers to re-evaluate the very essence of private vehicle ownership. *How will the market for private vehicles get affected when on-demand and shared mobility services become ubiquitous?* Such questions become even more complex with the consideration of autonomous vehicles in the fleet mix. The research question of AVs substituting private vehicles has started receiving considerable attention in recent times. Two major research approaches emerge from the literature. First, simulation-based approaches seek to replace on-road vehicles by testing different AV fleet sizes or assuming different market penetration rates that can meet observed travel demand. Second, behavioral approaches use stated preference surveys or use cost-benefit analysis to elicit user preferences for AVs. We combine econometrically robust behavioral models with an agent-based microsimulation framework in our integrated land use-transport simulator

SimMobility, and then examine the extent to which accessibility changes in 'car-lite' communities can influence car ownership directly and indirectly through residential relocation.

A scenario-based design is adopted, wherein different types of market responses to the policy are modeled through various assumptions of changes in model parameters. We chose a study area (Toa Payoh) that has a lower market share of private vehicles than the overall metropolitan area to obtain conservative estimates of policy impacts. Two additional study areas (Pasir Ris and Punggol & Sengkang) were chosen with differing vacancy and vehicle ownership rates to isolate the effects of each type of initial setting. Our findings indicate that none of the study areas undergo significant change in socio-demographics over a one-year horizon despite positive in-migration rates. However, in-movers are richer and initially less vehicle-free than out-movers. The final share of vehicle-free households and the transition rates from private vehicle ownership to vehicle-free do display positive trends for in-movers when the policy is instituted. However, when market effects are also considered, they become significantly less vehicle-free compared to the original study area population. In summary, the car-lite policy is able to influence in-movers successfully to reduce private vehicle ownership, but once market effects take hold and affect housing prices and bidding results, the net effect disappears.

We also notice spatial differences in policy effects across the study areas. If the real estate market is relatively open to begin with (i.e., has a reasonably high vacancy rate like Pasir Ris and Punggol & Sengkang), the study area experiences a net increase in the no-vehicle market shares, even with the inclusion of market effects. Therefore, the car-lite policy fails in inducing intended outcomes in tight markets like Toa Payoh. This is mainly driven by the comparative average household incomes of in-movers and the original population, as well as the perceived gains in accessibility of these in-movers relative to their current residential locations. In summary, Punggol & Sengkang seems to be the best study area to pilot a car-lite policy, due to its net positive vehicle-free shift despite market effects. However, the vehicle availability of in-movers needs to be moderated (or perhaps even restricted) through a sub-clause in the policy. Leaving in-mover behavior in the hands of the market results in them "having the cake and eating it too", as they enjoy the accessibility benefits of the study area in addition to retaining their private vehicles.

Our work opens up several avenues for future research. First, analyzing the sensitivity to the model parameter assumptions during scenario design can help increase the robustness

of our estimates of the policy impacts. Second, exploration of other scenarios that represent different market reactions would be useful in observing variability in policy impacts. We are currently working on including other supply-side reactions such as accelerated development behavior leading to an increase in housing stock inside the study area. Third, we could replace our speculative assumptions regarding accessibility improvements with activity-based accessibility measures for individuals that have shared and autonomous mobility options added to their choice sets. The *SimMobility Medium-Term* team is working on simulating activity patterns for such scenarios and generating corresponding ABA values, which can then be used directly in our *Long-Term* models. Finally, the simulation framework would benefit from inclusion of a more robust vehicle availability model, such as the one shown in previous chapters. While we are confident in our comparative results, a better-fitting model would increase our confidence in the estimated magnitudes of the impact of the car-lite policy.

There are a few limitations in this study. First, we do not consider the cost of vehicle ownership in our framework. The vehicle availability model does not contain cost as an explanatory variable due to lack of reliable data. Second, some studies indicate that vehicle ownership changes are closely associated with changes in household demographics and life cycle events (such as changing a job or residential relocation), which we do not consider in their entirety. We are addressing both these issues by conducting a retrospective survey of Singaporean households that tracks all major life cycle and demographic changes over the past three years (2016-2018). Additionally, the survey captures specific details of all vehicles owned by the household, including purchase and transaction costs, frequency and reason of use, and major users. Third, our WTP model is not conditional on availability of mobility options. While the ideal method would be to estimate a WTP model that considers joint housing-mobility bundles as choices, obtaining relevant data for such a model is quite challenging.

Overall, we hope that this study provides useful contributions to both the academic literature and planning practice by demonstrating how a technically robust ILUT simulation framework (*SimMobility Long-Term*) can be utilized as a policy analysis tool to anticipate the net effect of various competing forces, depending on the manner in which we introduce new technologies that improve accessibility. The importance of considering market effects when considering scenario differences is also highlighted.

Chapter 7

Conclusion

We summarize the key findings from the three studies conducted in this thesis in the first sub-section of this chapter. This is followed by providing suggestions for future research efforts based on some of the limitations in our studies. Finally, concluding remarks are provided for placing this thesis in the broader context of policy and planning.

7.1 Key findings

This thesis undertook three studies to deal with the changing essence of vehicle ownership as emerging transportation technologies and services begin to transform the mobility landscape. We summarize the key findings from these studies as follows.

7.1.1 Predicting the impact of new mobility

Traditional behavioral models, such as discrete choice models, perform well when the objective is to gain insights into the underlying mechanisms driving decisions and choices. However, machine learning models perform comparatively better at prediction exercises, provided the training data is an accurate representation of the world that is being modeled. With the advent of big data and advances in computing resources, we are becoming increasingly equipped to translate the advantages of machine learning techniques to traditional behavioral modeling.

This study proposes a framework that aims to combine the high predictive power of machine learning with the interpretability of discrete choice models. The framework goes through a multi-stage process of recommending the best discrete choice model and best

variable specification for use in an empirical application of modeling household vehicle ownership in Singapore. We identify the features that have the most impact on the predicted outcome of the model, which is a useful exercise for high-dimensional feature spaces. Non-linearities along those dimensions are then transformed into feature representations through piece-wise linear functions, that help bring the predictive performance of the multinomial logit model (i.e., the best discrete choice model in this application) at par with the best machine learning models.

Similar to most real-world data sets, our data set too suffers from class imbalance, where the data are distributed across the alternatives in a skewed manner. While our augmented model performs well on aggregate measures, it abysmally under-predicts low-sample alternatives. This is a major concern if we want to predict the impact of new mobility services, especially because existing data sets will not have any data on these up and coming alternatives (or minimal data, if present). We propose augmenting the low-sample alternatives with synthetically generated samples drawn from the multi-dimensional feature space. After creating this synthetic and balanced population, samples are randomly drawn to maintain the original distribution across alternatives.

This method helps address sparseness in the multi-dimensional feature space and provides a sample that is both random and representative. We test this framework to predict the market shares of off-peak cars in Singapore in 2012 after removing their samples from the 2008 data set. The model is calibrated on the OPC-less 2008 data set, following which specific assumptions are made about the utility function for off-peak cars to use the 2008-model for predicting the full set of 2012 choices. Our framework is found to reduce the RMSE of 2012 market share predictions by 60% on average, while maintaining high aggregate predictive performance and interpretability.

7.1.2 Household dynamics in vehicle availability and use decisions

Building on the Mobility-as-a-Service (MaaS) paradigm, we argue that modelers need to focus their attention on vehicle availability rather than traditional vehicle ownership. We created six alternatives of multiple vehicles that form a choice set of mobility bundles, which can emulate scenarios with widespread proliferation of access-based services. The multinomial logit model was used to gain insights into the decision-making process of choosing these mobility bundles. In order to circumvent biased or erroneous reporting of the house-

hold head in traditional surveys, we examined the job location effects of both the male and female highest income earners in each household. Additional effects pertaining to household demographics, and housing attributes (in the form of location and unit type) were also included.

We find that the male highest income earner tends to dominate decisions involving motorcycles and normal cars. However, the female highest income earner influences the decision to avail of a mobility bundle involving both a normal car and other vehicles such as motorcycles and off-peak cars. Higher-income households with children and/or seniors have diverse mobility needs, especially as their activity spaces tend to stretch farther away from their residential location. Therefore, they are seen to be more likely to prefer mobility bundles involving one normal car.

Living in the suburbs, which have low-density neighborhoods with poor access to public transport, increases households' preference for such bundles. However, we find that living in high-density and mixed-use neighborhoods with good public transport and walking accessibility tend to increase the uptake of off-peak cars. This indicates that investing in such neighborhoods could be a worthwhile way to make households feel that the restrictions in usage of the off-peak car are worth the cost savings.

While several studies examine vehicle usage at an aggregate level by measuring household vehicle miles traveled (VMT) over an extended period of time, rarely does the intra-household decision-making in allocation of private cars receive attention. This is an important aspect of activity scheduling and travel behavior, as certain individuals in the household may never have access to the car despite the household owning one. For example, young working males in multi-generational families might be more likely to use public transit to commute, while an older worker gets to use the car. Therefore, it is essential to consider intra-household interaction effects in exercises that seek to model which individual is most likely to become the primary user of the household's car.

We find strong and statistically significant effects related to gender, age, and income, as evidenced by a mid-aged male earning the lion's share of the total household income being the most likely to become the primary user of the car. Effects of job location and children's school location are also found to be statistically significant. We think that this is likely to be an indication of increased trip-chaining and multi-purpose tour-making, although specific checks for such behavior were not implemented.

7.1.3 The impact of car-lite policies on mobility bundle choices

The relevance of the previous two studies is certainly indubitable as standalone studies that encourage better understanding of behavioral decision-making related to mobility bundles, especially as new services enter the market. However, they are equally important in a larger planning framework, as cities grapple to formulate regulatory policies that provide citizens access to mobility in an equitable manner. We illustrate the use of an integrated land use-transportation (ILUT) modeling framework that combines behavioral econometrics with agent-based microsimulation. By using behavioral models to estimate willingness-to-pay values and hedonic prices for housing units, and vehicle availability preferences, we can evaluate the impacts of policies accounting for competing market forces.

We examine the effect of a car-lite policy in Singapore that seeks to increase accessibility inside a study region through the use of shared AVs or AMoD, or by encouraging TNC rides. The hypothesized outcome is an expected increase in vehicle-free households inside the study area owing to greater accessibility derived from alternative mobility options. However, place and spatial context both play a role in determining the effectiveness of this policy in achieving its desired outcome. Therefore, we consider multiple scenarios to test the effects of different market reactions to the policy, along with different study areas to account for the effect of the initial characteristics of the study area.

We notice a general positive trend in all study areas when the policy is instituted, albeit with different magnitudes. However, this net positive effect is found to suffer considerably with the consideration of market effects. We find that the car-lite policy is less likely to succeed in tight markets with low vacancy rates and high initial vehicle-free market shares, such as Toa Payoh. On the other hand, the likelihood of the policy succeeding increases when there is relatively more wiggle room on either front. In particular, Punggol & Sengkang seems to be the most successful study area (among the three options considered in this study) in achieving the desired outcome of the policy.

We also find unintended consequences of the policy in the form of gentrification, as evidenced by a rise in average household incomes. More importantly, the higher-income in-movers are more likely to bring their private vehicles along with them, thereby being less inclined to becoming vehicle-free despite enjoying increased accessibility benefits from the policy. Therefore, it is recommended that vehicle restrictions for in-movers, the nature of

the housing sub-market, and the number and type of new housing units inside the study area be considered during the design and implementation of such a car-lite policy in practice.

7.2 Limitations & future research

While this thesis provided insights into how modelers can start rethinking about the paradigm of vehicle ownership in a changing world with new mobility services, there is certainly room for improvement. First, we did not consider the endogeneity associated with housing and mobility choices. For example, households that prefer to use public transport might choose to live in neighborhoods that offer good access to public transport. This self-selection effect is unaccounted for in this thesis, as we deal only with the singular choice of mobility bundles. Second, the cost of vehicle ownership or availability is likely to be a strong influencer of the mobility bundle that is ultimately chosen. Since our primary data source is a traditional travel survey, we were unable to include this information as an independent variable in our behavioral models.

Third, the joint choice of a housing-mobility bundle is more likely to be influenced by wealth rather than income. For example, a retired couple might have negligible income but might have accumulated significant wealth over their lives, making them likely to be able to afford more expensive housing and mobility services. We are limited by our data sources yet again, which provide information about income only. While we tried to obtain an indirect estimate of wealth through the current residential unit, this can be improved with the consideration of a household budget constraint that accounts for their wealth and willingness-to-pay. Fourth, long-term household decisions such as residential location and vehicle availability choices can also be influenced by life stage events. For example, households are more likely to move to a bigger house or purchase a car after the birth of a child.

Most of these limitations can be addressed with the use of a mobility biography survey. We have already launched such a survey in Singapore that tracks several long-term decisions and life cycle events of households from 2016 to 2018, alongside obtaining attitudinal preferences towards mobility usage and more detailed information about private vehicle transactions. Using more complex modeling techniques can help us arrive at insights into the multi-dimensional long-term decision-making process. For example, the random param-

eters logit model can account for taste heterogeneity in the sample, while a latent class model can identify the differences in the direction of causality of long-term decisions among latent classes in the sample. A Markov Chain could also indicate how long-term decisions intermingle with life cycle events through a hierarchical model.

While these models would certainly serve as useful standalone contributions, using them in an ILUT microsimulation framework would enable us to examine the impacts of different policies on the population at a very disaggregate level. Hence, future research efforts should also be directed towards improving current model implementations in SimMobility using the models proposed in this thesis, and future models stemming from the mobility biography data. While microsimulation models such as SimMobility are needed to explore the complex behavioral consequences and market effects of alternative policies and regulations, the modeling framework needs to capture the key forces and interactions that drive market dynamics. Therefore, this is certainly an important and worthwhile topic to focus on in the future.

7.3 Concluding remarks

We are witnessing a new era of innovative technologies and services that will leave a lasting impression on the mobility landscape. Therefore, it is essential to reinterpret the traditional paradigm of vehicle ownership in the midst of access-based services. This thesis argues that new modeling strategies have to be adopted from different research areas to address such a complex problem. Through the use of three studies related to mobility choices, this thesis makes useful contributions to the methodological, conceptual, and praxis literatures. We hope that future studies will continue to build on our proposed methods and findings to adequately model long-term decisions.

Appendix A

Data preparation

A.1 Income imputation for HITS individuals

This section outlines the methodology for imputing incomes for individuals in the HITS 2012 sample whose income status is unknown. A similar procedure was adopted for the HITS 2008 sample and, thus, is omitted for brevity. The HITS 2012 sample has 12 income categories organized in an ordinal manner, ranging from '*No Income*' to '*$8,000 and above*'. In addition to these 12 categories, respondents are also offered the option to refuse to report their income status. Children below the age of 5 years are automatically excluded for further questioning at the beginning of the HITS survey and, thus, are assigned null values. The frequency shares of these 14 income categories are reported in Table A.1. I assigned zero income values to the 2,862 children whose income status is reported as '*null*' and the 15,623 individuals who have no income. Next, I combined the 12,110 individuals whose income status is known through the non-zero ordinal categories to form the training data set. Finally, the test data set was created using the 5,119 individuals who refused to report their income status.

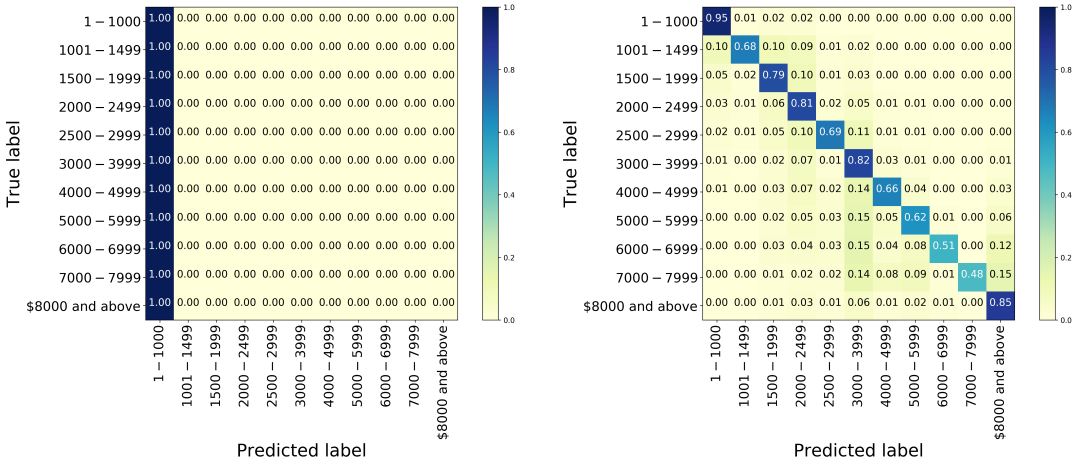
The goals of this exercise are to (a) predict the income category for the 5,119 individuals who refused to report their income status, and (b) impute continuous income values for all individuals in the HITS 2012 sample. The income category prediction is conducted through a random forest multiclass classification. The 12 non-zero ordinal categories were chosen as the class labels, while 19 individual-level features (representing age, citizenship status, gender, employment status, job occupation, job industry, and license ownership) and 19 household-level features (representing household size, ethnicity, housing type, and

Table A.1: Treatment of income categories

HITS category	Sample size	Treatment
Null values	2,862	Assign zero values
No Income	15,623	
\$1-\$1000	2,399	Training data set
\$1001-\$1499	979	
\$1500-\$1999	1,602	
\$2000-\$2499	1,796	
\$2500-\$2999	1,119	
\$3000-\$3999	1,564	
\$4000-\$4999	906	
\$5000-\$5999	649	
\$6000-\$6999	286	
\$7000-\$7999	172	
\$8000 and above	638	
Refused	5,119	Test data set

vehicle ownership) were combined to form a 38-dimensional feature vector. The random forest classifier achieved an overall accuracy score of 82.02% on the training data set after considering sampling weights. The baseline scenario is constructed using a zero-information approach that assumes the majority class for all samples, leading to a significantly lower test accuracy of 19.80%. Individual class predictions are also significantly better than the baseline predictions, as shown through the confusion matrices in Figure A-1.

After obtaining the income category for all individuals, I proceeded to create a log-normal distribution for individual incomes from which I will randomly draw values that fall within an individual's income category to assign continuous income values to all individuals. This is motivated by the frequency distribution as well as literature, both of which clearly show that individual income follows a log-normal distribution. The lower limits, upper limits, and mean values are calculated for each income category, following which a couple of modifications are made manually. First, the lower limit of the '*\$1-\$1000*' category is set to 50 in order to avoid implausibly low income draws. Second, the mean of the '*\$8000 and above*' category is set to 11,000 to limit unreasonably high income draws. The weighted mean and weighted standard deviation of the data are calculated using the equations below. Table A.2 shows that there are 12 income categories for which the mean value is treated as x_i and the sample size is treated as w_i in the following equations.



(a) Baseline (zero-R)

(b) Random Forest

Figure A-1: Confusion matrices for income category prediction on training data set

Table A.2: Income categories after imputation

Income category	Sample size	Lower limit	Upper limit	Mean value
0	18,485	0	0	0.0
$1-$1000	2,750	50	1,000	525.0
$1001-$1499	1,199	1,001	1,499	1,250.0
$1500-$1999	2,299	1,500	1,999	1,749.5
$2000-$2499	2,741	2,000	2,499	2,249.5
$2500-$2999	1,550	2,500	2,999	2,749.5
$3000-$3999	2,589	3,000	3,999	3,499.5
$4000-$4999	1,371	4,000	4,999	4,499.5
$5000-$5999	953	5,000	5,999	5,499.5
$6000-$6999	369	6,000	6,999	6,499.5
$7000-$7999	196	7,000	7,999	7,499.5
$8000 and above	1,212	8,000	50,000	11,000

$$\bar{x}^* = \frac{\sum_{i=1}^{12} w_i x_i}{\sum_{i=1}^{12} w_i} = 1,541.76 \tag{A.1}$$

$$\sigma^* = \sqrt{\frac{\sum_{i=1}^{12} w_i (x_i - \bar{x}^*)^2}{\sum_{i=1}^{12} w_i}} = 2,435.93 \tag{A.2}$$

The location (μ) and shape (Θ) parameters for the log-normal distribution are calculated using these values according to the following equations.

$$\mu = \log(\bar{x}^*) - \frac{1}{2} \log\left(1 + \frac{\sigma^{*2}}{\bar{x}^{*2}}\right) = 6.71 \tag{A.3}$$

$$\Theta = \sqrt{\log\left(1 + \frac{\sigma^{*2}}{\bar{x}^{*2}}\right)} = 1.12 \tag{A.4}$$

A log-normal distribution is then simulated according to the following equation. Recall that a random variable $\mathbf{X}$ is said to be log-normally distributed if its logarithm is normally distributed.

$$\mathcal{N}(\log \mathbf{X}; \mu, \Theta) = \frac{1}{\Theta\sqrt{2\pi}} \exp\left[-\frac{(\log \mathbf{X} - \mu)^2}{2\Theta^2}\right], \mathbf{X} > 0 \tag{A.5}$$

Income values are then randomly drawn from this simulated log-normal distribution and accepted as that individual's income if the value falls within the lower and upper limits of that individual's income category. Thus, continuous income values were obtained for each individual in the HITS 2012 sample (see Figure A-2).

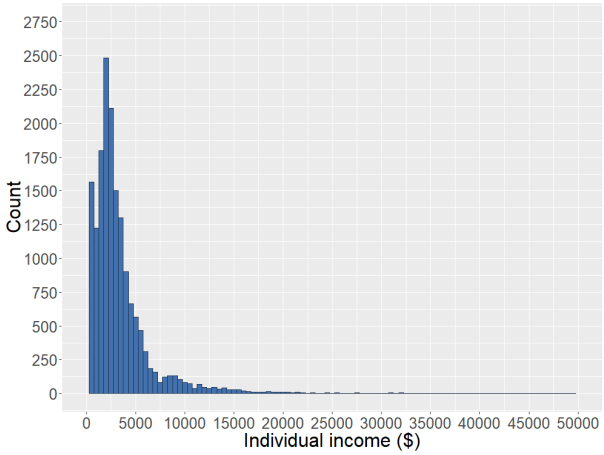


Figure A-2: Histogram of non-zero imputed individual incomes ($n = 17,229$)

A.2 Taxi imputation for HITS 2012 households

This section outlines the methodology for imputing taxi ownership among households in the HITS 2012 sample as taxi-owning households are underrepresented in the sample. I found that 166 households in the HITS 2012 sample (out of 9,635 households) own a single taxi and one household owns two taxis. After applying household sampling weights, the total number of taxis in the sample is 18,413. However, Comfort Company taxi data from 2012 indicates that there were 25,905 taxis in Singapore. Therefore, there is a deficit of 7,492 taxis in the weighted sample, which is equivalent to 61 taxis in the unweighted sample. It should also be noted that the ratio of the number of households with both car and taxi to

the number of households with only taxi is 0.14 in the weighted sample.

Based on sample households with access to taxi and personal information about taxi drivers, I constructed eligibility criteria to identify individuals that would be most likely to have access to taxis. First, individuals must hold a job with occupation as '*Plant & machine operator & assembler*' or '*Service & sales worker*'. Second, individuals must be employed in the '*Community, Social and Personal Services*' or '*Transport and Storage*' industry sector. Third, they must be 30 years of age or older. Finally, their income must not exceed $4,000. Using these criteria, I identified 1,001 individuals in the sample, out of which 831 did not have access to a taxi. These 831 individuals came from 767 distinct households, thereby simplifying the taxi imputation problem statement as follows: "*61 taxis need to be allocated among these 767 eligible households such that the weighted sample witnesses an increase of 7,492 taxis while the car-and-taxi household to only-taxi household ratio remains constant at 0.14*".

I used an iterative allocation procedure allowing for 1% error in both the additional taxi count (i.e., $7,492 \pm 1\%$) and the car-and-taxi household to only-taxi household ratio (i.e., $0.14 \pm 1\%$). Finally, the imputed HITS 2012 sample contained 231 households with access to taxi, among which one household had two taxis. After applying household weights, the number of taxis is the population is 25,893 and the car-and-taxi household to only-taxi household ratio is 0.14.

THIS PAGE INTENTIONALLY LEFT BLANK

Appendix B

Model performance during estimation

This section includes tables reporting the goodness-of-fit, prediction accuracy and predicted market shares for the best econometric (*multinomial logit* - MNL) and best machine learning (*random forest* - RF and *neural network* - NN) models in the estimation scenario. In this sccenario, the model is trained on a randomly selected 80% sample of the 2008 data set, and then tested on the remaining 20% of the 2008 data set. Tables B.1 and B.2 compare model assessment metrics between the standard and econometric variable specifications. Recall that the econometric specification involved transforming relevant features into economically consistent functions of the dependent variable. Tables B.3 and B.4 compare model assessment metrics between the econometric and the ML-enhanced specifications. Recall that the ML-enhanced specification builds on the econometric specification by accounting for non-linearities in the data set that are caused by correlations between the independent variables.

Table B.1: Gof and prediction accuracy with econometric specification

Metric	**Baseline**	**MNL**[a]	**RF**[a]	**NN**[a]	**MNL**[b]	**RF**[b]	**NN**[b]
			Estimation Scenario [c]				
Avg. train accuracy	0.55	0.74	1.00	0.76	0.75	1.00	0.78
Avg. test accuracy	0.55	0.73	0.73	0.73	0.74	0.73	0.74
Execution time (sec)	-	16.0	5.77	14.61	29.0	0.69	3.61
Avg. precision	0.30	0.71	0.72	0.69	0.73	0.72	0.74
Avg. recall	0.55	0.73	0.73	0.73	0.74	0.74	0.74
Avg. F-measure	0.39	0.71	0.71	0.70	0.72	0.71	0.72
McFadden's $\bar{\rho}^2$	-	0.632	-	-	0.644	-	-

[a] Standard Specification; [b] Econometric Specification

[c] Using HITS 2008 data with 80/20 train/test split

Table B.2: Predicted market shares with econometric specification

Category	**Actual**	**Baseline**	**MNL**[a]	**RF**[a]	**NN**[a]	**MNL**[b]	**RF**[b]	**NN**[b]
			Estimation Scenario [c]					
0	54.86	100	59.15	56.34	52.46	51.97	56.59	51.67
1	5.37	0	6.07	2.97	6.35	8.36	3.29	8.06
2	1.67	0	0.05	0	0	0	0	0
3	32.28	0	33.73	40.13	41.19	38.66	39.50	39.70
4	1.44	0	0.12	0	0	0.41	0	0
5	4.38	0	0.88	0.56	0	0.61	0.62	0.57
MAE	-	15.05	2.15	3.11	3.30	3.12	2.98	3.37
RMSE	-	22.85	2.50	3.85	4.28	3.56	3.62	3.91

[a] Standard Specification; [b] Econometric Specification

[c] Using HITS 2008 data with 80/20 train/test split

Table B.3: Gof and prediction accuracy with ML-enhanced specification

Metric	Baseline	MNL[a]	RF[a]	NN[a]	MNL[b]	RF[b]	NN[b]
			Estimation Scenario [c]				
Avg. train accuracy	0.55	0.75	1.00	0.78	0.75	1.00	0.76
Avg. test accuracy	0.55	0.74	0.73	0.74	0.73	0.74	0.74
Execution time (sec)	-	29.0	0.69	3.61	27.0	6.63	11.97
Avg. precision	0.30	0.73	0.72	0.74	0.71	0.73	0.71
Avg. recall	0.55	0.74	0.74	0.74	0.73	0.74	0.74
Avg. F-measure	0.39	0.72	0.71	0.72	0.71	0.72	0.72
McFadden's $\bar{\rho}^2$	-	0.644	-	-	0.618	-	-

[a] Econometric specification; [b] ML-enhanced specification

[c] Using HITS 2008 data with 80/20 train/test split

Table B.4: Predicted market shares with ML-enhanced specification

Category	Actual	Baseline	MNL[a]	RF[a]	NN[a]	MNL[b]	RF[b]	NN[b]
				Estimation Scenario [c]				
0	54.86	100	51.97	56.59	51.67	51.59	54.34	52.87
1	5.37	0	8.36	3.29	8.06	8.17	4.13	8.38
2	1.67	0	0	0	0	0	0	0
3	32.28	0	38.66	39.50	39.70	39.65	40.93	38.51
4	1.44	0	0.41	0	0	0.15	0	0
5	4.38	0	0.61	0.62	0.57	0.44	0.59	0.24
MAE	-	15.05	3.12	2.98	3.37	3.39	2.89	3.08
RMSE	-	22.85	3.56	3.62	3.91	3.93	4.00	3.51

[a] Econometric specification; [b] ML-enhanced specification

[c] Using HITS 2008 data with 80/20 train/test split

THIS PAGE INTENTIONALLY LEFT BLANK

References

Adnan, M., F. C. Pereira, C. M. L. Azevedo, K. Basak, M. Lovric, S. Raveau, Y. Zhu, J. Ferreira, C. Zegras, and M. Ben-Akiva (2016). Simmobility: A multi-scale integrated agent-based simulation platform. In *95th Annual Meeting of the Transportation Research Board*.

Advani, M. and G. Tiwari (2005). Evaluation of public transport systems: case study of delhi metro. *Transportation Research & Injury Prevention Programme*, 1.

Alías, F., J. Socoró, and X. Sevillano (2016). A review of physical and perceptual feature extraction techniques for speech, music and environmental sounds. *Applied Sciences 6*(5), 143.

Allahyari, M., S. Pouriyeh, M. Assefi, S. Safaei, E. D. Trippe, J. B. Gutierrez, and K. Kochut (2017). A brief survey of text mining: Classification, clustering and extraction techniques. *arXiv preprint arXiv:1707.02919*.

Anowar, S., N. Eluru, and L. F. Miranda-Moreno (2016). Analysis of vehicle ownership evolution in montreal, canada using pseudo panel analysis. *Transportation 43*(3), 531–548.

Anyanwu, M. N. and S. G. Shiva (2009). Comparative analysis of serial decision tree classification algorithms. *International Journal of Computer Science and Security 3*(3), 230–240.

Ao, Y., C. Chen, D. Yang, and Y. Wang (2018). Relationship between rural built environment and household vehicle ownership: An empirical analysis in rural sichuan, china. *Sustainability 10*(5), 1566.

Apley, D. W. (2016). Visualizing the effects of predictor variables in black box supervised learning models. *arXiv preprint arXiv:1612.08468*.

Ashqar, H. I., M. H. Almannaa, M. Elhenawy, H. A. Rakha, and L. House (2018). Smartphone transportation mode recognition using a hierarchical machine learning classifier and pooled features from time and frequency domains. *IEEE Transactions on Intelligent Transportation Systems* (99), 1–9.

Athey, S. (2018). The impact of machine learning on economics. In *The Economics of Artificial Intelligence: An Agenda*. University of Chicago Press.

Athey, S. and G. Imbens (2019). Machine learning methods economists should know about. *arXiv preprint arXiv:1903.10075*.

Athey, S. and G. W. Imbens (2017). The state of applied econometrics: Causality and policy evaluation. *Journal of Economic Perspectives 31*(2), 3–32.

Azevedo, C. L., N. M. Deshmukh, B. Marimuthu, S. Oh, K. Marczuk, H. Soh, K. Basak, T. Toledo, L.-S. Peh, and M. E. Ben-Akiva (2017). Simmobility short-term: An integrated microscopic mobility simulator. *Transportation Research Record 2622*(1), 13–23.

Bacciu, D., A. Carta, S. Gnesi, and L. Semini (2017). An experience in using machine learning for short-term predictions in smart transportation systems. *Journal of Logical and Algebraic Methods in Programming 87*, 52–66.

Bajari, P., D. Nekipelov, S. P. Ryan, and M. Yang (2015). Demand estimation with machine learning and model combination. Technical report, National Bureau of Economic Research.

Bansal, P., K. M. Kockelman, W. Schievelbein, and S. Schauer-West (2018). Indian vehicle ownership and travel behavior: A case study of bengaluru, delhi and kolkata. *Research in Transportation Economics 71*, 2–8.

Baptista, P., S. Melo, and C. Rolim (2014). Energy, environmental and mobility impacts of car-sharing systems. empirical results from lisbon, portugal. *Procedia-Social and Behavioral Sciences 111*, 28–37.

Barua, S., M. M. Islam, and K. Murase (2013). ProWSyn: Proximity weighted synthetic oversampling technique for imbalanced data set learning. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 317–328. Springer.

Bastian, A., M. Börjesson, and J. Eliasson (2016). Explaining "peak car" with economic variables. *Transportation Research Part A: Policy and Practice 88*, 236–250.

Basu, R., A. Araldo, A. P. Akkinepally, B. H. N. Biran, K. Basak, R. Seshadri, N. Deshmukh, N. Kumar, C. L. Azevedo, and M. Ben-Akiva (2018). Automated mobility-on-demand vs. mass transit: A multi-modal activity-driven agent-based simulation approach. *Transportation Research Record*.

Basu, R. and J. Ferreira (2019). Understanding household vehicle ownership in singapore through a comparison of econometric and machine learning models. *Transportation Research Procedia*, forthcoming.

Basu, R., A. Khatua, S. Ghosh, and A. Jana (2017). Harnessing twitter data for analyzing public reactions to transportation policies: Evidences from the odd-even policy in delhi, india. *Proceedings of the Eastern Asia Society For Transportation Studies (EASTS)*.

Batista, G. E., A. L. Bazzan, and M. C. Monard (2003). Balancing training data for automated annotation of keywords: a case study. In *WOB*, pp. 10–18.

Baum, C. L. (2009). The effects of vehicle ownership on employment. *Journal of Urban Economics 66*(3), 151–163.

Becker, H., M. Balać, F. Ciari, and K. W. Axhausen (2018). Assessing the welfare impacts of shared mobility and mobility as a service (maas). *Arbeitsberichte Verkehrs-und Raumplanung 1378*.

Becker, H., F. Ciari, and K. W. Axhausen (2017). Comparing car-sharing schemes in switzerland: User groups and usage patterns. *Transportation Research Part A: Policy and Practice 97*, 17–29.

Becker, H., F. Ciari, and K. W. Axhausen (2018). Measuring the car ownership impact of free-floating car-sharing–a case study in basel, switzerland. *Transportation Research Part D: Transport and Environment 65*, 51–62.

Beige, S. and K. W. Axhausen (2012). Interdependencies between turning points in life and long-term mobility decisions. *Transportation 39*(4), 857–872.

Beyan, C. and R. Fisher (2015). Classifying imbalanced data sets using similarity based hierarchical decomposition. *Pattern Recognition 48*(5), 1653–1672.

Bhat, C. R. (2005). A multiple discrete–continuous extreme value model: formulation and application to discretionary time-use decisions. *Transportation Research Part B: Methodological 39*(8), 679–707.

Bhat, C. R. and V. Pulugurta (1998). A comparison of two alternative behavioral choice mechanisms for household auto ownership decisions. *Transportation Research Part B: Methodological 32*(1), 61–75.

Bhat, C. R. and S. Sen (2006). Household vehicle type holdings and usage: an application of the multiple discrete-continuous extreme value (mdcev) model. *Transportation Research Part B: Methodological 40*(1), 35–53.

Bhat, C. R., S. Sen, and N. Eluru (2009). The impact of demographics, built environment attributes, vehicle characteristics, and gasoline prices on household vehicle holdings and use. *Transportation Research Part B: Methodological 43*(1), 1–18.

Blumenberg, E., A. Brown, and A. Schouten (2018). Car-deficit households: determinants and implications for household travel in the us. *Transportation*, 1–23.

Bösch, P. M., F. Becker, H. Becker, and K. W. Axhausen (2018). Cost-based analysis of autonomous mobility services. *Transport Policy 64*, 76–91.

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

Breiman, L. (2001). Random forests. *Machine learning 45*(1), 5–32.

Brown, A. E. (2017). Car-less or car-free? socioeconomic and mobility differences among zero-car households. *Transport Policy 60*, 152–159.

Brownstone, D. and H. Fang (2014). A vehicle ownership and utilization choice model with endogenous residential density. *Journal of Transport and Land Use 7*(2), 135–151.

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

Caulfield, B. (2012). An examination of the factors that impact upon multiple vehicle ownership: The case of dublin, ireland. *Transport Policy 19*(1), 132–138.

Cervero, R. and Y. Tsai (2004). City carshare in san francisco, california: second-year travel demand and car ownership impacts. *Transportation Research Record 1887*(1), 117–127.

Chawla, N. V., K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research 16*, 321–357.

Chen, X. M., M. Zahiri, and S. Zhang (2017). Understanding ridesplitting behavior of on-demand ride services: An ensemble learning approach. *Transportation Research Part C: Emerging Technologies 76*, 51–70.

Chiou, Y.-C., C.-H. Wen, S.-H. Tsai, and W.-Y. Wang (2009). Integrated modeling of car/motorcycle ownership, type and usage for estimating energy consumption and emissions. *Transportation Research Part A: Policy and Practice 43*(7), 665–684.

Choudhary, R. and V. Vasudevan (2017). Study of vehicle ownership for urban and rural households in india. *Journal of Transport Geography 58*, 52–58.

Clark, B., K. Chatterjee, and S. Melia (2016). Changes in level of household car ownership: the role of life events and spatial context. *Transportation 43*(4), 565–599.

Clark, B., G. Lyons, and K. Chatterjee (2016). Understanding the process that gives rise to household car ownership level changes. *Journal of Transport Geography 55*, 110–120.

Clark, S. D. (2009). The determinants of car ownership in england and wales from anonymous 2001 census data. *Transportation research part C: emerging technologies 17*(5), 526–540.

Clark, S. D. and S. Rey (2017). Temporal dynamics in local vehicle ownership for great britain. *Journal of Transport Geography 62*, 30–37.

Clewlow, R. R. (2016). Carsharing and sustainable travel behavior: Results from the san francisco bay area. *Transport Policy 51*, 158–164.

Combs, T. S. and D. A. Rodríguez (2014). Joint impacts of bus rapid transit and urban form on vehicle ownership: New evidence from a quasi-longitudinal analysis in bogotá, colombia. *Transportation Research Part A: Policy and Practice 69*, 272–285.

Curl, A., J. Clark, and A. Kearns (2018). Household car adoption and financial distress in deprived urban communities: A case of forced car ownership? *Transport Policy 65*, 61–71.

Dabiri, S. and K. Heaslip (2018). Inferring transportation modes from gps trajectories using a convolutional neural network. *Transportation research part C: emerging technologies 86*, 360–371.

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

Delbosc, A. (2017). Delay or forgo? a closer look at youth driver licensing trends in the united states and australia. *Transportation 44*(5), 919–926.

Delbosc, A. and G. Currie (2012). Choice and disadvantage in low-car ownership households. *Transport Policy 23*, 8–14.

Dias, F. F., P. S. Lavieri, V. M. Garikapati, S. Astroza, R. M. Pendyala, and C. R. Bhat (2017). A behavioral choice model of the use of car-sharing and ride-sourcing services. *Transportation 44*(6), 1307–1323.

Díez-Pastor, J. F., J. J. Rodríguez, C. I. García-Osorio, and L. I. Kuncheva (2015). Diversity techniques improve the performance of the best imbalance learning ensembles. *Information Sciences 325*, 98–117.

Ding, C., Y. Wang, T. Tang, S. Mishra, and C. Liu (2018). Joint analysis of the spatial impacts of built environment on car ownership and travel mode choice. *Transportation research part D: transport and environment 60*, 28–40.

Engel-Yan, J. and D. Passmore (2013). Carsharing and car ownership at the building scale: Examining the potential for flexible parking requirements. *Journal of the American Planning Association 79*(1), 82–91.

Fang, S.-H., Y.-X. Fei, Z. Xu, and Y. Tsao (2017). Learning transportation modes from smartphone sensors based on deep neural network. *IEEE Sensors Journal 17*(18), 6111–6118.

Firnkorn, J. and M. Müller (2011). What will be the environmental effects of new free-floating car-sharing systems? the case of car2go in ulm. *Ecological Economics 70*(8), 1519–1528.

Firnkorn, J. and M. Müller (2012). Selling mobility instead of cars: new business strategies of automakers and the impact on private vehicle holding. *Business Strategy and the environment 21*(4), 264–280.

Fisher, A., C. Rudin, and F. Dominici (2018). Model Class Reliance: Variable importance measures for any machine learning model class, from the" Rashomon" perspective. *arXiv preprint arXiv:1801.01489*.

Gal, A., A. Mandelbaum, F. Schnitzler, A. Senderovich, and M. Weidlich (2017). Traveling time prediction in scheduled transportation with journey segments. *Information Systems 64*, 266–280.

Gao, X. and G. M. Lee (2019). Moment-based rental prediction for bicycle-sharing transportation systems using a hybrid genetic algorithm and machine learning. *Computers & Industrial Engineering 128*, 60–69.

Gazzah, S. and N. E. B. Amara (2008). New oversampling approaches based on polynomial fitting for imbalanced data sets. In *2008 The Eighth IAPR International Workshop on Document Analysis Systems*, pp. 677–684. IEEE.

Giblin, S. and A. McNabola (2009). Modelling the impacts of a carbon emission-differentiated vehicle tax system on co2 emissions intensity from new vehicle purchases in ireland. *Energy Policy 37*(4), 1404–1411.

Giesel, F. and C. Nobis (2016). The impact of carsharing on car ownership in german cities. *Transportation Research Procedia 19*, 215–224.

Glaeser, E. L. and M. E. Kahn (2004). Sprawl and urban growth. In *Handbook of regional and urban economics*, Volume 4, pp. 2481–2527. Elsevier.

Greene, W. H. and D. A. Hensher (2010). *Modeling ordered choices: A primer*. Cambridge University Press.

Gucwa, M. (2014). Mobility and energy impacts of automated cars. In *Proceedings of the Automated Vehicles Symposium, San Francisco*.

Guerra, E. (2015). The geography of car ownership in mexico city: a joint model of households' residential location and car ownership decisions. *Journal of Transport Geography 43*, 171–180.

Guyon, I. and A. Elisseeff (2003). An introduction to variable and feature selection. *Journal of machine learning research 3*(Mar), 1157–1182.

Haboucha, C. J., R. Ishaq, and Y. Shiftan (2017). User preferences regarding autonomous vehicles. *Transportation Research Part C: Emerging Technologies 78*, 37–49.

Hagenauer, J. and M. Helbich (2017). A comparative study of machine learning classifiers for modeling travel mode choice. *Expert Systems with Applications 78*, 273–282.

Haixiang, G., L. Yijing, J. Shang, G. Mingyun, H. Yuanyue, and G. Bing (2017). Learning from class-imbalanced data: Review of methods and applications. *Expert Systems with Applications 73*, 220–239.

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

Haykin, S. S., S. S. Haykin, S. S. Haykin, K. Elektroingenieur, and S. S. Haykin (2009). *Neural networks and learning machines*, Volume 3. Pearson education Upper Saddle River.

He, H., Y. Bai, E. A. Garcia, and S. Li (2008). ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, pp. 1322–1328. IEEE.

He, H., R. Ponce-Lopez, J. Shaw, D.-T. Le, J. Ferreira, and P. C. Zegras (2019). Representing accessibility: Evidence from vehicle ownership choices and property valuations in singapore. *Transportation Research Record*, 0361198119825831.

Hess, D. B. and T. M. Almeida (2007). Impact of proximity to light rail rapid transit on station-area property values in buffalo, new york. *Urban studies 44*(5-6), 1041–1068.

Hörl, S., F. Ciari, and K. W. Axhausen (2016). Recent perspectives on the impact of autonomous vehicles. *Arbeitsberichte Verkehrs-und Raumplanung 1216*.

Huang, X., X. J. Cao, J. Yin, and X. Cao (2017). Effects of metro transit on the ownership of mobility instruments in xi'an, china. *Transportation Research Part D: Transport and Environment 52*, 495–505.

Jahangiri, A. and H. A. Rakha (2015). Applying machine learning techniques to transportation mode recognition using mobile phone sensor data. *IEEE transactions on intelligent transportation systems 16*(5), 2406–2417.

Jiang, Y., P. Gu, Y. Chen, D. He, and Q. Mao (2017). Influence of land use and street characteristics on car ownership and use: Evidence from jinan, china. *Transportation Research Part D: Transport and Environment 52*, 518–534.

Jiang, Y., J. Zhang, Y. Wang, and W. Wang (2018). Capturing ownership behavior of autonomous vehicles in japan based on a stated preference survey and a mixed logit model with repeated choices. *International Journal of Sustainable Transportation*, 1–14.

Jong, G. D., J. Fox, A. Daly, M. Pieters, and R. Smit (2004). Comparison of car ownership models. *Transport Reviews 24*(4), 379–408.

Kahneman, D., J. L. Knetsch, and R. H. Thaler (1990). Experimental tests of the endowment effect and the coase theorem. *Journal of Political Economy 98*(6), 1325–1348.

Katzev, R. (2003). Car sharing: A new approach to urban transportation problems. *Analyses of social issues and public policy 3*(1), 65–86.

Kauffman, R. J., K. Kim, S.-Y. T. Lee, A.-P. Hoang, and J. Ren (2017). Combining machine-based and econometrics methods for policy analytics insights. *Electronic Commerce Research and Applications 25*, 115–140.

Klein, N. J. and M. J. Smart (2017). Millennials and car ownership: Less money, fewer cars. *Transport Policy 53*, 20–29.

Klincevicius, M. G., C. Morency, and M. Trépanier (2014). Assessing impact of carsharing on household car ownership in montreal, quebec, canada. *Transportation Research Record 2416*(1), 48–55.

Knittel, C. R. and E. Murphy (2019). Generational trends in vehicle ownership and use: Are millennials any different? Technical report, National Bureau of Economic Research.

Kotsiantis, S. B., I. Zaharakis, and P. Pintelas (2007). Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering 160*, 3–24.

Kuhnimhof, T., J. Armoogum, R. Buehler, J. Dargay, J. M. Denstadli, and T. Yamamoto (2012). Men shape a downward trend in car use among young adults—evidence from six industrialized countries. *Transport Reviews 32*(6), 761–779.

Kuhnimhof, T., D. Zumkeller, and B. Chlond (2013). Who made peak car, and how? a breakdown of trends over four decades in four countries. *Transport Reviews 33*(3), 325–342.

Lagrell, E., E. Thulin, and B. Vilhelmson (2018). Accessibility strategies beyond the private car: A study of voluntarily carless families with young children in gothenburg. *Journal of Transport Geography 72*, 218–227.

Lane, C. (2005). Phillycarshare: First-year social and mobility impacts of carsharing in philadelphia, pennsylvania. *Transportation Research Record 1927*(1), 158–166.

Lane, P. C., D. Clarke, and P. Hender (2012). On developing robust models for favourability analysis: Model choice, feature sets and imbalanced data. *Decision Support Systems 53*(4), 712–718.

Lavieri, P. S., V. M. Garikapati, C. R. Bhat, and R. M. Pendyala (2017). Investigation of heterogeneity in vehicle ownership and usage for the millennial generation. *Transportation Research Record 2664*(1), 91–99.

Le, D.-T., G. Cernicchiaro, C. Zegras, and J. Ferreira Jr (2016). Constructing a synthetic population of establishments for the simmobility microsimulation platform. *Transportation Research Procedia 19*, 81–93.

Le Vine, S. and J. Polak (2017). The impact of free-floating carsharing on car ownership: Early-stage findings from london. *Transport Policy*.

Le Vine, S., C. Wu, and J. Polak (2018). A nationwide study of factors associated with household car ownership in china. *IATSS Research 42*(3), 128–137.

Lee, K.-I., K.-J. Kim, and S.-J. Kwon (2005). A study on characteristics of subway utilization and pedestrians' accessibility at new towns in korea. *Journal of Asian Architecture and Building Engineering 4*(1), 85–95.

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

Li, T., J. Dodson, and N. Sipe (2018). Examining household relocation pressures from rising transport and housing costs–an australian case study. *Transport Policy 65*, 106–113.

Liao, F., E. Molin, H. Timmermans, and B. van Wee (2018). Carsharing: the impact of system characteristics on its potential to replace private car trips and reduce car ownership. *Transportation*, 1–36.

Lima, R. F. and A. C. M. Pereira (2015). A fraud detection model based on feature selection and undersampling applied to web payment systems. In *2015 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*, Volume 3, pp. 219–222. IEEE.

Liu, Y., J.-M. Tremblay, and C. Cirillo (2014). An integrated model for discrete and continuous decisions with application to vehicle ownership, type and usage choices. *Transportation Research Part A: Policy and Practice 69*, 315–328.

Loose, W. (2010). The state of european car-sharing. *Project Momo Final Report D 2*.

Louvet, N. (2014). One-way carsharing: which alternative to private cars. *The case study of Autolib'in Paris*.

Loyola-González, O., J. F. Martínez-Trinidad, J. A. Carrasco-Ochoa, and M. García-Borroto (2016). Study of the impact of resampling methods for contrast pattern based classifiers in imbalanced databases. *Neurocomputing 175*, 935–947.

LTA (2013). Walk2Ride Programme.

Lu, Y., M. Adnan, K. Basak, F. C. Pereira, C. Carrion, V. H. Saber, H. Loganathan, and M. E. Ben-Akiva (2015). Simmobility mid-term simulator: A state of the art integrated agent based demand and supply model. In *94th Annual Meeting of the Transportation Research Board, Washington, DC*.

Luke, R. (2018). Car ownership perceptions and intentions amongst south african students. *Journal of Transport Geography 66*, 135–143.

Lv, Y., Y. Duan, W. Kang, Z. Li, and F.-Y. Wang (2015). Traffic flow prediction with big data: a deep learning approach. *IEEE Transactions on Intelligent Transportation Systems 16*(2), 865–873.

Ma, X., H. Yu, Y. Wang, and Y. Wang (2015). Large-scale transportation network congestion evolution prediction using deep learning theory. *PloS one 10*(3), e0119044.

Macfarlane, G. S., L. A. Garrow, and P. L. Mokhtarian (2015). The influences of past and present residential locations on vehicle ownership decisions. *Transportation research part A: policy and practice 74*, 186–200.

Maltha, Y., M. Kroesen, B. Van Wee, and E. van Daalen (2017). Changing influence of factors explaining household car ownership levels in the netherlands. *Transportation Research Record 2666*(1), 103–111.

Manski, C. F. (1977). The structure of random utility models. *Theory and decision 8*(3), 229–254.

Manville, M. (2017). Bundled parking and vehicle ownership: Evidence from the american housing survey. *Journal of Transport and Land Use 10*(1), 27–55.

Martin, E. and S. Shaheen (2016). Impacts of car2go on vehicle ownership, modal shift, vehicle miles traveled, and greenhouse gas emissions: an analysis of five north american cities. *Transportation Sustainability Research Center, UC Berkeley 3*.

Martin, E., S. A. Shaheen, and J. Lidicker (2010). Impact of carsharing on household vehicle holdings: Results from north american shared-use vehicle survey. *Transportation Research Record 2143*(1), 150–158.

McBride, E., J. H. Lee, A. M. Lundberg, A. W. Davis, and K. G. Goulias (2016). Behavioural micro-dynamics of car ownership and travel in the seattle metropolitan region from 1989 to 2002. *European Journal of Transport and Infrastructure Research 16*(4).

McDonald, N. C. (2015). Are millennials really the "go-nowhere" generation? *Journal of the American Planning Association 81*(2), 90–103.

McFadden, D. (1973). Conditional logit analysis of qualitative choice behavior.

McFadden, D. (1977). Quantitative methods for analysing travel behavior of individuals: Some recent developments.

Menon, N., N. Barbour, Y. Zhang, A. R. Pinjari, and F. Mannering (2019). Shared autonomous vehicles and their potential impacts on household vehicle ownership: An exploratory empirical assessment. *International Journal of Sustainable Transportation 13*(2), 111–122.

Meyer, J., H. Becker, P. M. Bösch, and K. W. Axhausen (2017). Autonomous vehicles: The next jump in accessibilities? *Research in Transportation Economics 62*, 80–91.

Milakis, D., M. Kroesen, and B. van Wee (2018). Implications of automated vehicles for accessibility and location choices: Evidence from an expert-based experiment. *Journal of Transport Geography 68*, 142–148.

Millard-Ball, A. (2005). *Car-sharing: Where and how it succeeds*, Volume 108. Transportation Research Board.

Millard-Ball, A. and L. Schipper (2011). Are we reaching peak travel? trends in passenger transport in eight industrialized countries. *Transport reviews 31*(3), 357–378.

Mishra, G. S., R. R. Clewlow, P. L. Mokhtarian, and K. F. Widaman (2015). The effect of carsharing on vehicle holdings and travel behavior: A propensity score and causal mediation analysis of the san francisco bay area. *Research in Transportation Economics 52*, 46–55.

MoE (2019). Allocation of Places in Primary Schools.

Mullainathan, S. and J. Spiess (2017). Machine learning: an applied econometric approach. *Journal of Economic Perspectives 31*(2), 87–106.

Namazu, M. and H. Dowlatabadi (2018). Vehicle ownership reduction: A comparison of one-way and two-way carsharing systems. *Transport Policy 64*, 38–50.

Nolan, A. (2010). A dynamic analysis of household car ownership. *Transportation research part A: policy and practice 44*(6), 446–455.

Oakil, A. T. M., D. Ettema, T. Arentze, and H. Timmermans (2014). Changing household car ownership level and life cycle events: an action in anticipation or an action on occurrence. *Transportation 41*(4), 889–904.

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

Pakusch, C., G. Stevens, A. Boden, and P. Bossauer (2018). Unintended effects of autonomous driving: A study on mobility preferences in the future. *Sustainability 10*(7), 2404.

Paleti, R., C. R. Bhat, and R. M. Pendyala (2013). Integrated model of residential location, work location, vehicle ownership, and commute tour characteristics. *Transportation Research Record 2382*(1), 162–172.

Paredes, M., E. Hemberg, U.-M. O'Reilly, and C. Zegras (2017). Machine learning or discrete choice models for car ownership demand estimation and prediction? In *2017 5th IEEE International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS)*, pp. 780–785. IEEE.

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

Potoglou, D. and Y. O. Susilo (2008). Comparison of vehicle-ownership models. *Transportation Research Record 2076*(1), 97–105.

Rahim Taleqani, A., J. Hough, and K. E. Nygard (2019). Public opinion on dockless bike sharing: A machine learning approach. *Transportation Research Record*, 0361198119838982.

Ritter, N. and C. Vance (2013). Do fewer people mean fewer cars? population decline and car ownership in germany. *Transportation Research Part A: Policy and Practice 50*, 74–85.

Sáez, J. A., J. Luengo, J. Stefanowski, and F. Herrera (2015). SMOTE-IPF: Addressing the noisy and borderline examples problem in imbalanced classification by a re-sampling method with filtering. *Information Sciences 291*, 184–203.

Salon, D. (2009). Neighborhoods, cars, and commuting in new york city: A discrete choice approach. *Transportation Research Part A: Policy and Practice 43*(2), 180–196.

Schaefers, T., S. J. Lawson, and M. Kukar-Kinney (2016). How the burdens of ownership promote consumer usage of access-based services. *Marketing Letters 27*(3), 569–577.

Shalev-Shwartz, S., Y. Singer, N. Srebro, and A. Cotter (2011). Pegasos: Primal estimated sub-gradient solver for svm. *Mathematical programming 127*(1), 3–30.

Shen, Q., P. Chen, and H. Pan (2016). Factors affecting car ownership and mode choice in rail transit-supported suburbs of a large chinese city. *Transportation Research Part A: Policy and Practice 94*, 31–44.

Sifringer, B., V. Lurkin, and A. Alahi (2018). Let me not lie: Learning multinomial logit. *arXiv preprint arXiv:1812.09747*.

Soltani, A. (2017). Social and urban form determinants of vehicle ownership; evidence from a developing country. *Transportation Research Part A: Policy and Practice 96*, 90–100.

Soteropoulos, A., M. Berger, and F. Ciari (2019). Impacts of automated vehicles on travel behaviour and land use: an international review of modelling studies. *Transport reviews 39*(1), 29–49.

Spieser, K., K. Treleaven, R. Zhang, E. Frazzoli, D. Morton, and M. Pavone (2014). Toward a systematic approach to the design and evaluation of automated mobility-on-demand systems: A case study in singapore. In *Road vehicle automation*, pp. 229–245. Springer.

Stasko, T. H., A. B. Buck, and H. O. Gao (2013). Carsharing in a university setting: Impacts on vehicle ownership, parking demand, and mobility in ithaca, ny. *Transport Policy 30*, 262–268.

Tahir, M. A., J. Kittler, K. Mikolajczyk, and F. Yan (2009). A multiple expert approach to the class imbalance problem using inverse random under sampling. In *International Workshop on Multiple Classifier Systems*, pp. 82–91. Springer.

Thakur, P., R. Kinghorn, and R. Grace (2016). Urban form and function in the autonomous era. In *38th Australasian Transport Research Forum (ATRF), 2016, Melbourne, Victoria, Australia*.

Tomek, I. (1976). Two modifications of CNN. *IEEE Trans. Systems, Man and Cybernetics 6*, 769–772.

Tran, M. T., J. Zhang, M. Chikaraishi, and A. Fujiwara (2016). A joint analysis of residential location, work location and commuting mode choices in hanoi, vietnam. *Journal of Transport Geography 54*, 181–193.

van Eggermond, M. A., A. Erath, and K. W. Axhausen (2016). Vehicle ownership and usage in switzerland: Role of micro-and macroaccessibility. In *2016 TRB Annual Meeting: Compendium of Papers*, pp. 16–4761. The National Academies of Sciences, Engineering, and Medicine.

Wadud, Z. (2017). Fully automated vehicles: A cost of ownership analysis to inform early adoption. *Transportation Research Part A: Policy and Practice 101*, 163–176.

Walks, A. (2018). Driving the poor into debt? automobile loans, transport disadvantage, and automobile dependence. *Transport policy 65*, 137–149.

Wang, J., R. Chen, and Z. He (2019). Traffic speed prediction for urban transportation network: A path based deep learning approach. *Transportation Research Part C: Emerging Technologies 100*, 372–385.

Wang, S. and J. Zhao (2018). Framing discrete choice model as deep neural network with utility interpretation. *arXiv preprint arXiv:1810.10465*.

Wasikowski, M. and X.-w. Chen (2010). Combating the small sample class imbalance problem using feature selection. *IEEE Transactions on knowledge and data engineering 22*(10), 1388–1400.

Wells, P. and D. Xenias (2015). From 'freedom of the open road'to 'cocooning': Understanding resistance to change in personal private automobility. *Environmental Innovation and Societal Transitions 16*, 106–119.

Yu, H., Z. Wu, S. Wang, Y. Wang, and X. Ma (2017). Spatiotemporal recurrent convolutional networks for traffic prediction in transportation networks. *Sensors 17*(7), 1501.

Zakharenko, R. (2016). Self-driving cars will change cities. *Regional Science and Urban Economics 61*, 26–37.

Zegras, C. (2010). The built environment and motor vehicle ownership and use: Evidence from santiago de chile. *Urban Studies 47*(8), 1793–1817.

Zhang, J. (2017). *Life-oriented behavioral research for urban policy*. Springer.

Zhang, J., M. Kuwano, B. Lee, and A. Fujiwara (2009). Modeling household discrete choice behavior incorporating heterogeneous group decision-making mechanisms. *Transportation Research Part B: Methodological 43*(2), 230–250.

Zhang, J., B. Yu, and M. Chikaraishi (2014). Interdependences between household residential and car ownership behavior: a life history analysis. *Journal of Transport Geography 34*, 165–174.

Zhang, W. and S. Guhathakurta (2018). Residential location choice in the era of shared autonomous vehicles. *Journal of Planning Education and Research*, 0739456X18776062.

Zhang, W., S. Guhathakurta, and E. B. Khalil (2018). The impact of private autonomous vehicles on vehicle ownership and unoccupied vmt generation. *Transportation Research Part C: Emerging Technologies 90*, 156–165.

Zhao, J., W. Deng, Y. Song, and Y. Zhu (2013). What influences metro station ridership in china? insights from nanjing. *Cities 35*, 114–124.

Zhao, P. and Y. Zhang (2018). Travel behaviour and life course: Examining changes in car use after residential relocation in beijing. *Journal of Transport Geography 73*, 41–53.

Zhao, Z. and J. Zhao (2018). Car pride and its behavioral implications: an exploration in shanghai. *Transportation*, 1–18.

Zhu, Y., M. Diao, J. Ferreira, and C. Zegras (2018). An integrated microsimulation approach to land-use and mobility modeling. *Journal of Transport and Land Use 11*(1).

Zhu, Y. and J. Ferreira Jr (2014). Synthetic population generation at disaggregated spatial scales for land use and transportation microsimulation. *Transportation Research Record 2429*(1), 168–177.