AN INTEGRATED FRAMEWORK FOR EXPLORING FINITE MIXTURE HETEROGENEITY IN TRAVEL DEMAND AND BEHAVIOR

A Dissertation
Presented to
The Academic Faculty

by

Sung Hoo Kim

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Civil and Environmental Engineering

Georgia Institute of Technology
August 2021

AN INTEGRATED FRAMEWORK FOR EXPLORING FINITE MIXTURE HETEROGENEITY IN TRAVEL DEMAND AND BEHAVIOR

Approved by:

Dr. Patricia L. Mokhtarian, Advisor
School of Civil and Environmental Engineering
Georgia Institute of Technology

Dr. Shatakshee Dhongde
School of Economics
Georgia Institute of Technology

Dr. Giovanni Circella
School of Civil and Environmental Engineering
Georgia Institute of Technology

Dr. Joan L. Walker
Department of Civil and Environmental Engineering
University of California, Berkeley

Dr. Jorge A. Laval
School of Civil and Environmental Engineering
Georgia Institute of Technology

Date Approved: June 28, 2021

To my family

ACKNOWLEDGEMENTS

Reflecting on my long journey to complete this dissertation, I've come to realize that I couldn't have gotten here without the support of people around me. I know I am now just at the beginning of a rich and challenging future, but I want to pause for a moment and would like to express my gratitude to those who have helped me.

First, I want to give my deepest thanks to my advisor, Dr. Patricia Mokhtarian. I cannot imagine myself without her patience and guidance. She has been a teacher, a life mentor, and a friend for me. Her teaching and intellectual challenges helped me become a scholar and her life mentoring made me overcome any hardships I faced. I will miss weekly meetings with her.

My committee members have been very supportive of me. Dr. Giovanni Circella has always encouraged me to do whatever I want to do in my research. Interactions with him helped equip me with critical thinking and made my graduate school life joyful. Dr. Jorge Laval, Dr. Shatakshee Dhongde, and Dr. Joan Walker provided numerous insightful comments on my dissertation. As well, I gained inspiration from Dr. Laval's lectures and Dr. Dhongde's and Dr. Walker's research.

I want to express my appreciation to the School of Civil and Environmental Engineering (CEE), the Georgia Institute of Technology (GT), and funding agencies. I believe that knowledge I learned from GT transportation faculty has been the ground of my dissertation. My life was made easier with the tremendous help of Marjorie Jorgensen and the CEE office staff. I could focus on my research thanks to funding from the Georgia

Department of Transportation, the center for Teaching Old Models New Tricks (TOMNET), and the President's Fellowship of Georgia Tech.

Many professors have guided me on how to build my career path and provided lifelong advice. I am deeply grateful to professors at Yonsei University – Dr. Jin-Hyuk Chung, Dr. Hyung Jin Kim, Dr. Bongsoo Son, and Dr. Jinhee Kim – and professors I interacted with at Georgia Tech – Dr. Sangho Choo, Dr. Joonho Ko, and Dr. Wonho Suh.

I would like to thank my lab mates, colleagues, and friends. I will cherish many memories of lab meetings and hanging-out with, Dr. Ali Etezady, Dr. Atiyya Shaw, Xinyi Wang, Grace Chen, Dr. Alex Malokin, Dr. Yongsung Lee, Dr. Sungtaek Choi, Dr. Shin-Hyung Cho, Dr. Gwen Kash, and Dr. Farzad Alemi. Dr. Hyun Woong Cho, Dr. Daejin Kim, and many other Korean colleagues helped me settle down in Atlanta and at Georgia Tech. I cannot list all friends here, but they have been my mental supporters and thus I could finish this journey.

Lastly, I am so fortunate to have my father, mother, and sister. They have always been my true life mentors whenever I need advice and my supporters whatever I do. I owe an unrepayable debt to the many sacrifices they have made for me to get to this place. This dissertation is truly dedicated to their unconditional love.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

In recent years we have faced a plethora of social trends and new technologies such as shared mobility, micro-mobility, and information and communication technologies, and we will be facing many more in the future (e.g. self-driving cars, disruptive events). In this context, the perennial mission of transportation behavior analysts and modelers – to model behavior/demand so as to understand behavior, help craft responsive policies, and accurately forecast future demand – has become far more challenging.

Specifically, behavioral realism and predictive ability are two key goals of modeling (travel) behavior/demand, and a key strategy for achieving those goals has been to introduce some type of *heterogeneity* in modeling. Thus, this thesis aims to improve our behavioral modeling by accounting for heterogeneity, with clues from the ideas of *data/market segmentation*, *finite mixture*, and *mixture modeling*. The objectives of the thesis are: (1) to build a framework for modeling finite mixture heterogeneity that connects seemingly less related models and various methodological ideas across domains, (2) to tackle various heterogeneity-related research questions in travel behavior and thus show the empirical usefulness of the models under the framework; and (3) to examine the potential, challenges, and implications of the framework with conceptual considerations and practical applications. Five inter-related studies in this thesis illuminate some part(s) of the framework and delineate how key concepts in the framework are connected to each other.

CHAPTER 1 and CHAPTER 2 start with discussions about the necessity of studying heterogeneity, related key concepts, and an overview of modeling finite mixture

heterogeneity. Through a comprehensive and systematic review, the study (1) provides a broader understanding of the usage landscape of finite mixture modeling, (2) sheds light on various typologies related to methodological approaches to treat heterogeneity, and (3) discusses alternative model configurations. Transportation researchers may benefit from this study by understanding the general idea of finite mixture heterogeneity and where we are now in this modeling. As well, analysts can use this study as a compass while designing their models.

CHAPTER 3 discusses parameter heterogeneity, which is the most popular type of heterogeneity. Specifically, the chapter connects three alternative approaches to treating finite-valued parameter heterogeneity: deterministic segmentation, endogenous switching, and latent class models. The study (1) expands the typology of mixture modeling by embracing "observed classes", and (2) connects the finite mixture model with the switching model family by way of detailed discussions about their similarities and differences from conceptual and empirical standpoints. Specifically, with equation-rich discussions the study points out the distinctive usefulness of each approach: the often-better performance of the latent class model over competing models, and the proper framework for estimating treatment effects offered by the endogenous switching model (including an in-depth interpretation of treatment effects). Analysts may benefit from this study by understanding the connections between two modeling families (thus supporting model selection appropriate to satisfying their ends) and obtaining the correct equations for calculating treatment effects, especially when the dependent variable is log-transformed.

CHAPTER 4 deals with the confirmatory latent class approach, which has been less discussed in the literature. The study illustrates the usefulness of the confirmatory latent

class approach with an empirical application (modeling leisure trip frequencies by car and air). Specifically, the zero-inflated model is embraced under the finite mixture heterogeneity framework, given the expanded typology of heterogeneity. Analysts may gain inspiration from this study on how to operationalize behavioral models when dealing with data showing a particular pattern and when having some behavioral hypotheses on such a pattern.

CHAPTER 5 expands the latent class model by combining it with the endogenous switching model. It relaxes the latent class model's implicit assumption of independence between the unobserved influences on class membership and outcome. With two empirical applications (modeling the willingness to share autonomous vehicle rides with strangers and the adoption of ridehailing for social-purpose trips), the study shows how the proposed models may give different insights compared to standard latent class models, even when parameter estimates and goodness-of-fit measures appear to be similar. Specifically, when conducting scenario analysis, the proposed method provides distinct marginal and conditional (on class) expectations, whereas the standard model only focuses on conditional expectations. The study opens the door to an avenue for evaluating "treatment effects" in the latent class modeling context, which analysts may wish to pursue in the future.

CHAPTER 6 conceptually connects latent class modeling to the mixture of experts (MoE) approach arising from the machine learning domain. This study uses MoE as a data-driven exploratory tool to identify nonlinear and interaction effects (which are special types of parameter heterogeneity) and uses what we learn from MoE to improve the performance of conventional models. Through experiments with synthetic data and an empirical

application (to mode choice), the study shows that MoE can automatically detect nonlinear/interaction effects and can be used to inform our model specifications. To our knowledge, this study is the first in the transportation domain to use the "indirect application" (as it is known in the psychometrics field) of latent class modeling. Hence, the study expands the usage of finite mixture structures and thus helps to diversify applications for analysts.

The journey of this thesis concludes with discussions about challenges, potential technical advances, and outlook for the framework (CHAPTER 7). The dissertation is expected to give conceptual/methodological insights into the framework for modeling finite mixture heterogeneity and how various methodologies are connected under the framework. As well, the studies provide rich discussions about study-specific empirical findings and their implications. Thus, the dissertation can help improve our behavior/demand models by serving as a navigational compass for analysts. The conceptual contributions, methodological novelties, and selected empirical findings of this thesis are summarized in the following tables.

	Conceptual contributions	Methodological novelties
Ch2 & Ch7	• The first comprehensive review of heterogeneity and mixture modeling in transportation • Develops a typology of heterogeneity in travel behavior/demand • Examines key elements of mixture modeling and thus portrays the general usage of the method • Discusses potential technical advances • Presents critical issues and challenges of the finite mixture approach that have not been discussed	• Applies topic modeling to papers of specific interest in the Scopus database (as opposed to typical applications in which topic modeling is applied to any type of study in a broad domain)
Ch3	• The first conceptual connections/ comparisons among three finite segmentation models: deterministic, endogenous switching, and latent class • Extends the concept of finite mixture/ segmentation	• Discusses model-specific usage and implications (performance, interpretations) when applied in the context of vehicle-miles driven (VMD) modeling • Provides complete equations for calculating various treatment effects when a log-transformation is applied (which has not been covered in travel behavior research)
Ch4	• Embraces zero-inflated models under the confirmatory latent class approach • Discusses the usefulness of the confirmatory latent class approach	• The first application that probabilistically decomposes different types of zeros in the context of modeling long-distance trip frequency
Ch5	• The first introduction of the idea of combining latent class and endogenous switching models in the transportation domain • Delineates subtle conceptual differences between latent class models with and without an error structure • Derives marginal effects of the model • Discusses the issue of evaluating latent class models	• The first application of latent class modeling in the context of modeling the willingness to share automated vehicle (AV) rides with strangers and adoption of ridehailing for social-purpose trips • Conducts statistical inference based on (parallelized) bootstrapping • Illustrates the usefulness of the method with scenario analyses
Ch6	• The first study in travel behavior research that introduces the ideas of indirect application of latent class modeling and the mixture of experts (MoE) architecture • Proposes the idea of using MoE as a data-driven tool to identify nonlinear/interaction effects	• The first application of MoE in the travel behavior and choice modeling communities • Demonstrates the approximation abilities of MoE by experimenting with synthetic data

	Empirical findings (selected)
Ch2 & Ch7	• Found that heterogeneity and mixture modeling have gained popularity over the years in transportation research publications • Identified six subdomains in transportation that use mixture modeling: *discrete choice modeling*, *general behavior analysis*, *crash/safety analysis*, *traffic analysis*, *travel time distribution*, and *electric vehicles* • Summarized types of heterogeneity and related applications in the literature: *variable distributions*, *parameters*, *model specification*, *attribute processing*, *functional forms*, *decision rules*, *causal structure/order*, *constraint/choice set* • Illuminated that supervised learning and unsupervised learning applications tend to have divergent numbers of classes in their respective final solutions; many studies determined such a number qualitatively rather than quantitatively
Ch3	• Urban residents were more sensitive to the availability of transit, whereas non-urban residents were more sensitive to local amenities • The lower-VMD latent class was influenced to drive less when living in more job-dense or better transit-service areas, whereas the higher-VMD class was not significantly influenced by these factors • Propensities associated with residential location choice and VMT generation shared common unobserved factors
Ch4	• Identified profiles of those in the structural zero-trip regime (as opposed to the trip-making regime): they were the oldest and have the lowest household income for both air and car travel • The presence of children acted as a barrier to belonging to the trip-making regime for air travel, but it was a facilitator of doing so for car travel; it was negatively associated with the number of trips in both modes • As distance to airport increased, both entry into the trip-making regime and number of trips were inhibited for air travel, but car travel exhibited the opposite effects
Ch5	• Found significant correlations between unobserved influences on latent segmentation and behavioral processes (for both empirical contexts of modeling the willingness to share AV rides with strangers and adoption of ridehailing for social-purpose trips) • Males, more educated, and those who have used ridehailing services were more willing to share AV rides with strangers • Identified a latent group of people who are "structurally unwilling" to share AV rides with strangers; (none of the tested factors affects their willingness)
Ch6	• Experiments with synthetic data showed that MoE can capture nonlinear/interaction effects without prior knowledge of those effects • MoE identified significant nonlinear effects of time and cost on mode choice in an empirical application • Found significant interaction effects in this empirical application: in particular, travel time interacted with gender and trip purpose; travel cost interacted with gender and seat grade of train • Conventional logit models were substantially improved by re-specifying them on the basis of what MoE learned from the data

CHAPTER 1. INTRODUCTION

1.1 Motivation

People make numerous transportation-related choices, always under a constellation of circumstances, but such circumstances – urban environments, transportation systems, and society – have been radically changing. We have faced a plethora of social trends and new technologies such as shared mobility, micro-mobility, and information and communication technologies, and we will be facing many more in the future (e.g. self-driving cars, disruptive events). In this context, the perennial mission of behavioral analysts and transportation modelers – to model behavior/demand so as to understand behavior, help craft responsive policies, and accurately forecast future demand – has become far more challenging.

Behavioral realism and predictive ability are two key goals of modeling (travel) behavior. It may seem that these goals are the same, in that better understanding of behavioral mechanisms and causality should lead to better predictions of outcomes. However, arguably, they are not completely congruent (e.g. Shmueli, 2010), and trying to achieve both at once may sometimes feel like "chasing two rabbits". Many social science fields use statistical models "almost exclusively for causal explanation" (Shmueli, 2010, p. 289), whereas some fields mainly focus on the utilitarian need for predictive accuracy. In the transportation domain, both goals co-exist, in sometimes separate but often overlapping realms. This reality is perhaps exemplified by the existence of separate standing committees of the US Transportation Research Board, respectively devoted to Traveler Behavior and Values (AEP30) and to Transportation Demand Forecasting (AEP50).

Membership on the former committee is typically dominated by academic researchers, while membership on the latter is a purposeful mixture of researchers with planners/ practitioners/industry in roughly equal proportions. To some extent, this dual nature can be traced to the position of the transportation field on the boundary between social science and engineering, where its social science perspective often leads to a search for behavioral explanations while its engineering perspective often calls for accurate forecasting in support of infrastructure development, planning, and policy.

Myriads of models have been developed in various fields such as statistics, economics, psychology, and data science. However, such models were developed for different purposes and contexts and, even if some models are eventually performing similar mathematical tasks, they may have different names and/or application approaches. A major challenge for many researchers and practitioners is, "what model(s) is (are) appropriate for addressing this question?" which in turn stems from several sub-questions, such as: (1) what basic assumptions should be considered in order to have a *useful* model[1] and (2) to what extent is compromise acceptable (including tradeoffs between behavioral realism and prediction ability, if any)?

A key strategy for achieving either behavioral realism or predictive ability has been to introduce some type of *heterogeneity* in modeling. *Heterogeneity* has become a popular concept in (but not limited to) the behavior modeling field, in that it is conceptually more realistic for explaining human behavior. Many questions can arise at this point: What types of heterogeneity exist in behavioral modeling? How do we incorporate heterogeneity in

[1] This question follows from the famous saying by statistician George Box, "All models are wrong, but some are useful," which is an expanded version of the quote in Box (1976).

modeling travel demand and behaviors? What is the empirical usefulness of such modeling? What are the challenges and implications in applications? What are possible challenges and avenues for methodological advances? Addressing these questions may ultimately help improve our behavior/demand models. Thus, this thesis is a journey pursuing the answers to those questions, with clues from the ideas of *segmentation*, *finite mixture,* and *mixture modeling*.

1.2 Setting up the context

1.2.1 Heterogeneity

We may need to step back and question, why heterogeneity? The concept of heterogeneity (or related synonyms) is ubiquitous and thus we can find it in almost every domain. Figure 1-1 shows the appearance of the word "heterogeneity" in the literature.[2] Figure 1-1a conveys a general idea that the word "heterogeneity" is more commonly used than the word "homogeneity" in recent decades. As well, Figure 1-1b illustrates, overall, that "heterogeneity" is far more often used than "homogeneity" in the academic articles of numerous domains. While one reason for these patterns may be that there is a lower perceived need to articulate the dominant, often implicit paradigm of homogeneity (whereas heterogeneity, when considered, generally needs to be explicitly named and contrasted with the dominant paradigm), the figure clearly suggests that the idea of heterogeneity is shouldering a greater role than homogeneity in research, and increasingly

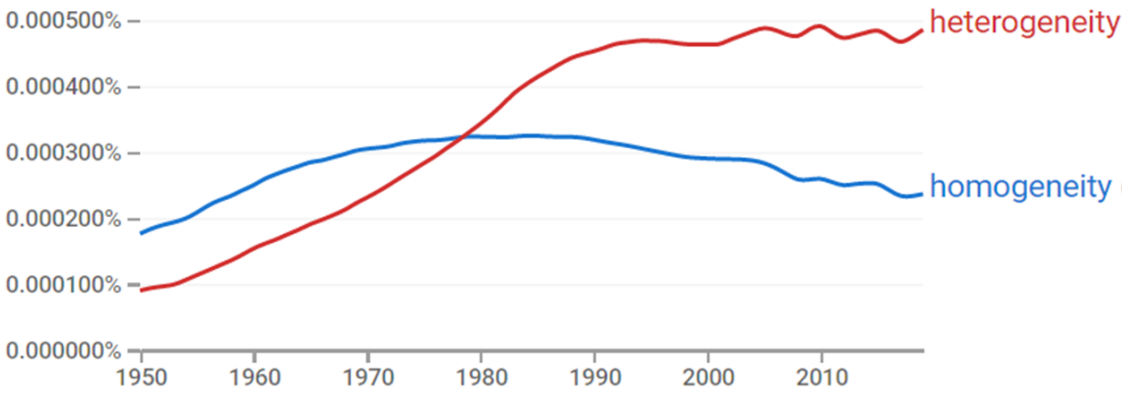
[2] The purpose of this figure is to give a general idea of how these words have appeared in the literature. The exact meaning of heterogeneity/homogeneity may differ substantially depending on the context, and appearance in the literature itself does not necessarily mean that those concepts are the core of the literature. As well, many variant words including synonyms are not considered here.

more so in recent years. It is instructive to dip into the thinking throughout history of some selected influential scholars with respect to heterogeneity (bolding added for emphasis).

- "The asymmetry may arise from the fact that the units grouped together in the measured material are **not really homogeneous**. It may happen that we have a mixture of 2, 3, …, *n* homogenous groups…" (Pearson, 1894, *the first finite mixture model*)
- "In considering the **sub-groups** of a population – especially in dealing with local races in man, animals or plants – a problem of the following character has not infrequently arisen: It is found that a sub-class, for example a local sample, differs considerably from the general population." (Pearson, 1906)
- "**Market segmentation**, on the other hand, consists of viewing a **heterogeneous** market (one characterized by divergent demand) as a number of smaller homogeneous markets in response to differing product preferences among important market segments." (Smith, 1956, *the first idea of market segmentation*)
- "But all evolutionary biologists know that **variation** itself is nature's only irreducible essence. Variation is the hard reality, not a set of imperfect measures for a central tendency. Means and medians are the abstractions." (Gould, 1985)
- "Accounting for **heterogeneity** and **diversity** and its implications for economics and econometrics is a central message of this [Nobel] lecture and a main theme of my life's work." (Heckman, 2000, *Nobel Memorial Prize in Economic Sciences*)
- "The original formulation of RUM [random utility maximization] as a behavioral hypothesis started from the standard model, with randomness attributed to unobserved **heterogeneity** in tastes, experience, and information on the attributes of alternatives." (McFadden, 2000, *Nobel Memorial Prize in Economic Sciences*)
- "Predictive accuracy is substantially improved when blending **multiple predictors**. *Our experience is that most efforts should be concentrated in **deriving substantially different approaches**, rather than refining a single technique*." (Bell et al., 2007, *Netflix Prize*)

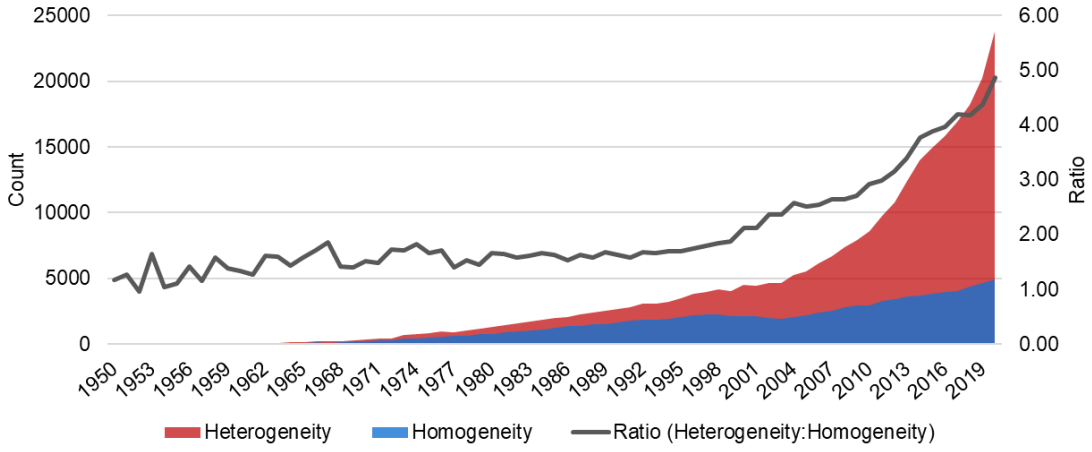
- "This approach, called '**ensemble of specialists**',… overall forecasting accuracy would improve if one used a **separate model for each group** instead of a single one for the whole dataset." (Smyl, 2020, *M4 Competition*)

(a) Google Books Ngram Search



Note: this displays a graph showing how the words have occurred in a corpus of books over the years (retrieved on May 24, 2021; Source: Google Books)

(b) Number of articles having the words (Scopus database)



Note: Searched the words in title/keyword/abstract. Focused on "article" type documents

Figure 1-1. Appearance of the words "heterogeneity" and "homogeneity" over time

The meaning of heterogeneity itself, of course, might be very heterogeneous depending on the context and domain. This thesis will articulate the meaning of heterogeneity in travel demand/behavior studies in a later section (2.3.1). However, from

a broader perspective, we can observe that there has been a consensus about the importance of considering heterogeneity in numerous fields. In addition, many scholars and studies have been finding a pathway to leverage our knowledge and/or improve their research/models by accounting for heterogeneity. This motivates the focus of this thesis on heterogeneity as a pathway to better behavior/demand modeling.

1.2.2 Finite mixture modeling

As discussed further below, in essence heterogeneity can have two fundamental natures: continuously varying across the population, or taking on only a finite number of different versions. This thesis specifically focuses on the finite nature of heterogeneity, anchored in mixture modeling but relating it to model cousins as appropriate.

Finite mixture modeling is a statistical approach to modeling a variety of random phenomena, and it has a long history. As noted in McLachlan and Peel (2001), one of the first major analyses using mixture modeling was in the late 1800s. Specifically, Pearson (1894) fitted a distribution of the body length of crabs using a mixture of two normal distributions, indicating the possibility of two sub-species. Notable features of the mixture model are that it has a probabilistic nature and it can disentangle latent structure in the data (or subgroups in the population). Hence, the basic idea of mixture modeling is to posit the existence of within-subgroup homogeneity but between-subgroup heterogeneity in the population and to model those heterogeneous patterns/distributions/behaviors. A general form of finite mixture models is as follows:

$$f(y) = \sum_{z=1}^{Z} f(y, z) = \sum_{z=1}^{Z} P(z|\boldsymbol{W}) f_z(y|z, \boldsymbol{X}) \,, \tag{1.1}$$

where y is an outcome variable (target variable, dependent variable), $\boldsymbol{X}$ is a vector of variables explaining y (covariates), $\boldsymbol{W}$ is a vector of variables explaining subgroup membership z (segmentation bases), z is a *discrete* segment or subgroup indicator ($z = 1, 2, \dots, Z$), $P(\cdot)$ denotes a mixture density function (or segment membership probability), and $f_z(\cdot)$ denotes an outcome function for segment z. We will describe functional forms of $P(\cdot)$ and $f_z(\cdot)$ in later sections (Sections 2.3.4 and 2.3.5). It is useful to see a graphical illustration of homogeneous and heterogeneous behavioral processes that could be handled by the finite mixture modeling paradigm (Figure 1-2). As opposed to assuming homogeneity in the data generation process (or behavior generation process, in the context of behavior studies), finite mixture modeling posits that there are multiple subpopulations having different behavior generation processes. Putting this in statistical terms, we aim to find the joint density of y and z (and thence the marginal density of y, obtained by summing over, or marginalizing out, z), which can be expressed as a product of the marginal probability of belonging to a segment and the conditional density of outcome given segment. By decomposing the joint density into two parts, it brings benefits of interpretation and the potential of technical extensions.

When discussing finite mixture modeling, for a better understanding of the methodology it is instructive to make some distinctions with respect to some relevant concepts: (1) *continuous* versus *discrete/finite* mixture and (2) *disaggregation* versus *segmentation*. These two contrasts will be discussed in turn.

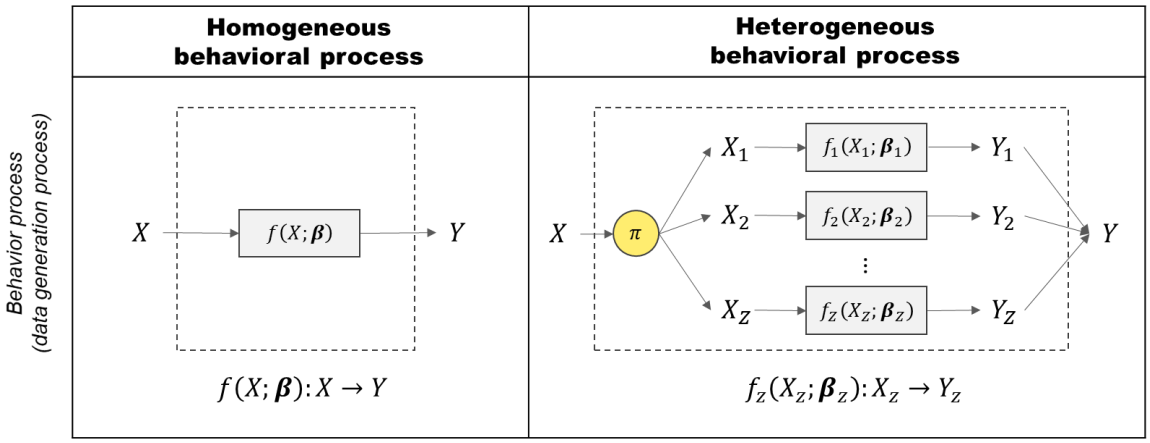


Figure 1-2. Conceptual illustration of homogeneous and heterogeneous behavioral processes

1.2.2.1 Continuous versus discrete/finite mixture

In the statistics literature, the weighted average of several functions is called a *mixed function*, and the density that provides the weights is called the *mixing distribution* (Train, 2009). Although we focus on finite mixtures, more broadly, mixture distributions can be either continuous or discrete (cf. Walker and Ben-Akiva, 2011; Vij and Krueger, 2017)[3]. Continuous mixture models can appear in two different forms (Figure 1-3): mixtures over *parameter space*, and mixtures over *latent variable space*. A common example of a mixture over parameter space is the case in which the travel time coefficient in a mode choice model is specified to have a continuous distribution, $g_\beta(\beta)$, in the

[3] It is possible to have another typology, where we focus on the distribution type instead of on the overall model structure: parametric mixture distribution, nonparametric distribution, and semi-(non)parametric distribution (cf. Vij and Krueger, 2017). Typical latent class models can be classified in the nonparametric family in that "the support of the distribution is defined as a fixed number of points in a high-dimensional coefficient space" (Vij and Krueger, 2017, p. 78). However, finite mixture modeling can also be considered a semi-parametric approach, which is between the fully parametric and nonparametric approaches (McLachlan and Peel, 2001).

population. In the general case where the vector of model parameters can have continuous distributions, the mixture model can be expressed as follows:

$$f(y) = \int f(y|\boldsymbol{X};\boldsymbol{\beta}) g_{\beta}(\boldsymbol{\beta}) d\boldsymbol{\beta} \,. \tag{1.2}$$

This means that $f(y|\boldsymbol{X};\boldsymbol{\beta})$ is weighted by $g_{\beta}(\boldsymbol{\beta})$ and integrated over $\boldsymbol{\beta}$, since $g_{\beta}(\boldsymbol{\beta})$ is a continuous density. The mixed logit model, in which $f(y|\boldsymbol{X};\boldsymbol{\beta})$ is the logit probability function, is a popular form of this type of mixture model (cf. Hensher and Greene, 2003; Train, 2009).

With respect to mixing over the latent variable space, the model can be expressed as follows:

$$f(y) = \int f(y|z,\boldsymbol{X};\boldsymbol{\beta}) g_{z}(z|\boldsymbol{W};\boldsymbol{\alpha}) dz \,, \tag{1.3}$$

where z is a latent (unobserved) variable. This means that $f(y|z,\boldsymbol{X};\boldsymbol{\beta})$ is weighted by $g_{z}(z|\boldsymbol{W};\boldsymbol{\alpha})$. The integrated choice and latent variable model (ICLV, or hybrid choice model; cf. Vij and Walker, 2016) is a particular type of this mixture model (Walker, 2001), in which z is a continuous-valued variable such as an attitude, and $f(y|z,\boldsymbol{X};\boldsymbol{\beta})$ is a discrete choice model.

Finite mixture models can be viewed as degenerate special cases of both kinds of continuous mixture models, in which the mixing distribution is finite-valued and therefore the integrals of Eq. (1.2) and Eq. (1.3) are replaced by the summation in Eq. (1.1). The latent class model is the counterpart of Eq. (1.2) in which the parameter vector $\boldsymbol{\beta}$ only takes

on a finite number of values, each with a non-zero probability; it is the counterpart of Eq. (1.3) in which the latent variable z is a finite-valued marker of class membership, where the probability that z takes on a given value is being modeled by a function of observed variables $\boldsymbol{W}$ and parameters[4] $\boldsymbol{\alpha}$. Walker and Ben-Akiva (2011) compared the continuous behavioral mixture, Eq. (1.3), and discrete behavioral mixture models, Eq. (1.1).

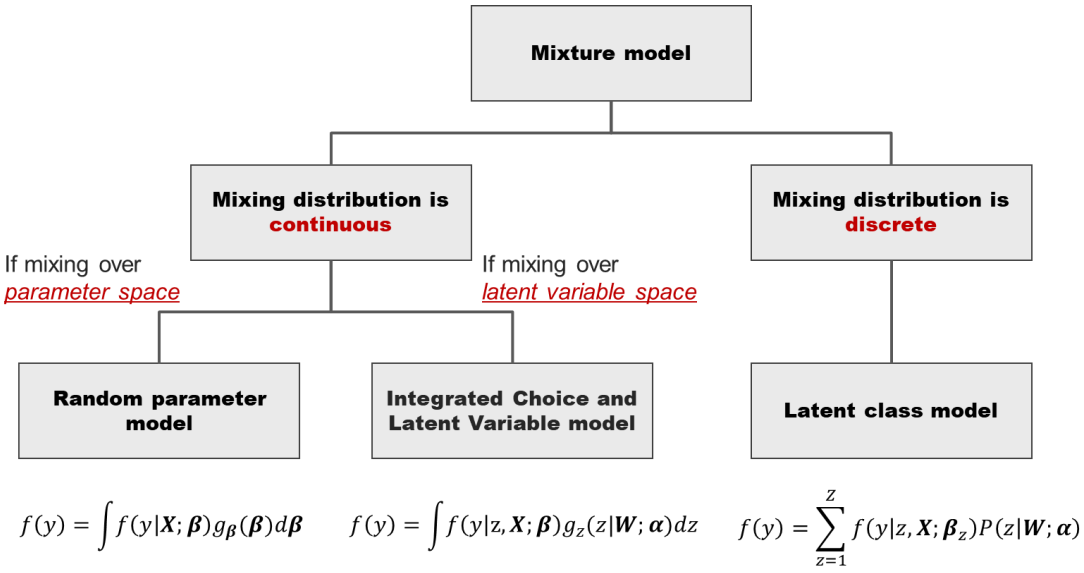


Figure 1-3. A tree of continuous and discrete/finite mixture modeling

In this thesis, we focus on the discrete/finite mixture due to its three unique benefits. *First*, it brings **convenience** both conceptually and technically. Conceptually, it is cognitively easier to think of a finite number of sets of parameters rather than considering the distribution(s) of parameters. Along with this, it often provides a more tangible explanation of an individual's behavior. Consider the mixed logit and latent class models for mode choice modeling as a simple example. With a discrete mixture, it is easier to

[4] The parameters $\boldsymbol{\alpha}$ can be considered hyperparameters as well, since we are reparameterizing constant class membership probabilities as functions of $\boldsymbol{W}$. This will be addressed further in Section 2.3.4.

understand that a particular type of person may have a willingness to pay of $A as opposed to $B for those in another class. On the other hand, mixed logit models may indicate that willingness to pay exhibits a certain distribution in the population, but it gives limited information as to where a person with certain characteristics would fall in such a distribution. To obtain such information from a continuously-distributed random parameter, the analyst must resort to parameterizing the parameters characterizing the distribution of the outcome model's parameter (i.e., hypothesizing heterogeneity in the mean and/or variance of a given model parameter's continuous distribution, and specifying such heterogeneity as a certain function of $\boldsymbol{W}$), which only *proliferates* the parameter's distribution rather than *simplifying* it. The technical convenience of finite mixture models arises because in continuous mixture models, the integral does not offer a closed form solution and thus it requires simulation for estimation. This means that in general it may require a longer time for estimation (although this can depend on model specification).

Second, the discrete mixture is **nicely connected with some other useful concepts or modeling approaches**. One such important concept is *market segmentation* (see Section 2.1.2), which has been long and successfully used in the marketing field, and to which discrete mixture models are well suited (Wedel and Kamakura, 2012). In addition, discrete mixture models can even be associated with other segmentation models such as switching models (CHAPTER 3), and the ensemble method by weighting in machine learning (CHAPTER 6).

Lastly and importantly, whereas continuous mixture models focus on how parameters/latent variables are distributed, discrete mixture models in fact **expand our modeling capability by examining more diverse types of heterogeneity** beyond

parameter heterogeneity (as will be covered in Section 2.3.1). However, the two approaches each have their own advantages, and there are often tradeoffs between them (e.g. continuous mixture models require distributional assumptions, but discrete mixture models are subject to decisions on the number and nature of classes). Therefore, the thesis does not assert superiority of finite mixture models over continuous mixture models; rather it aims to deepen our understanding of the finite mixture case and its implications.

1.2.2.2 Disaggregation versus segmentation

Another important distinction is to understand the meaning of the "segmentation" which is achieved by finite mixture modeling. Two key concepts are *the level of (analysis unit) (dis)aggregation* and *the level of (data) segmentation*. The terminology can be confusing because the word "disaggregation" is sometimes used as a synonym for "segmentation" (e.g., "the sample was disaggregated by income category"), but at the same time, "level of (dis)aggregation" can be used to describe one trait of the unit of analysis. The two dimensions of (dis)aggregation and segmentation have different implications for data analysis and its interpretation.

We are exposed to a large spectrum of data – potentially about countries, states, cities, neighborhoods, transit agencies, employee groups, households, individuals, vehicles, and so on. One way of characterizing a data set is by the unit of observation: what type of entity is being measured by each data point? The unit of observation falls at some level of **disaggregation**: the aggregate level (i.e. each data point represents some aggregated group of actors), disaggregate level (i.e. each data point represents a single actor such as a person), and individual level (i.e. multiple data points are captured from the same

actor). In theory, data available at a finer-grained level of disaggregation can be aggregated to higher levels, provided there are "enough" finer-grained cases being aggregated to provide a good measure of the coarser-grained case they represent, and "enough" cases at the coarser level to enable reliable statistical analysis at that level. Obviously, the definition of an "actor" and the level of disaggregation can be context-dependent. As a concrete example, suppose we want to estimate the elasticity of driving distance with respect to income. In the aggregate-level analysis, we may collect data on average income and total vehicle-miles traveled (VMT) in major cities in the US, and model VMT as a function of income. Note that the aggregation could also be at the state or county level, and so on. At the disaggregate level, we may collect data on the income and VMT of individuals and estimate a similar model. At the individual level, we can even focus on a particular person and model a VMT-income relationship across multiple years of measurement.

Another dimension for a study's modeling strategy is the (often data-driven) **segmentation** of the data, i.e. the subdivision of the data into smaller groups. For example, if we had a sample of New York state residents, for modeling transit use behavior, we might be tempted to split the sample into residents of the New York metropolitan area and the rest of the residents. Note that this segmentation does not change the level of the observation unit: it would be the individual (disaggregate level) whether we segmented the data or not. In the continuum of segmentation, at one end is no segmentation (or pooled data), meaning that the data are analyzed as one group representing the whole "population". The other end is highly-segmented and, at the extreme, each data point can be considered as its own segment. In theory, data segmentation can be applied to any level of (dis)aggregation of the data. Finite mixture modeling, which is the focus of this thesis, is

an approach of segmentation. In this study, we particularly focus on the segmentation scheme (via mixture modeling) with disaggregate-level data.

1.2.3 Data used in this thesis

For empirical applications, the thesis mainly uses survey data hereafter referred to as the "GDOT data" (Figure 1-4; CHAPTER 6 is an exception in that it employs synthetic data and a publicly available dataset, both of which will be described in that chapter). The survey was designed and the data were collected as a part of a research project titled "The Impact of Emerging Technologies and Trends on Travel Demand in Georgia", funded by the Georgia Department of Transportation (GDOT). In keeping with the project's name, the survey aimed to explore the impacts of emerging technologies and trends on travel behavior in Georgia (2017-2018), and accordingly the population of interest was adult residents of Georgia. The survey employed a combination of two sampling approaches: (1) recruiting respondents through address-based stratified random sampling in the 15 Metropolitan Planning Organization (MPO) areas in Georgia (the "main" sample), and (2) recontacting survey participants who took the 2016–17 National Household Travel Survey (Westat, 2018) in Georgia (which included residents of non-MPO areas) and agreed to be surveyed further (the "NHTS" sample). The working dataset includes ~3,300 cases, but the final sample size varies by the study (each chapter will describe the working data and key variables in that study). More details about the survey and the data are available in the final report of the project (Kim et al., 2019b).

The GDOT data were enriched through appending additional information based on geocodes of respondents' home locations. External data sources include the American

Community Survey (ACS), Longitudinal Employer-Household Dynamics (LEHD), Alltransit, Google Place API, and Google Map API. In addition, several variables were created by using *factor analysis* (e.g. attitudinal constructs, Table 1-1) and *principal component analysis* (e.g. a proxy for local accessibility, Table 1-2).

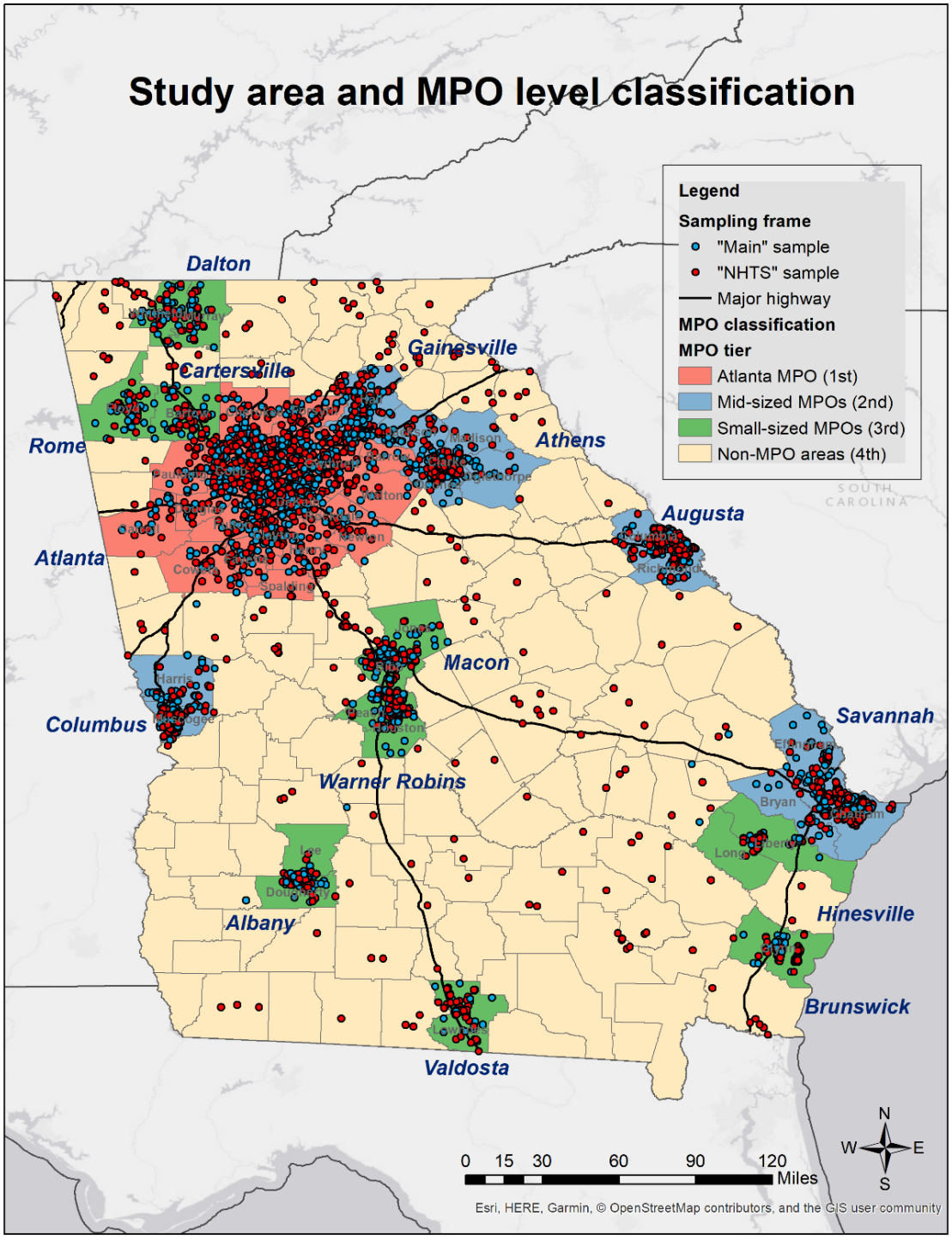


Figure 1-4. Geographical distribution of the GDOT data

Table 1-1. Attitudinal Factors and Corresponding Factor Loadings

Factor	Statement	Pattern matrix loading
Non-car alternatives	I like the idea of walking as a means of travel for me.	0.666
	I like the idea of bicycling as a means of travel for me.	0.628
	I like the idea of public transit as a means of travel for me.	0.336
Tech-savvy	Learning how to use new technologies is often frustrating for me.	−0.866
	I am confident in my ability to use modern technologies.	0.801
Commute benefit	My commute is a useful transition between home and work (or school).	0.677
	My travel to/from work (or school) is usually pleasant.	0.579
	I wish I could instantly be at work (or school)—the trip itself is a waste of time.	−0.428
Modern urbanite	I like the idea of having stores, restaurants, and offices mixed among the homes in my neighborhood.	0.417
	My phone is so important to me, it's almost part of my body.	0.350
Work-oriented	At this stage of my life, having fun is more important to me than working hard.	−0.572
	I'm too busy to have as much leisure time as I'd like.	0.527
	It's very important to me to achieve success in my work.	0.298
Materialistic	I usually go for the basic ("no-frills") option rather than paying more money for extras.	−0.565
	The functionality of a car is more important to me than the status of its brand.	−0.431
	I would/do enjoy having a lot of luxury things. [c]	0.426
	I like to wait a while rather than being first to buy new products.	−0.357
	I prefer to minimize the amount of things I own.	−0.341
Polychronic	I prefer to do one thing at a time.	−0.834
	I like to juggle two or more activities at the same time.	0.697
Pro-environmental	Cost or convenience takes priority over environmental impacts (e.g., pollution) when I make my daily choices.	−0.914
	I am committed to an environmentally friendly lifestyle.	0.481
Pro-exercise	The importance of exercise is overrated.	−0.669
	I am committed to exercising regularly.	0.663
Family/friends-oriented	Family/friends play a big role in how I schedule my time.	0.612
	It's okay to give up a lot of time with family and friends to achieve other worthy goals.	−0.468
Pro-suburban	I prefer to live in a spacious home, even if it's farther from public transportation or many places I go to.	0.609
	I see myself living long-term in a suburban or rural setting.	0.387
Waiting-tolerant	Having to wait is an annoying waste of time.	−0.831
	Having to wait can be a useful pause in a busy day.	0.533
Travel liking	I generally enjoy the act of traveling itself.	0.618
	I like exploring new places.	0.593
Sociable	I consider myself to be a sociable person.	0.563
	I'm uncomfortable being around people I don't know.	−0.507
Pro–car-owning	I definitely want to own a car.	0.748
	I am fine with not owning a car, as long as I can use/rent one any time I need it.	−0.576
	I like the idea of driving as a means of travel for me.	0.535
	As a general principle, I'd rather own things myself than rent or borrow them from someone else.	0.404

Note: Factor loadings under 0.3 in magnitude are suppressed.

Table 1-2. PCA pattern loadings on number of amenities near geocoded home

Number of amenities	Pattern loading	Number of amenities	Pattern loading
Bar	0.803	Bakery	0.821
Convenience store	0.741	Café	0.838
Doctor	0.734	Dentist	0.711
Florist	0.670	Parking	0.606
Home goods store	0.857	Bank	0.738
Liquor store	0.698	Book store	0.655
Restaurant	0.887	Clothing store	0.835
Beauty salon	0.883	Hair care	0.860
Gas station	0.661	Library	0.545
Park	0.636	Pharmacy	0.634
School	0.794	Supermarket	0.568
Store	0.726		

Note: Numbers of amenities near the home location are collected via the Google Map API.

1.3 Knowledge gaps and research objectives

The ideas of heterogeneity, segmentation, and finite mixture modeling have been around in the transportation domain for some time and are becoming more popular recently (cf. Section 2.2). Then, what will be the merits of this thesis? Years of delving into these topics point to some knowledge gaps.

First, the concepts of heterogeneity and finite mixture modeling have gained popularity, but the literature lacks a framework that integrates various types of heterogeneity and their model configurations. So the questions are: How has the transportation domain used finite mixture modeling? What types of heterogeneity have been discussed? What is the proper model configuration for a specific type of heterogeneity and what are the alternatives? (CHAPTER 2 – CHAPTER 6) Second, there has been little effort to connect methodological ideas scattered in various domains. For example, the relationship may seem slight at first, but the finite mixture model can be connected with

several classical econometric models (CHAPTER 3 and CHAPTER 4) and even models developed in the machine learning domain (CHAPTER 3 and CHAPTER 6). Knowledge of linkages among related models could bring benefits of enriching interpretations and the potential to shed light on new pathways to new approaches (e.g. CHAPTER 5 and CHAPTER 6). Lastly, there are plenty of conceptual/methodological details and issues "under the hood", but discussions about those are scarce (and accordingly, each chapter, and especially CHAPTER 7, will aim to invite unique discussions). In this regard, the goals of this thesis are threefold:

1. To build a framework for modeling finite mixture heterogeneity that connects seemingly less related models and various methodological ideas across domains;
2. To tackle various heterogeneity-related research questions in travel behavior and thus show the empirical usefulness of the models under the framework;
3. To examine the potential, challenges, and implications of the framework with conceptual considerations and practical applications.

1.4 Thesis outline and contributions

This section outlines the structure of the thesis (Figure 1-5) and describes the key contents of each chapter. In the thesis, each core chapter illuminates some part(s) of the framework and delineates how key concepts in the framework are connected to each other (and to other chapters as well). A summary of the conceptual contributions and methodological novelties of the thesis is offered in Table 1-3, and a summary of selected empirical findings appears in Table 1-4.

CHAPTER 1 provided the motivation of this thesis and set up the background and thesis objectives. This section outlines the rest of the thesis.

CHAPTER 2 aims to provide a broader understanding of the usage landscape of finite mixture modeling, and also insights into detailed elements of the approach through a comprehensive and systematic review. The chapter does not simply summarize the literature; rather, it aims to provide conceptual insights (e.g. the typology of heterogeneity in Section 2.3.1). Section 2.1 skims related concepts in other domains. Section 2.2 presents how we obtained a pool of relevant papers and how the studies are distributed by year and topic. Section 2.3 dives into the key elements of the finite mixture model and how transportation studies have used the method. To the best of our knowledge, this is the first comprehensive conceptual/review study exploring heterogeneity and mixture modeling in the transportation domain.

CHAPTER 3 focuses on parameter heterogeneity, which is the most popular type of heterogeneity. Specifically, the chapter connects three alternative approaches to treating parameter heterogeneity: deterministic segmentation, endogenous switching, and latent class models. The study compares them from theoretical and conceptual standpoints (Section 3.3) and with empirical applications (modeling vehicle-miles driven; Section 3.4). In addition, the chapter provides some important discussions related to the models, especially notes about estimating treatment effects (Section 3.5). To our best knowledge, this is the first study to connect those three alternative models, including theoretical, conceptual, and empirical comparisons.

CHAPTER 4 deals with the confirmatory latent class approach. The confirmatory approach is introduced in Section 2.3.2, as distinguished from the exploratory approach which is more common in the literature. This chapter illustrates the usefulness of the confirmatory latent class approach with an empirical application (modeling the frequency

of overnight domestic leisure trips by car and air). Specifically, the zero-inflated model is embraced under the finite mixture heterogeneity framework given the expanded typology of heterogeneity. To our best knowledge, this is the first study to probabilistically classify two types of zero-trip cases (structural vs. incidental, which will be described in the chapter) in the context of modeling long-distance trip frequency.

CHAPTER 5 expands the latent class model by combining it with another, previously encountered (CHAPTER 3), model family – the endogenous switching model. It relaxes an (often implicit) assumption of independence of the latent class model, by allowing the unobserved influences on class membership and outcome to be correlated. In doing so, however, the model deviates from the standard finite mixture model; hence, the chapter discusses the implications of this idea (Section 5.2). With two empirical applications (modeling willingness to share autonomous vehicle rides with strangers and adoption of ridehailing for social-purpose trips), the chapter shows how the proposed models may give markedly different pictures compared to the standard latent class models, even when parameter estimates and goodness-of-fit measures appear to be similar (Sections 5.3 and 5.4). As far as we know, this is the first study in the transportation domain that introduces the idea of combining latent class and endogenous switching models.

CHAPTER 6 examines the potential of using the mixture of experts (MoE) method as a data-driven exploratory tool to capture nonlinear and interaction effects (which are special types of parameter heterogeneity). Section 6.2 describes how the MoE fits into the framework of finite mixture heterogeneity (a so-called “indirect application” of finite mixture modeling). Section 6.3 verifies the usefulness of the MoE method with synthetic data and Section 6.4 applies the method to empirical data (mode choice). To our best

knowledge, this is the first study in the transportation domain that connects latent class modeling and MoE and proposes the idea of using MoE as a data-driven tool to capture nonlinear/interaction effects.

CHAPTER 7 summarizes the thesis and invites further discussions. It suggests several avenues for future technical advances and presents some issues regarding mixture modeling. It concludes with remarks on the use of the methodology and improvements of our behavior/demand modeling.

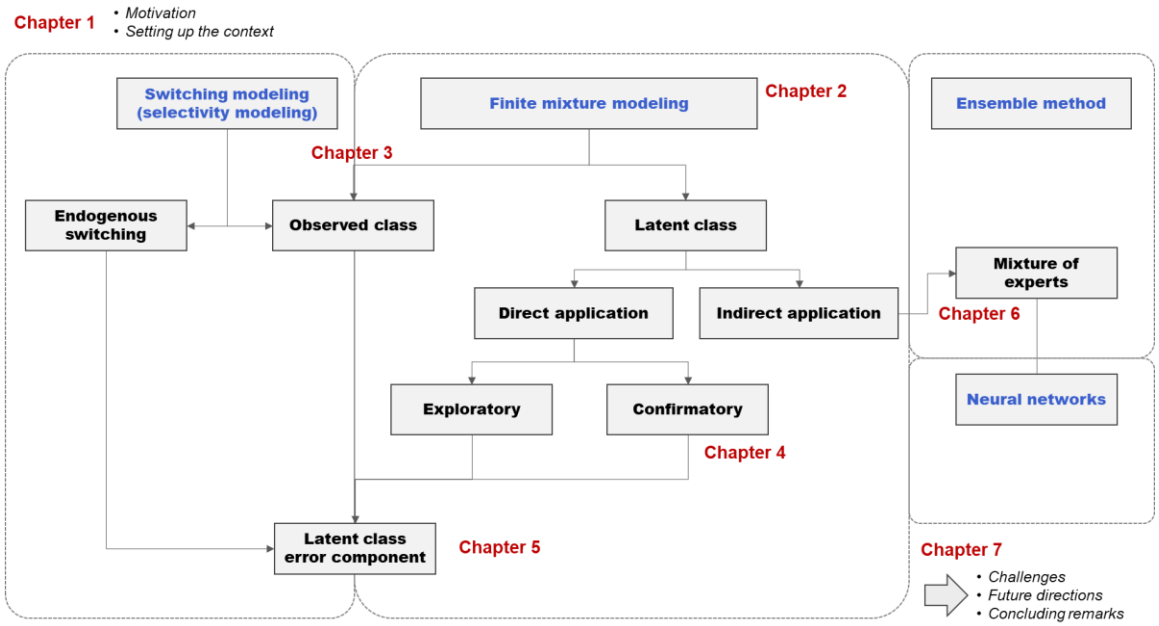


Figure 1-5. Schematic relationships among components of the thesis

Table 1-3. Summary of conceptual contributions and methodological novelties

	Conceptual contributions	Methodological novelties
Ch2 & Ch7	• The first comprehensive review of heterogeneity and mixture modeling in transportation • Develops a typology of heterogeneity in travel behavior/demand • Examines key elements of mixture modeling and thus portrays the general usage of the method • Discusses potential technical advances • Presents critical issues and challenges of the finite mixture approach that have not been discussed	• Applies topic modeling to papers of specific interest in the Scopus database (as opposed to typical applications in which topic modeling is applied to any type of study in a broad domain)
Ch3	• The first conceptual connections/ comparisons among three finite segmentation models: deterministic, endogenous switching, and latent class • Extends the concept of finite mixture/ segmentation	• Discusses model-specific usage and implications (performance, interpretations) when applied in the context of vehicle-miles driven (VMD) modeling • Provides complete equations for calculating various treatment effects when a log-transformation is applied (which has not been covered in travel behavior research)
Ch4	• Embraces zero-inflated models under the confirmatory latent class approach • Discusses the usefulness of the confirmatory latent class approach	• The first application that probabilistically decomposes different types of zeros in the context of modeling long-distance trip frequency
Ch5	• The first introduction of the idea of combining latent class and endogenous switching models in the transportation domain • Delineates subtle conceptual differences between latent class models with and without an error structure • Derives marginal effects of the model • Discusses the issue of evaluating latent class models	• The first application of latent class modeling in the context of modeling the willingness to share automated vehicle (AV) rides with strangers and adoption of ridehailing for social-purpose trips • Conducts statistical inference based on (parallelized) bootstrapping • Illustrates the usefulness of the method with scenario analyses
Ch6	• The first study in travel behavior research that introduces the ideas of indirect application of latent class modeling and the mixture of experts (MoE) architecture • Proposes the idea of using MoE as a data-driven tool to identify nonlinear/interaction effects	• The first application of MoE in the travel behavior and choice modeling communities • Demonstrates the approximation abilities of MoE by experimenting with synthetic data

Table 1-4. Summary of selected empirical findings

	Empirical findings (selected)
Ch2 & Ch7	• Found that heterogeneity and mixture modeling have gained popularity over the years in transportation research publications • Identified six subdomains in transportation that use mixture modeling: *discrete choice modeling*, *general behavior analysis*, *crash/safety analysis*, *traffic analysis*, *travel time distribution*, and *electric vehicles* • Summarized types of heterogeneity and related applications in the literature: *variable distributions*, *parameters*, *model specification*, *attribute processing*, *functional forms*, *decision rules*, *causal structure/order*, *constraint/choice set* • Illuminated that supervised learning and unsupervised learning applications tend to have divergent numbers of classes in their respective final solutions; many studies determined such a number qualitatively rather than quantitatively
Ch3	• Urban residents were more sensitive to the availability of transit, whereas non-urban residents were more sensitive to local amenities • The lower-VMD latent class was influenced to drive less when living in more job-dense or better transit-service areas, whereas the higher-VMD class was not significantly influenced by these factors • Propensities associated with residential location choice and VMT generation shared common unobserved factors
Ch4	• Identified profiles of those in the structural zero-trip regime (as opposed to the trip-making regime): they were the oldest and have the lowest household income for both air and car travel • The presence of children acted as a barrier to belonging to the trip-making regime for air travel, but it was a facilitator of doing so for car travel; it was negatively associated with the number of trips in both modes • As distance to airport increased, both entry into the trip-making regime and number of trips were inhibited for air travel, but car travel exhibited the opposite effects
Ch5	• Found significant correlations between unobserved influences on latent segmentation and behavioral processes (for both empirical contexts of modeling the willingness to share AV rides with strangers and adoption of ridehailing for social-purpose trips) • Males, more educated, and those who have used ridehailing services were more willing to share AV rides with strangers • Identified a latent group of people who are "structurally unwilling" to share AV rides with strangers; (none of the tested factors affects their willingness)
Ch6	• Experiments with synthetic data showed that MoE can capture nonlinear/interaction effects without prior knowledge of those effects • MoE identified significant nonlinear effects of time and cost on mode choice in an empirical application • Found significant interaction effects in this empirical application: in particular, travel time interacted with gender and trip purpose; travel cost interacted with gender and seat grade of train • Conventional logit models were substantially improved by re-specifying them on the basis of what MoE learned from the data

CHAPTER 2. HOW WE HAVE USED MIXTURE MODELING

Paper title: *Finite mixture (or latent class) modeling in transportation: Trends, usage, potential, and future directions*

This chapter explores studies using finite mixture modeling in transportation. It starts with related concepts in various disciplines and then examines publication trends by year and subdomains based on topic modeling. The chapter provides a comprehensive review of how transportation studies have used finite mixture modeling, develops a framework encompassing finite mixture modeling and related subjects, and discusses key elements of the framework with various typologies (e.g. a typology of heterogeneity).

2.1 The arena of segmentation, finite mixture modeling, and other relevant concepts in various disciplines

Transportation studies constitute a wide spectrum of research, but a majority of them employ analytic approaches involving some type of statistical modeling. Numerous studies have pointed out some types of *heterogeneity* that could be an important consideration in modeling. For example, unobserved heterogeneity in crash data is a critical issue for modeling in safety analysis, and has thus led to the proposal of various models to deal with it (cf. Lord and Mannering, 2010; Mannering and Bhat, 2014; Mannering et al., 2016). In travel behavior studies, a key avenue of research has addressed how to capture heterogeneity with respect to a given behavioral process (e.g. Brownstone et al, 2000; Ben-Akiva et al., 2002; Greene and Hensher, 2003). The interest in heterogeneity and corresponding modeling approaches has been exponentially growing over the past several decades (as will be shown in Section 2.2).

The mixture modeling framework is extremely flexible, and thus it is a versatile tool for probabilistically and simultaneously taking various types of heterogeneity into account in behavioral modeling. This methodology has been applied to analysis techniques such as cluster analysis (unsupervised learning), regression/classification, and model systems (structural equation modeling). Thus, concepts of segmentation, heterogeneity, and mixture modeling have appeared in an arena of various fields beyond transportation, including (but not limited to) (bio-)statistics, econometrics, psychometrics, machine learning, marketing research, social/behavioral science, and transportation. Here, we briefly and selectively overview concepts and methodologies discussed in various domains that are relevant to the themes.

2.1.1 Multigroup analysis in psychometric models

Naturally, the existence of heterogeneity in human behavioral processes would be of fundamental interest to the discipline of psychology, and indeed, "group differences" or "multigroup analysis" has long been of interest in psychometrics. For instance, Meredith (1964) aimed to answer the question, "Under what conditions is it reasonable to expect that the factor structure inherent in a given set of variables will be invariant over populations?" In psychometric models, the focal point of multigroup analysis is to test multigroup invariance in model elements such as factor loadings, factor covariances, regression paths, and latent factor means (cf. Byrne et al., 1989; Ansari et al., 2000; Byrne, 2013). There have been several methodological approaches to treating heterogeneity in the population. For example, Joreskog (1971) described a procedure of testing for multigroup invariance (when the sample is drawn from several populations). Muthen (1989) introduced multiple indicator multiple cases (MIMIC) analysis as a method for describing heterogeneity.

Although he took another avenue to handle heterogeneity, he also hinted that "An alternative view to homogeneity is that data come from a mixture of populations with their own sets of parameter values. This relates to statistical modeling called finite mixture analysis" (p. 558). However, most multigroup analyses have been based on a certain deterministic group indicator of interest (e.g. gender, age). Early proposals adopting (finite) mixture modeling in psychometric models include Yung (1997) and Jedidi (1997a; 1997b). Yung (1997) proposed finite-mixture confirmatory factor-analysis models; Jedidi et al. (1997a; 1997b) proposed a latent-class version of the multigroup structural equation model (SEM)[5], questioning the basic assumption of a known group indicator.

2.1.2 Market segmentation in marketing research

The concept of *market segmentation* has a long history in marketing research. Since the pioneering introduction of Smith (1956), market segmentation has become a dominant concept in marketing research and practice (Kotler and Armstrong, 2010; Wedel and Kamakura, 2012). Initially, segmentation was mainly performed using two approaches: direct segmentation on the basis of one variable at a time or perhaps the cross-tabulation of two or (seldom) more variables, and cluster-based segmentation, in which an initial cluster analysis was conducted on multiple variables, and then the data were segmented (for further analysis) on the basis of the resulting clusters (Wind, 1978). Both of these approaches are deterministic (manifest) and exogenous: the segment membership of each case is known *a priori*, having been determined outside of the model or process of interest.

[5] This is a so-called finite-mixture SEM. This approach appears to be less well-known in the transportation domain – the authors are aware of only a few studies (Astroza et al., 2019; Allen et al., 2019; Pendyala et al., 2020; Kim et al., in progress).

Finite mixture modeling was an important analytical breakthrough. Since the first application of mixture modeling in the 1970s, it has been considered one of the most influential methodological developments in the field (Wedel and Kamakura, 2012). In the marketing literature, such models have been often called "latent class models"[6] – terminology that is also familiar in the transportation domain. Early works in marketing research, which employed this type of mixture modeling, include DeSarbo and Cron (1988), Kamakura and Russell (1989), and Swait (1994). Particularly, Swait (1994) presented a useful conceptual model that embeds the latent class concept into the discrete choice process formulated by Nobel economist Daniel McFadden.

2.1.3 Model structure and ensemble methods in machine learning

A variety of machine learning techniques have been proposed in the last several decades. Among the myriads of model classes, some are relevant to our themes. As commented in Bishop (2006), if we focus on the mechanism of *decision tree models* (namely, to partition the input space into a set of rectangles and then fit a model for each segment), we may notice that the basic idea for solving the problem is segmentation. In addition, when examining the model structures of latent class models and neural networks, we may recognize that their structures are fairly similar to each other in that there are hidden (or latent) nodes in layers. Indeed, the latent class model can be considered as a particular form of a single hidden layer feedforward *neural network* with MNL activation functions (McFadden, 2001; Vermunt and Magidson, 2003).

[6] According to Green et al. (1976), the term "latent class" seems to stem from the early work of Lazarsfeld (1950).

Lastly, one particular relationship pertains to a modeling approach (or model architecture) rather than to a single model per se. Considerable effort has been expended to improve the performance of models in the machine learning domain (cf. Kotsiantis et al., 2006). A particular approach is to combine models (ensemble methods). Loosely speaking, this approach aims to predict outcomes based on multiple (local) models instead of a single (global) model. There are various ensemble methods (cf. Rokach, 2010); one such method is the so-called *mixture of (local) experts* (MoE) proposed by Jacobs et al. (1991a). The basic idea is to split the input space into homogeneous regions, with different *experts* (i.e. *models* or also called *learners*) being "responsible for" (i.e. operating in) the different regions (Masoudnia and Ebrahimpour, 2014; Baldacchino et al., 2016). This approach is called "divide-and-conquer" from a problem-solving perspective (which is comparable to the concept of market segmentation). Then, results from the different experts are combined by a *gating network* (usually employing the so-called softmax function – known as multinomial logistic regression in other domains – which functionalizes the membership models of latent class models in most studies).

Recalling the concepts of finite mixture modeling and market segmentation, these approaches are not far from each other. That is, although some details in application and context may not be identical, the basic idea for solving the problem is fairly similar. We can translate "homogeneous input spaces" into "latent classes" or "market segments", "gating network" into "membership/segmentation model", and "local expert/learner" into "class-specific outcome model". This will be covered in CHAPTER 6 in detail. Hence, the mixture modeling framework of interest in this study blends fundamental conceptual (e.g. marketing research) ideas with a promising analytic approach. This accordingly

suggests that the methodology has great potential to improve the interpretability as well as the performance of the models to which it is applied.

2.2 Landscape and trends in transportation

2.2.1 Methodology

Although studies of research trends are not uncommon, only a few attempts to identify such trends through modeling can be observed in the transportation field. This has happened recently (2016-2020), since text mining and/or topic modeling has started to gain attention in the field. As shown in Table 2-1, four out of six studies used resources of the Transportation Research Board (TRB), including compendia of the Annual Meeting and the *Transportation Research Record* journal. A critical difference of this study from others is that this study specifically targets selected papers sharing a theme of interest, whereas others pooled *all* articles in the target sources. Hence, the previous studies identified macroscopic topics/trends in transportation at large, whereas this study aims to uncover topics/trends with respect to papers focusing on our interest (applications of latent class or finite mixture modeling).

First, we find a large volume of journal articles, to encompass the landscape of transportation literature related to our interest. There are multiple sources of information on research articles (e.g. Google Scholar, Scopus, Web of Science, Transport Research International Documentation, individual journal websites); we use the Scopus database to identify our pool of articles. This study specifically uses the Scopus API since (1) it is one of the few sources publicly available (for non-commercial purposes) and (2) Scopus is a

major source-neutral abstract and citation database[7]. We search for our keywords in three fields: article title, keywords, and abstract. We do not search for keywords in the body of the paper, to focus on papers for which our keywords are highlighted and to avoid irrelevant papers (e.g. terms can be mentioned in the body of the paper in a way that is peripheral to the paper's theme). However, this search strategy could be on the conservative side, because some studies may use the methods of interest without mentioning them in the title, keywords, or abstract. We limit our analysis to the pool of articles published through 2020 (no lower bound)[8]. We directed our search query to focus on major peer-reviewed transportation journals as shown in Table 2-2. The keywords we use are:

- "latent class", "latent segmentation", "endogenous segmentation";
- "mixture model(s)", "mixture model(l)ing", "finite mixture".

We pool the selected papers using the union operation – in other words, our pool contains any papers having at least one of the keywords searched. This list of specified keywords may potentially include less relevant papers (e.g. when keywords appear in the abstract by coincidence) and/or exclude highly relevant papers (e.g. papers in other journals, but involving a transportation application).

[7] For more details, refer to https://dev.elsevier.com/sc_apis.html, https://www.elsevier.com/solutions/scopus
[8] The up-to-2019 data were collected in January, 2020. Data for 2020 publications were collected at the end of December 2020.

Table 2-1. Summary of topic modeling applications to transportation research

Author	Year	Method	Target sources	Number of articles	Period
Das et al.	2016	Latent Dirichlet allocation	TRB annual meeting (compendia)	15,357	2008-2014
Das et al.	2017	Structural topic modeling	TRB annual meeting (compendia)	15,357	2008-2014
Sun & Yin	2017	Latent Dirichlet allocation	22 selected transportation journals	17,163	1990-2015
Boyer et al.	2017	Latent Dirichlet allocation	TRB annual meeting (compendia)	32,090	1998-2016
Kuhn	2018	Structural topic modeling	Aviation incident reports	25,706	2011-2015
Das et al.	2020	Latent Dirichlet allocation; structural topic modeling	*Transportation Research Record*	30,784	1974-2019

Table 2-2. The pool of transportation journals searched

- *Transportation*
- *Transportation Research Part A: Policy and Practice*
- *Transportation Research Part B: Methodological*
- *Transportation Research Part C: Emerging Technologies*
- *Transportation Research Part D: Transport and Environment*
- *Transportation Research Part E: Logistics and Transportation Review*
- *Transportation Research Part F: Traffic Psychology and Behaviour*
- *Transportation Research Record*
- *Transportation Science*
- *Transportmetrica A: Transport Science*
- *Transportmetrica B: Transport Dynamics*
- *Transportation Letters*
- *Travel Behaviour and Society*
- *International Journal of Sustainable Transportation*
- *Journal of Advanced Transportation*
- *Journal of Public Transportation*
- *Journal of Transport and Health*
- *Journal of Transport and Land Use*
- *Journal of Transport Geography*
- *Journal of Transportation Engineering Part A: Systems*
- *Research in Transportation Economics*
- *Transport Policy*
- *Transportation Planning and Technology*
- *Journal of Intelligent Transportation Systems*
- *Accident Analysis and Prevention*
- *Analytic Methods in Accident Research*

2.2.2 Yearly trends

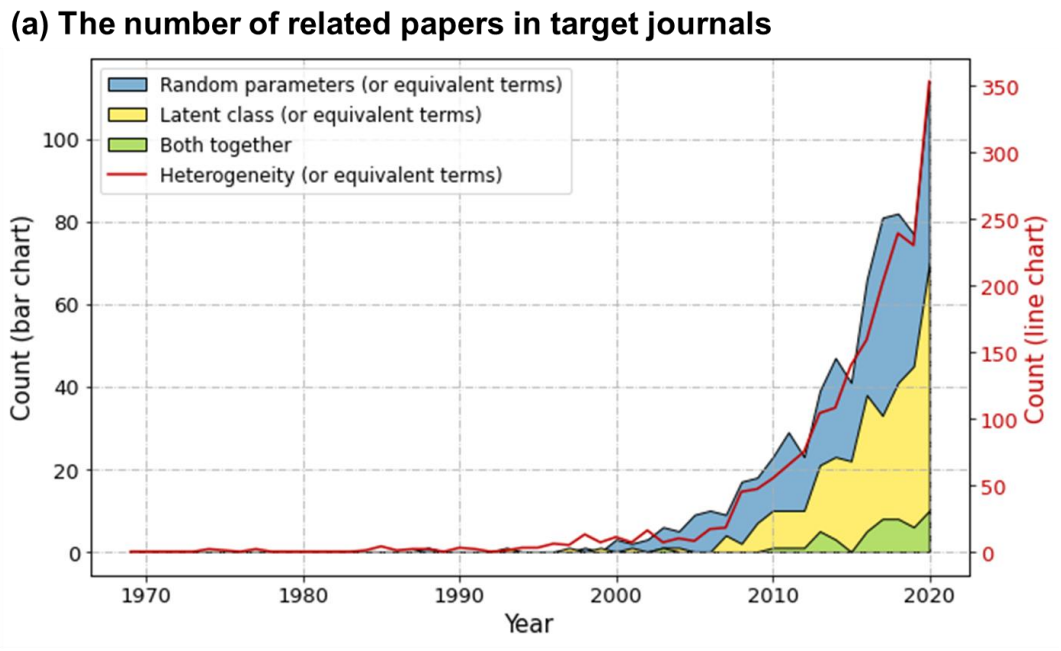
Finite mixture modeling or latent class modeling is gaining in popularity, and such studies appear in a portfolio of transportation journals. In particular, *Transportation Research Part A, Accident Analysis and Prevention, Transportation Research Record,* and *Transportation* are the main transportation journals publishing such papers. We compare the publication trend of our focus (latent class or mixture modeling) with the trends for two adjacent and overlapping topics: heterogeneity and random parameter models. Figure 2-1a plots the annual counts of identified papers having the selected keywords. We can observe a huge surge, beginning around 2008, in the number of publications that are related to the concept of "heterogeneity"[9] (here, the red line graph of "heterogeneity", which is the right y-axis, is superimposed on the area charts associated with the left y-axis). In addition, random parameters and latent class models have also surged. However, the total number of scientific papers across all fields has also been exponentially growing over the same period. Hence, we normalize the number of related papers by the total number of papers in the target journals in each year (Figure 2-1b). For example, as of 2020, the respective shares of papers pertaining to latent class modeling and heterogeneity were 1.5% and 7% out of all papers in the selected journals. The smallness of these shares is not surprising in view of the wide variety of topics and methodological approaches available in the transportation field, but the important point is that the shares of related papers are exhibiting increasing trends. Thus, the figure confirms that identifying and accounting for heterogeneity is becoming an important stream of research. Furthermore, the rising attention given to

[9] For this query, we searched for the keywords "heterogeneity" and "heterogen(e)ous". Since these terms are rather general, substantially diverse papers could be identified.

"random parameters" and "latent class" mirrors increasing publication trends similar to that for heterogeneity studies, thus implying that they might be major tools to be employed. Although the two approaches keep rapidly growing, the random parameter approach seems to be relatively more popular in terms of the number of papers. Some studies mentioned both concepts together, signifying that they compared the two approaches (we will revisit this in Section 2.3.7). Given that concepts of "heterogeneity" or "market segmentation" have been discussed in transportation journals since the late 1970s, it is likely that the spread of newer tools (i.e. random parameters or latent class modeling) has helped increase the number of papers discussing heterogeneity.[10,11]

[10] This spread in transportation research might be partly attributable to some seminal papers about latent class models (Bhat, 1997), mixed logit models (Hensher and Greene, 2003), and comparisons of the two methods (Greene and Hensher, 2003), allowing time for the concepts to be disseminated, absorbed, applied, and for the application papers to be published.

[11] Another interesting manifestation of the penetration of statistical modeling for heterogeneity into the transportation domain can be found in a popular textbook used in many courses on statistical and econometric methods for transportation. Washington et al. added a separate chapter on random-parameter models in the second edition (2010) that was not in the first edition (2003). In their recent third edition (2020), the chapter on random-parameter models has been further elaborated and a chapter on latent class models has been newly added.



Note: Plots (a) and (b) are unstacked area charts. Heights for the three colored regions are all calculated from the zero point on the vertical axis; i.e. the three regions are overlaid.

Figure 2-1. Publication of related papers over the years in target journals: (a) the number of papers, (b) the share of papers normalized by total number of papers

2.2.3 Topic modeling

Latent class or mixture modeling approaches have been used in a variety of applications. To identify and classify research topics that employed the mixture modeling approach, we employ the nonnegative matrix factorization (NMF) technique. NMF aims to find the positive factorization of a given nonnegative matrix (i.e. a matrix with no negative elements; Xu et al., 2003; Shahnaz et al., 2006). Specifically, we begin with the nonnegative word-abstract matrix, $X_{m\times n}$, where *m* is the number of words, *n* is the number of abstracts, and the *ij*th element of $X_{m\times n}$, x_{ij}, is a number between 0 and 1 that represents the value of word *i* to characterizing abstract *j*. The more abstracts that contain word *i*, the less valuable that word is in distinguishing among abstracts, and the smaller x_{ij} will be. We assume there are k latent topics in our pool of abstracts, which are considered as proxies for the (relevant) contents of the papers themselves. The goal is to factorize $X_{m\times n}$ into two nonnegative matrices: a word-topic matrix, $W_{m\times k}$, and a topic-abstract matrix, $H_{k\times n}$. The objective function to be minimized is as follows: $J = \|X - WH\|^2$, where $\|\cdot\|^2$ denotes the squared sum of all the elements in the matrix. Figure 2-2 shows a schematic diagram of NMF. The darker colors indicate greater values. In this example, Abstract 2 is more relevant to Topic 2 and Word 3 is more relevant to Topic 1. Although methodologies take different mathematical forms, the basic idea of NMF is analogous to that of principal component analysis (PCA) or exploratory factor analysis (EFA), which are widely employed in the transportation domain. For example, when applying EFA to attitudinal statements, we aim to find latent psychological constructs (here topics) and estimate the loadings, or associations, of statements (here words) with constructs and with the factor scores of individuals (here abstracts).

When constructing a matrix X, we transform a raw word-abstract matrix into a matrix of tf-idf (term frequency-inverse document frequency) values. This concept is often used in text mining to measure the importance of words in each document (Beel et al., 2016). It works by determining the relative frequency of words in a document compared to the inverse proportion of that word over the entire document corpus (Ramos, 2003); hence the tf–idf value increases proportionally with the frequency of a word's appearance in the document, but is offset by the number of documents in the corpus that contain the word.

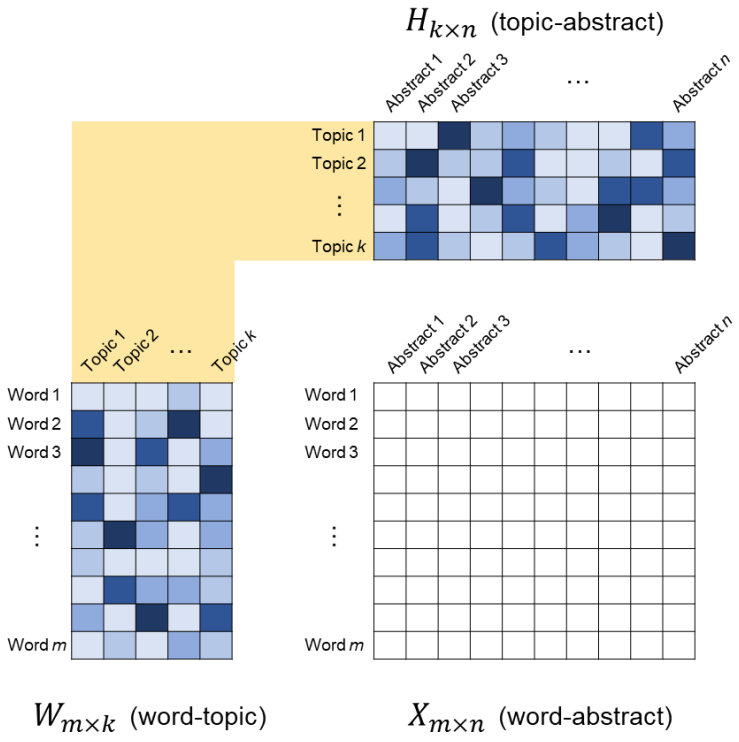
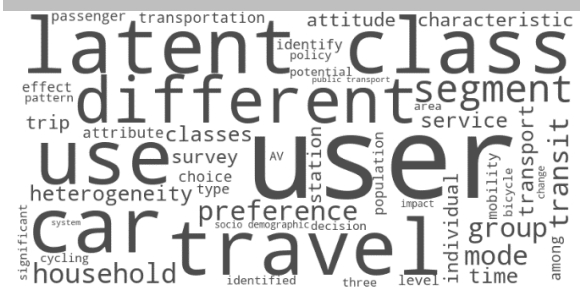
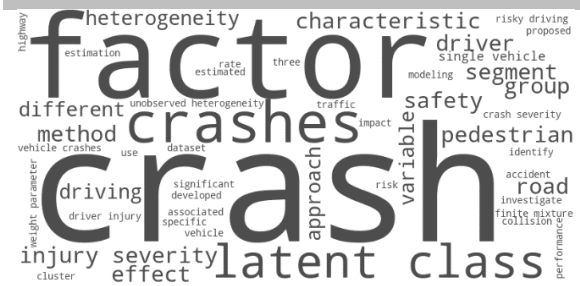
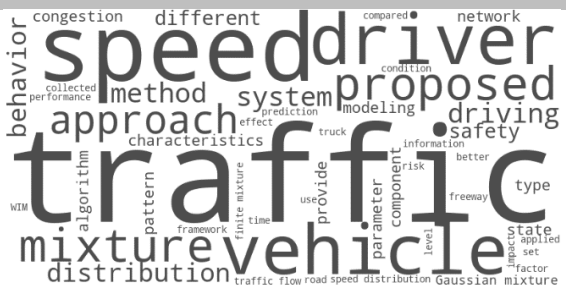
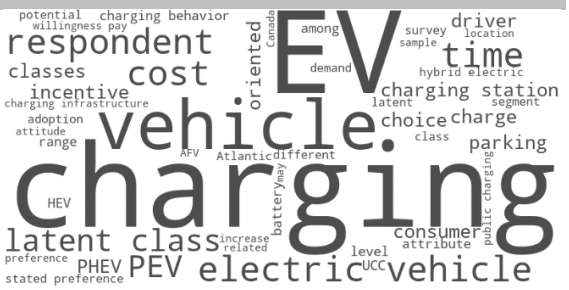


Figure 2-2. Schematic diagram of NMF

Table 2-3. Word clouds for each topic

Topic 1: Behavior analysis (26%)	*Topic 2: Discrete choice modeling (21%)*
	
Topic 3: Crash/safety analysis (17%)	***Topic 4: Traffic analysis (16%)***
	
Topic 5: Travel time distribution (14%)	***Topic 6: Electric vehicles (6%)***
	

We experimented with topic models having various numbers of topics (values of *k*), and ultimately identified six topics among 284 papers of interest. Based on the word-topic matrix, we can identify the top keywords in each topic (Table 2-3). In the pool, studies "under the umbrella" of behavior analysis are the most common (topics 1, 2). Especially, discrete choice modeling is a major community using this methodology. Crash/safety

analysis is the main application domain (17%) using the method, followed closely by traffic analysis (16%). Each of the topics involves general issues of heterogeneity in analyses. For behavior analysis, it has been emphasized that human behaviors are heterogeneous. In addition, many studies have commented on heterogeneity in safety (Mannering and Bhat 2014; Mannering et al., 2016) and travel-time related data (e.g. Kim and Mahmassani, 2014; Yang and Wu, 2016; Zou et al., 2017). In a nutshell, mixture modeling has become a major methodological tool, having diverse applications in transportation. We will examine the literature and its keynotes in greater detail in later sections.

2.3 How have we used mixture modeling? Diving into each key element

This section explores how transportation studies have used the finite mixture modeling framework. For more detailed and specific discussions, we put more emphasis on transportation studies, further specializing in travel demand and behavior analysis. The following subsections successively discuss the key elements of mixture modeling: the types of heterogeneity considered and the types of approaches used (confirmatory versus exploratory), the types of problem to which mixture modeling is applied, the membership function, the outcome model, selection of the number of classes, model comparisons, and software and estimation approaches.

2.3.1 Type of heterogeneity

A number of different kinds of heterogeneity have been identified, particularly relevant to the field of behavior modeling. In this thesis, we classify heterogeneity as pertaining to one or more of the following aspects of a behavioral model: *variable distributions*, *parameters*, *model specification*, *attribute processing*, *functional form*,

decision rule, *causal structure/order*, and *constraint/choice set*.[12] As shown in Figure 2-3, we suggest that these types of heterogeneity (rectangles) arise from four key sources (rounded rectangles): data, parameters, function, and conceptualization. The types of heterogeneity are briefly summarized in Table 2-4. Note that they are not mutually exclusive: some types can be obtained as special cases of others (but we keep them separate because they are typically treated separately in the literature), and multiple types of heterogeneity can appear in the same model. Below, we will examine each of the representations of heterogeneity found in the literature.

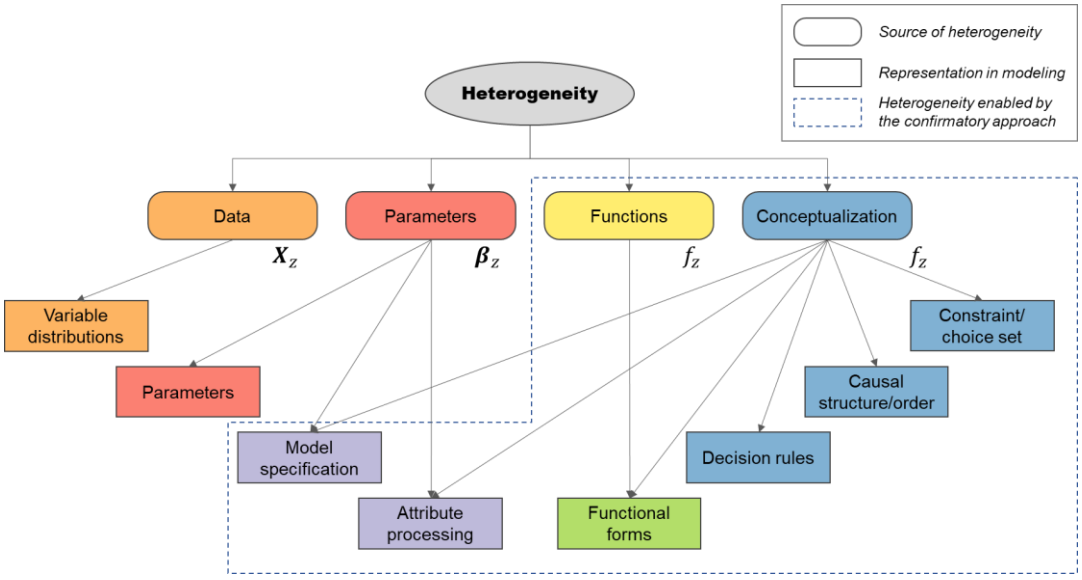


Figure 2-3. Typology of forms of heterogeneity addressed by the mixture modeling framework

[12] Other papers have discussed various kinds of heterogeneity. Multiple typologies of heterogeneity could be possible from different points of view; for example, heterogeneity can be discussed with respect to its source in a structural equation model context (Ansari et al., 2000), or the way it is treated (e.g. Kim and Mokhtarian, 2018). Note that the typology of heterogeneity presented here has built upon discussions in several studies cited in this thesis (e.g. Gopinath, 1995; Walker, 2001; Walker and Ben-Akiva, 2011; Vij et al., 2013; Hess, 2014). For example, Gopinath (1995) discussed that unobserved heterogeneity can stem from decision protocols, choice sets, and taste variations; Hess (2014) commented that attribute processing and decision-rule heterogeneity can be treated with mixture modeling. Some other studies commented on selected types of heterogeneity in the context of discussing the usefulness of mixture modeling. To our knowledge, discussion of the extensive list of aspects of heterogeneity presented here has not previously appeared in a single place.

Table 2-4. Descriptions of types of heterogeneity

Heterogeneity	Description
Variable distributions	A fundamental source of heterogeneity is the data distribution. There are subgroups in the population (no matter how we define them) and each segment has its own distributions of characteristics. For example, we may classify individuals into urban and non-urban residents; and then urban and non-urban resident segments may have different distributions of characteristics (e.g. gender, income, age, attitudinal dispositions). Mixture modeling can be used to probabilistically classify individuals on the basis of selected characteristics, or indicators.
Parameters	Individuals have heterogeneous sensitivities to, or preferences for, factors associated with the outcome. The most popular form of heterogeneity explored in transportation is parameter heterogeneity (referred to as "taste heterogeneity" or "taste variation" in the choice modeling context).
Model specification/ Attribute processing	Each class has a different model specification with respect to the set of attributes to be considered in the model. For example, a certain factor may not be considered in the decision-making of a certain segment.
Functional form	Each segment follows a different data generation process which is represented by a different functional form (not just different parameter values *per se*). For example, for some segments the choice may follow a multinomial logit form best, while for others a cross-nested logit form is more appropriate.
Decision rule	Each segment has a different behavioral mechanism, beyond functional forms *per se*. For example, some people may make the choice that gives maximal utility, whereas others may make the choice that gives minimal regret.
Causal structure/ order	The causal relationships among behavioral indicators may differ across segments. For example, some people may determine vehicle ownership and then residential location, whereas others may do so in the reverse order.
Constraint/ choice set	Each segment has its own constraints on decision-making. Or each segment may have a different choice set. For example, a certain segment would never consider bicycle for a mode choice.

We reserve the discussion of *heterogeneity in variable distributions* to Section 2.3.3.[13] Turning to the second type, most applications of mixture modeling

[13] This is because heterogeneity in variable distributions has a distinct nature. Other types of heterogeneity

highlight heterogeneity in ***parameters***. The idea is that individuals have heterogeneous sensitivities to, or preferences for, factors. This is the most popular form of heterogeneity explored in the transportation literature (taste heterogeneity/variation in the choice modeling context, and in an even broader perspective, "structural change" in time-series analysis is a type of parameter heterogeneity). Hence, although many studies do not explicitly say which type of heterogeneity they are investigating, a majority of them implicitly examined parameter heterogeneity. Early works in travel-related choice modeling include modeling mode choice behaviors (Bhat, 1997; Walker, 2001) and route choice behaviors (Greene and Hensher, 2003). Many studies have provided conceptual and empirical evidence regarding how treating parameter heterogeneity (via mixture modeling) is helpful. In particular, in behavior analyses, it is well suited for explaining complicated human decisions. Several studies tried to understand behaviors as arising from endogenously segmented travel/life-related styles (e.g. Walker and Li, 2007; Vij et al., 2013; Vij and Walker, 2014; Prato et al., 2017). Hence, the latent class modeling framework has been applied to a variety of subjects such as (but not limited to) mode choice (Wen et al., 2012; Ma et al., 2015; Vij et al., 2017; Qin et al., 2017; Saxena et al., 2019), flight choice (e.g. Wen and Lai, 2010; Seelhorst and Liu, 2015; Araghi et al., 2016), vehicle ownership (e.g. Anowar et al., 2014; Kim and Mokhtarian, 2018), vehicle type (e.g. Beck et al., 2014; Hackbarth and Madlener, 2016; Ferguson et al., 2018), location choice (Olaru et al., 2011; Fatmi et al., 2017), and some types of binary decisions regarding, or interest in, services/activities (e.g. Wen et al., 2016; Lin et al., 2017; Wolbertus and Gerzon, 2018;

deal with how behavioral outcome models differ across segments (related to "supervised learning"), whereas the variable distributions are about distribution differences per se (related to "unsupervised learning").

Lin et al., 2018). Numerous safety analyses have also benefited from incorporating parameter heterogeneity since it has been claimed that unobserved heterogeneity is present in crash data. For example, latent class modeling has been an important tool for treating parameter heterogeneity in modeling crash/injury severity level (e.g. Xie et al., 2012; Eluru et al., 2012; Chu, 2014; Behnood et al., 2014; Yasmin et al., 2014; Shaheed and Gkritza, 2014; Adanu et al., 2018; Fountas et al., 2018; Yu et al., 2019) and crash count (e.g. Zou et al., 2013; Yasmin and Eluru, 2016; Park et al., 2016).

Studies of *attribute processing* (or information processing) strategies in choice modeling have also employed the mixture modeling framework (e.g. Hensher and Greene, 2010; Hess et al., 2013b, Hensher, 2014). In this type of heterogeneity, each class is conceived of considering (or attending to) a different combination of attributes in the choice process (i.e. when there are p attributes, 2^p combinations are possible and thus there can be as many as 2^p classes). Loosely speaking, this can be also considered as a branch of parameter heterogeneity in that different model specifications imply different parameter values (e.g. individuals in one segment take into account built environment attributes and thus have non-zero weights on those attributes, whereas individuals in another segment do not, and thus give them zero weight). Or, as Hess et al. (2013b) and Hensher et al. (2013) pointed out, attribute non-attendance and regular taste heterogeneity could be confounded while employing the mixture modeling framework. In any case, the latent class modeling framework (generally with constant-only membership models) has served as a formal specification of models considering attribute attendance. Collins et al. (2013) applied various specifications of attribute processing models (formulated with mixture modeling) to flight choice behaviors. Hensher et al. (2013) modeled route choice with a latent class

structure that allows heterogeneity in attribute processing rules (e.g. full attribute attendance, attribute non-attendance, aggregation of common-metric attributes). Pan et al. (2019) modeled electric vehicle (EV) charging choice with a latent class model accounting for attribute non-attendance. They reported that, with respect to the Akaike Information Criterion (AIC), the model was better than a counterpart that assumed the same model specification across segments.

We can also imagine that different population segments may have different *a priori model specifications* apart from variations in attribute attendance. Even if an attribute is a factor for all segments, for some conceptual reasons, it may be formulated in different ways across segments (e.g. log-transformed, powered, or untransformed). For instance, Sun et al. (2012) conjectured three types of risk attitudes (risk avoider, risk taker, and risk neutral) related to route choice, and then constructed three classes having different utility functions to reflect such attitudes. El Zarwi et al. (2017) modeled the adoption and diffusion of new transportation services (e.g. Uber/Lyft) and formulated three types of technology adopters (innovator/early adopters, imitators, and non-adopters). Based on the technology diffusion literature, the systematic utility of the innovator class is formulated as a function of characteristics of the decision-maker and attributes of the new technology; the utility of imitators is a function of social influence as well as the aforementioned attributes; and the utility of the non-adopter class is formulated only with constant terms.

A rationale behind heterogeneity in *functional forms* is that there could be alternative mathematical representations/functions of how segments generate outcomes (aside from different sensitivities). For example, Koutsopoulos and Farah (2012) applied latent class modeling to car-following behaviors. They assumed that there are three

possible (latent) states for vehicles (acceleration, deceleration, and do nothing) and that each state follows different models (lognormal distributions for the first two classes and normal distribution for the do-nothing class, in this study). Some studies posited that there are random and regular user groups of particular activities and applied latent class modeling (to shopping, Bhat et al., 2004 and use of EVs, Kim, Yang, Raouli, Timmermans, 2017). Bhat et al. (2004) modeled two different functional forms (proportional hazard accommodating the effect of observed and unobserved characteristics for the erratic-shopper class and non-parametric hazard for the regular-shopper class). Kim et al. (2017) modeled EV inter-charging time with the exponential distribution (for the random user group) and the Erlang-2 distribution (for the regular user group). Or, we can consider conventional zero-inflated models (Lambert, 1992) as a particular form of latent class modeling addressing heterogeneity in functional forms. In this case, one class follows a usual behavioral generation process (e.g. represented by probit, Poisson or negative binomial), but another class structurally generates zero instances or amounts of the behavior (cf. CHAPTER 4). Ma et al. (2016) modeled injury severity with a two-class model: one assuming the multinomial logit and the other assuming the ordered logit (OL) functional form. They called the method a "hybrid finite mixture model" since the model includes two types of models. After comparing with a mixture of multinomial and mixture of ordered logit models, they (p.70) concluded that "…the FMMNL [finite mixture MNL] model provides a very flexible modeling structure but the interpretation of the factors is difficult, whereas the FMOL [finite mixture OL] model is simpler to interpret but less capable of mining complicated patterns of influence of factors. The proposed HFM [hybrid

finite mixture] model exhibits an appropriate balance between modeling flexibility and interpretation difficulty."

Particularly in the choice modeling context, there have been discussions of various *decision rules* (decision protocol or process heuristics) that could potentially govern human decisions (cf. Gopinath, 1995). Thanks to the seminal work of McFadden (2001) connecting economic theory (random utility maximization, RUM) with a statistical model, RUM has served as the dominant decision rule paradigm of choice modeling. However, due to the complexity of the human decision process, it has been argued that human behaviors cannot be explained solely by RUM. Hence, various alternatives have also been proposed such as lexicography (Payne et al., 1992) and random regret minimization (Chorus and Timmermans, 2008)[14]. Although numerous studies looked for the best decision heuristic given the empirical context, several studies posited that multiple decision rules can exist in the sample and endogenously clustered the sample into segments that follow different decision rules with the aid of a mixture modeling framework. Hess et al. (2012) provided four case studies, with each case study employing a mixture modeling framework to combine two different kinds of decision rules: (1) RUM versus lexicography, (2) RUM versus reference-dependent choice, (3) RUM versus elimination by aspects, and (4) RUM versus random regret minimization. The study demonstrated that the mixture modeling framework is flexible enough to accommodate behavioral process heterogeneity. Srinivasan et al. (2009) modeled mode choice under the assumption that individuals follow either utility maximization or disutility minimization decision rules. In this application, a

[14] For more details about decision rules and relevant discussions, please refer to Ben-Akiva and Lerman (1985).

sizable majority (68%) turned out to follow the disutility minimization rule. Zhang et al. (2009) utilized latent class structures to examine heterogeneous household decision-making mechanisms. They built three separate two-class models that contain combinations of two decision-making mechanisms out of three: multi-linear utility, maximum utility, and minimum utility models. Boeri et al. (2014) conducted a choice experiment obtaining preferences among alternative traffic calming projects. They constructed two latent classes: one following random utility maximization and the other following random regret minimization. In their application, the share of the utility maximization class was dominant (57.3% versus 42.7%). Hensher et al. (2018) examined two process heuristics in the discrete choice modeling context: extremeness aversion and extended expected utility attribute transformation. Cranenburgh and Alwosheel (2019) used a latent class model with three classes of decision rules: random utility maximization, random regret minimization, and random.

A large body of studies in transportation investigates *causality* or *structural relationships* among variables. Most of the previous studies have assumed a homogeneous model structure for the sample, but we can expect there could be subsamples that follow different structural relationships. Here, mixture modeling renders a framework for incorporating such an assumption in modeling. Chakour and Eluru (2014) employed a mixture modeling framework to segment two types of decision order: choosing train station first and then access mode (station-mode), and choosing access mode first and then train station (mode-station). In particular, the second choice is specified as a function of exogenous variables that include attributes of the first choice. They found that, based on modeling segment membership as a function of work status, walk time to closest station

and departure time, the mode-station choice segment consisted of 36% of the sample. Angueira et al. (2015, 2019) analyzed interrelationships between vehicle type choice and distance traveled. With the aid of a latent segmentation approach, two segments (vehicle type choice affects distance traveled; distance traveled affects vehicle type choice) were identified. The first segment consisted of 89% of the sample. Anowar et al. (2019) constructed a latent class joint choice model (mode choice and departure time) to understand university students' behavior. They assumed two possible decision processes: mode choice first and then departure time choice (class 1), and the reverse choice order (class 2). Under this confirmatory latent class setting, they found that the model outperformed the baseline models (separate models assuming each order homogeneously), and that the departure-time-choice-first group has a higher share (64.65%) in the sample. Astroza et al. (2019) explored heterogeneous structural relationships among residential location, vehicle ownership, and use of shared mobility. In other words, the study assumed that each segment has a different structural relationship among the three individual choices in the bundle (e.g. residential location affects vehicle ownership for one segment, and vice versa for another). In their analysis, a majority of the sample (53%) presented the following structural relationships: residential location affects vehicle ownership and both decisions affect the use of shared mobility.

Some groups of data (in particular, people) may have ***constraints*** on producing certain outcomes. An example is different ***choice sets***. Ben-Akiva and Boccara (1995) is an early application in marketing that used mixture modeling for heterogeneity in choice sets. Some people may have the options of choosing car, public transit, or bike, whereas bike is not a feasible option for others. In the transportation literature, very few studies

have explored this type of heterogeneity. In behavior studies, some scholars have linked preference heterogeneity with the choice set (Vij and Walker, 2014). For example, Vij et al. (2013) discussed choice set generation related to modality styles. In their experiment, a three-class latent class model with heterogeneous choice sets across segments outperformed counterparts with uniform choice sets as well as mixed logit with error components, with respect to Bayesian and Akaike Information Criteria (BIC and AIC).

2.3.2 Confirmatory versus exploratory approaches

As aforementioned, one of the main reasons why mixture modeling has become popular is that it could be a suitable methodological approach for examining a number of types of heterogeneity. To capture *certain* types of heterogeneity with mixture modeling, however, analysts need to customize the model structure. In this regard, it is helpful to introduce two types of approaches using mixture modeling: *exploratory* and *confirmatory*[15] (Table 2-5).

Most typical finite mixture models are considered **exploratory** (Hoijtink, 2001; Laudy et al., 2005). A key question expected to be addressed by the exploratory approach is: ***how many latent classes are there and what types/characteristics of classes are there, given the data?*** To answer this question, the number of classes is empirically explored and determined. For example, analysts do not start the modeling with a statement such as "we think there are three classes in this empirical dataset". As well, there are no class-specific hypotheses imposed on the classes. This is an obvious consequence of not having a prior

[15] "Exploratory" and "confirmatory" are properties of general modeling *procedures* in scientific research, not of a particular *methodology* itself. The two terms have been widely used in psychology/psychometrics. Wagenmakers et al. (2012) provides useful philosophical discussions focused on psychology studies, but they are also relevant to other fields of scientific research.

belief about the classes. Hence, the general procedure for the exploratory approach is as shown in Figure 2-4. First, select a functional form given the data, and determine model specifications. Then, test numbers of classes ranging from one to *K* and see if there is a satisfactory solution (Section 2.3.6 will cover this). A premise is that we do not have *a priori* hypotheses/knowledge about the latent classes, and are thus exploring possible solutions.

On the other hand, it is relatively less often characterized as such, but some studies have employed the **confirmatory** approach in a mixture modeling context; this approach seems to have originated from psychometrics (cf. Hoijtink, 2001; Finch and Bronk, 2011). As discussed by Hess (2014) in a choice modeling context, confirmatory mixture modeling imposes certain *a priori* constraints, not only on the number of segments, but also on their distinctive natures. A key question expected to be addressed by the confirmatory approach is: ***are the hypotheses on the number of classes and their class-specific assumptions supported by the data?*** To answer this question, an analyst designs a behavioral mechanism or mathematical representation of each class based on prior assumptions, then tests the model upon the empirical data to see if the hypotheses are corroborated by the data (see the modeling flow chart in Figure 2-4). For example, suppose we have *K* model candidates and then want to identify how many and what type of individuals follow certain models. Then, researchers can combine the *K* models with mixing proportions (which could be just constants or a function of covariates) and then estimate such mixing proportions as well as the parameters of the *K* models. In this case, the number of classes is pre-determined; the purpose of using mixture modeling is then, as aforementioned, to identify the shares of classes and (potentially) to characterize the classes with covariates.

It is worth noting two caveats about this distinction. First, the two approaches are on a continuum rather than a dichotomy.[16] The two approaches share common characteristics in practice. Both are partly *theory-driven* in that theories and conceptual validity are involved in (1) model specification and (2) the decision on the final solution. As well, both are partly *data-driven* in that model validity is subject to the empirical data. Second, the word "exploratory" may not be perfect in the context of latent class modeling. The main reason is that the exploratory approach is not, in fact, *fully* exploratory. In principle, a fully exploratory approach should be able to search all combinations of types of heterogeneity (as discussed in the preceding section) with numbers of classes. For example, in a mode choice problem, it should explore multiple combinations of decision rules (e.g. random utility maximization, random regret minimization), functional form (e.g. multinomial logit, probit, mixed logit), and so on with a hyperparameter governing the number of classes. This is an almost infeasible search problem. However, while assuming that many types of heterogeneity are not operative, the conventional approach still has an exploratory nature in terms of (1) the philosophy that the number of classes is empirically determined rather than assumed in advance, and (2) the fact that no class-specific constraints/assumptions are imposed on model parameters.

Then, why do we need to discuss this distinction? Why don't we just apply the exploratory approach? The benefits of this conceptual distinction and understanding the

[16] Similar arguments can be found in the psychology/psychometrics literature: "…the fact [is] that almost no psychological research is conducted in a purely confirmatory fashion… psychological studies can be placed on a continuum from purely exploratory, where the hypothesis is found in the data, to purely confirmatory, where the entire analysis plan has been explicated before the first participant is tested" (Wagenmakers et al., 2012, p. 633) and "Most uses of 'confirmatory' factor analyses are, in actuality, partly exploratory and partly confirmatory in that the resultant model is derived in part from theory and in part from a respecification based on the analysis of model fit" (Gerbing and Hamilton, 1996, p. 71).

usefulness of the confirmatory approach are twofold. First, the confirmatory approach is a way of incorporating (human) expert knowledge in mixture modeling (and probably any other modeling as well). As a result, we may have more meaningful results accruing from more reasonable assumptions. In addition, from the technical viewpoint, we may reduce the search space of model configurations. This is an important merit because, in reality, we cannot search every possibility. Second, as indicated above, we may notice that the usual latent class model is not truly exploratory under the extended typology of heterogeneity. In practice, we do not impose class-specific hypotheses regarding functions and conceptualizations (see Figure 2-3) and thus the modeling can only find parameter heterogeneity. On the other hand, the confirmatory approach is customized to test other types of heterogeneity across classes. In other words, as shown in Figure 2-3, the blue dashed line presents the expanded territory of heterogeneity addressed by the confirmatory approach. We suggest that this is the key virtue of the confirmatory approach, and thus highlights a unique benefit of discrete mixtures as opposed to continuous mixtures. However, recalling the earlier argument that exploratory and confirmatory approaches are two ends of a continuum, any study may contain both characteristics and the placement of the study along the continuum is context- and application- dependent instead of simply constituting a certain approach.

Table 2-5. Summary of exploratory and confirmatory approaches under mixture modeling

	Exploratory approach	***Confirmatory approach***
Common characteristics	• Partly *theory-driven*: theories and conceptual validity are involved in (1) model specification and (2) the decision on the final solution • Partly *data-driven*: model validity is subject to the empirical data	
Key question	How many latent classes and what types/characteristics of classes are there, given the data?	Are the hypotheses on the number of classes and their class-specific assumptions supported by the data?
Number of classes	Empirically determined	Hypothesized in advance
Class-specific assumptions	Not imposed	Behavioral mechanism or mathematical representation of each class is designed based on prior assumptions
Type of heterogeneity	Parameter heterogeneity	Any type of heterogeneity
Search space of model configurations	Extremely large (even infinite, if truly exploratory)	Search space is reduced based on knowledge/assumptions of the analyst

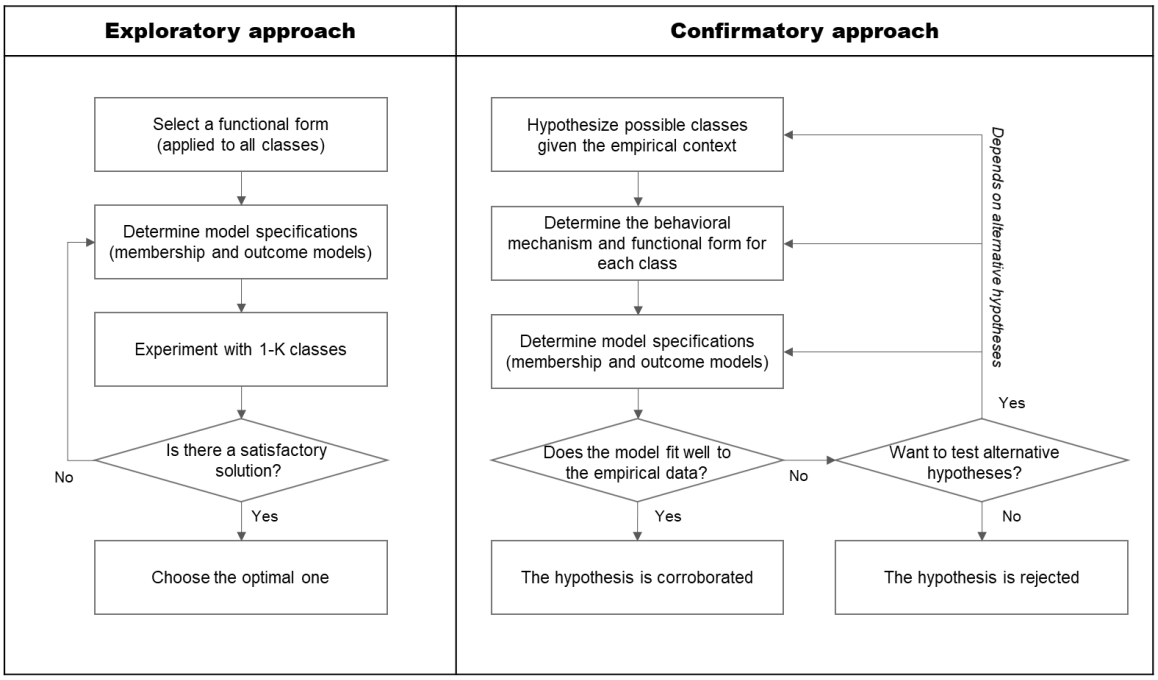


Figure 2-4. Modeling flow charts of the two approaches

2.3.3 Types of problem

Finite mixture modeling has been applied in various types of contexts. In particular, the problem type shapes the narrative flow and the characteristics of the study. From a broad and practical viewpoint, types of problems that mixture modeling aims to address include (in machine learning parlance) ***supervised learning*** (***classification/ regression***) and ***unsupervised learning*** (***cluster analysis***) problems. Among the identified pool of papers, about 65% are supervised learning problems and the rest (35%) deal with unsupervised learning. Although the underlying frameworks are similar, the types of results obtained from the methodological framework are not necessarily identical.

For applications of classification and regression, a majority of studies are particularly interested in the different functional forms of outcome models or distinct taste or sensitivity to factors exhibited by different segments. In other words, such studies aim to simultaneously estimate segmentation (i.e. clustering) and outcome models. Supervised learning can be subclassified by the type of target or outcome variable. The distinction between classification and regression is the standard one: classification deals with categorical outcomes (including binary, nominal, ordinal variables; such models are often referred to as *latent class choice models*, particularly in the choice modeling community), whereas regression deals with continuous or count outcomes (often referred to as *latent class regression models*). In other words, the type of target variable determines the type of outcome model (we will revisit this in Section 2.3.5). For either case, a common narrative flow in the papers is to describe how the latent segments are distributed (often with profiles) and how the outcome functions differ across segments.

In an unsupervised learning context, the purpose of using mixture modeling is to cluster the data with respect to indicators. It does not aim to simultaneously produce different behavioral outcome models for each segment. In this case, the clustering method is often referred to by some specific terms including *Gaussian mixture models (GMM)*, *latent class cluster analysis*, *latent profile analysis*, and so on. From one perspective, cluster analysis based on mixture modeling can be considered as a special case of latent class regression, where the outcome equation(s) have only constant term(s).[17] Unlike other popular clustering methods (e.g. distance-based clustering such as conventional k-means or hierarchical clustering), clustering based on mixture modeling is based on the *distribution* of data rather than distances between data points, and has a probabilistic nature in that observations only have probabilities of belonging to each segment, rather than being deterministically assigned to one and only one segment. In essence, constant-only regression models are fit for different classes as identified through segmentation variables.

Although clustering is embedded in classification/regression problems when using mixture modeling, there are some differences in practical applications. First, studies using cluster analysis do not necessarily model or discuss how the different segments have different functions to explain/predict certain target variable(s), whereas that is usually the main focus of classification/regression problems. In other words, for studies classified as cluster analysis, identifying segments is a key goal (often *the* key goal) in its own right.

[17] Following the notation of Eq. (1.1), the basic formulation of latent clustering can be expressed as $f(y) = \sum_{z=1}^{Z} P(z|\boldsymbol{W}) \prod_{h=1}^{H} f_z(y_h|z)$, where h indexes the number of indicators to be clustered (like dependent variables, $h = 1 \dots H$). For example, a common GMM is $f(y) = \sum_{z=1}^{Z} P(z) f_z(y|z)$ where f_z are normal densities (i.e. a constant-only regression with a normal error term); a common latent class cluster analysis on H binary indicators may be expressed as $f(y) = \sum_{z=1}^{Z} P(z|\boldsymbol{W}) \prod_{h=1}^{H} f_z(y_h|z)$ where $f_z(y_h|z)$ are constant-only binary logit models.

Another difference concerns the general tendency of the number of classes. Classification/ regression problems tend to end up with fewer classes than unsupervised learning problems do, for practical reasons (e.g. estimation and interpretability). This will be covered in Section 2.3.6.

However, it has been also common to follow a "two-step approach" in the literature. One type of two-step approach, adopted by a few studies, first deterministically segments the sample and then applies latent class modeling to the multiple subsamples. The implicit assumption is that each subsample may have different latent classes and corresponding behavior functions. In this case, most applications choose a certain number of classes (two-class is the most popular; we also revisit this in Section 2.3.6) and then constrain all subsamples to have the same number of classes (e.g. Olaru et al., 2011; Adanu et al., 2018; Lin et al., 2018; Potoglou et al., 2020; Choi and Mokhtarian, 2020). We speculate that this is for convenience of comparison across subsamples.

Another type of two-step approach (Figure 2-5) consists of uncovering latent segments with the aid of some type of mixture modeling, and then constructing separate outcome models for each segment.[18] For example, Machado et al. (2018) identified six latent transit customer types based on their perceptions and then built a separate structural equation model for each class to find heterogeneous behavioral models. Piendl et al. (2019) classified shipment types into four latent classes and then estimated class-specific shipment size choice models. Ahmed et al. (2020) clustered workers based on socio-demographics

[18] Authors did not always disclose details of how segment-specific models were estimated in a separate second step when segment membership is unobserved, but it is common to assign each case to the highest-predicted-membership-probability segment, whether correcting for misclassification bias or not (see Bakk et al., 2013; Bolck et al., 2004).

via latent class cluster analysis and then applied hierarchical logit models of destination choice to the resulting four classes.

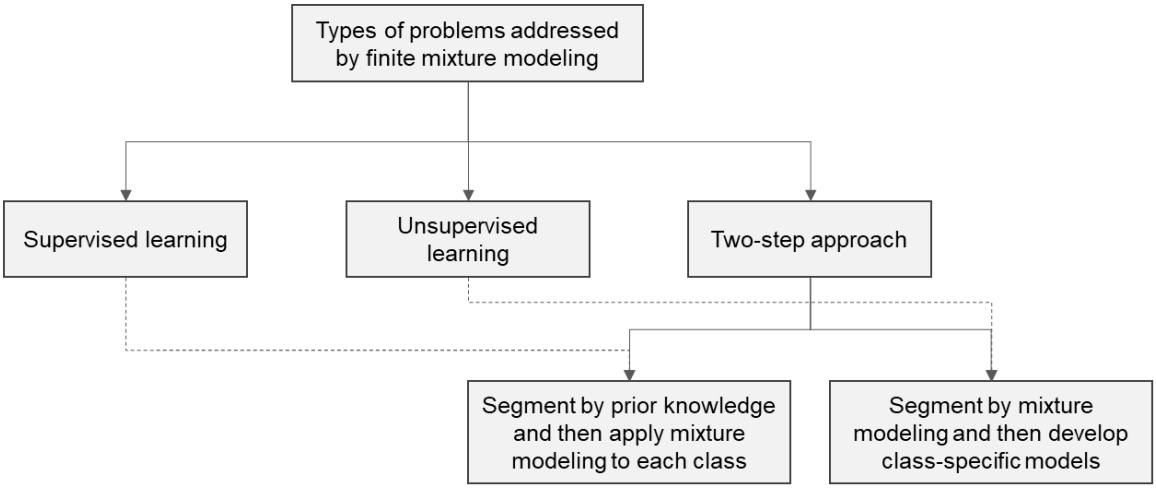


Figure 2-5. Types of problems addressed by using finite mixture modeling

Typically, latent class choice/regression models estimate both class membership and class-specific outcome models simultaneously, and thus the latter type of two-step approach may not be efficient or may have an inferior goodness-of-fit. The latter can happen because when we simultaneously estimate membership and outcome models, latent classes are identified *with respect to* the outcome of interest via maximizing log-likelihood during estimation, whereas clustering in the two-step approach finds segments based on the distribution of variables and thus does not necessarily help explain the outcome better. However, this two-step approach could potentially bring some practical benefits. First, it simplifies the estimation process: it could reduce estimation time or help reduce the chance of having estimation issues. Second, it could make interpretation easier. Lastly, simultaneous modeling means that the specifications of the segmentation and outcome models can affect each other, and if the analyst does not want this situation (e.g. if it is

desirable to identify a set of clusters that remains stable across a number of different models or analysis purposes), a two-step approach may be beneficial.

2.3.4 Membership model

Sections 2.3.4 and 2.3.5 respectively examine the two model components in finite mixture modeling: membership (i.e. the model component that characterizes the segmentation, f_m) and outcome models (f_o in Figure 2-6).

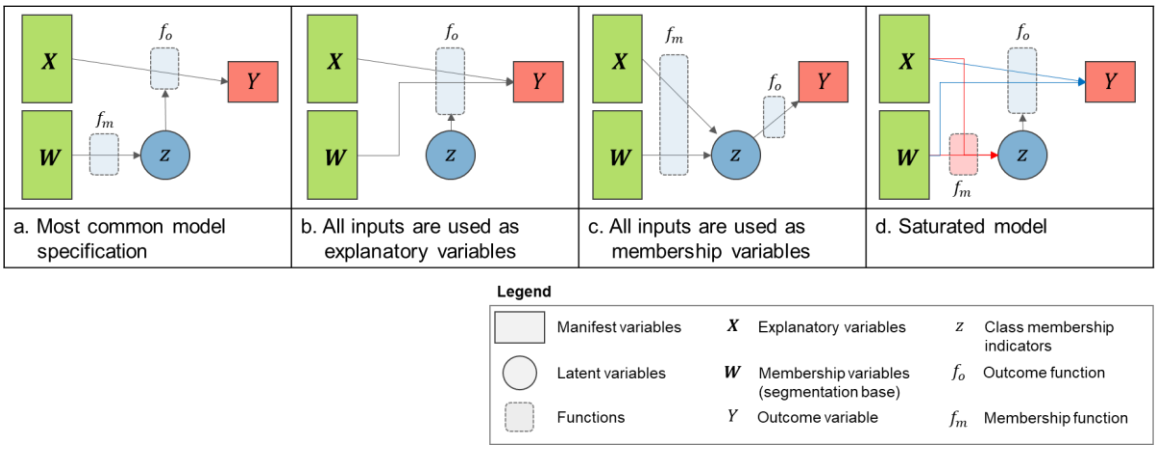


Figure 2-6. Illustration of model specifications in finite mixture modeling

2.3.4.1 Membership function

It is common to focus more on outcome models than on the membership model, but the membership model has its own value, since it helps us understand the nature of the classes. Our first interest is in the functional form of membership models. Some mixture modeling applications directly estimated membership probabilities (*mixing coefficients*). The basic implicit constraints are $0 \leq \pi_z \leq 1$ and $\sum_{z=1}^{Z} \pi_z = 1$, where z is a class membership indicator and π_z is a mixing coefficient. These constraints are required

because we assume classes are collectively exhaustive and mutually exclusive given the sample. Although this approach, of directly estimating mixing coefficients, was the original idea of mixture modeling in the early stages, it is still used (e.g. Arentze, 2015; Park et al., 2016). However, recently it seems more common to use certain *link functions* as membership models (denoted as f_m in Figure 2-6), with the most popular link function being a logit link. There are two clear benefits of using a link function. First, from a technical perspective, we naturally satisfy the conditions of $0 \leq \pi_z \leq 1$ and $\sum_{z=1}^{Z} \pi_z = 1$ by reparameterization (e.g. $\pi_{z'} = \exp(\tau_{z'}) / \sum_{z=1}^{Z} \exp(\tau_z)$). For example, even when we construct a membership model with only constants, when a link function is employed those constants (e.g. τ_z) do not need to be constrained, unless there are specific reasons to do so. Second, we can reparameterize membership probabilities with covariates of interest (as shown in Figure 2-6a). In theory, other types of link functions are possible (e.g. probit), but they are rarely used. We speculate that this is because using a probit link function adds a computational burden, unless it is a two-class situation (recall that the probit probability is not a closed form function). However, the probit link function has been used in the switching model (e.g. Ding et al., 2015; CHAPTER 3), which is closely related to finite mixture models (this can be illustrated by changing z from a latent variable to a manifest variable in Figure 2-6; CHAPTER 3 covers more details about this connection). Another unique attempt in the transportation literature was to adopt a fuzzy c-means method to construct membership probabilities (Ishaq et al., 2014).

Parameterizing the membership function with covariates not only improves the model, but also enriches interpretation. Without parameterization of the membership function, every individual has the same fixed segment membership probabilities. On the

other hand, characterizing segments using information such as individual traits provides more behavioral insights. In marketing research, the class membership variables, denoted as $\boldsymbol{W}$ in Figure 2-6, are also known as *segmentation bases* (Wedel and Kamakura, 2012), and behavioral studies in transportation have widely used covariates in the membership model. The pie chart in Figure 2-7 exhibits the distribution of membership model specifications in the selected literature. Among 284 papers, 50% of them employed membership variables while 38% modeled class membership with only constants (e.g. Hess et al., 2013b; Li, 2018). One of the major reasons for constructing the membership model with only constant(s) is to minimize the complexity of the estimation process (e.g. Zou et al., 2014; Molin and Maat, 2015; Pan et al., 2019). Furthermore, if there is no clear conceptualization of the classes, or if empirical tests indicate that most of the segmentation variables are insignificant (e.g. Peer et al., 2014; Tirachini et al., 2017; Ferguson et al., 2018; Oliva et al., 2019; Zhou et al., 2020), then it could be appropriate to assume that every individual case has the same membership probabilities.

2.3.4.2 Membership variables

There are many possibilities for representing a membership model. Table 2-6 exhibits possible segmentation bases in travel behavior/demand applications. Figure 2-7 (bar graph) presents the distribution of membership variables used in the literature. The most popular segmentation base in the literature is some *individual characteristics* (72% of papers using membership variables employed demographic traits). This is not surprising, because not only do demographics constitute common segmentation bases in marketing research but also they are typically the most basic information available in many datasets. Numerous demographics have been used for segmentation. In the selected literature,

gender, age, and income are particularly popular segmentation bases, and many studies have included those three components in membership models (e.g. Wen and Lai, 2010; Astroza et al., 2017; Krueger et al., 2018; Kim et al., 2019a; Li et al., 2020). Other personal characteristics have also been used such as education (e.g. Bailey and Axsen, 2015; Mouter et al., 2017; Kroesen, 2019) and race (e.g. Maness and Cirillo, 2016; Ardeshiri and Vij, 2019; Guerra and Daziano, 2020), and work/occupation related information (e.g. Chakour and Eluru, 2014; Erdogan et al., 2015; Angueira et al., 2019). In terms of household characteristics, household size/composition (e.g. Olaru et al., 2011; Boeri et al., 2014; van de Coevering et al., 2018) and vehicle ownership (e.g. Nayum et al., 2013; Hackbarth and Madlener, 2016; Anowar et al., 2019) have been popular.

In travel behavior studies, several studies have emphasized that *attitudes* are important factors shaping latent segments. Swait (1994) proposed an early conceptual model reflecting this approach, by conjecturing that general perceptions/attitudes as well as socio-demographics shape the membership likelihood of latent classes. Beck et al. (2014, p. 178) compared two models (latent class models without attitudes and with attitudes in the membership model) and reported that "The inclusion of the attitudinal data, however, allows a more robust class to emerge with clearly defined properties. The attitudinal data is crucial in understanding the class differences, a crucial requirement that allows policy decisions to be considered in a more informed framework and to thus avoid incorrect interpretations of results." Olaru et al. (2011) estimated two location choice models to compare the relative importance of demographics and attitudes. In their application, the membership model that incorporated attitudes outperformed the one that contained demographics. Argahi et al. (2016) used attitudinal constructs to model class

membership and identified three classes of “price hunters, “luggage lovers”, and “ecoflyers” for flight choice. Molesworth and Koo (2016) segmented the sample into two classes based on two factors related to trust in technology. Bailey and Axsen (2015) employed variables of technology-oriented lifestyle, biospheric/altruistic values, and privacy concern to aid in segmenting the sample. Ma et al. (2015) employed three factor scores related to attitudes toward modes. Kim and Mokhtarian (2018) posited that attitudes could act as moderators on the effects of other factors; they identified two (auto-oriented, and urbanite) segments based on attitude measures and found that the two segments have different sensitivities to built environment characteristics. In a commute mode choice context, Choi and Mokhtarian (2020) hypothesized (and confirmed) the existence of classes based on attitudes toward work, multitasking, and productive uses of travel time. Tran et al. (2020) examined how classes based on environmentalism and attitude towards physical activity are associated with mode choice.

Certain *geographical* or *built environment characteristics* have been crucial segmentation bases. Characteristics could be different scales of measures. For example, they can be regional factors such as state/province indicators (Wafa et al., 2015; Abotalebi et al., 2019; Kormos et al., 2019; Astroza et al., 2019), city/metropolitan indicators (e.g. Bhat et al., 2004; Angueira et al., 2015; Krueger et al., 2018; Kim et al., 2019a), and census division indicators (e.g. Maness and Cirillo, 2016). On a smaller scale, neighborhood type, particularly if it is an urbanized area, has been employed for segmentation (e.g. Kroesen, 2015; Prato et al., 2017). Or some studies utilized indices such as “D variables”, which describe land use characteristics (e.g. Sobhani et al., 2013; Ferguson et al., 2018; Anowar and Eluru, 2018).

Behavioral indicators also have been used to characterize classes. In the literature, various factors have been used such as mode use habits (Fu, 2020), trip frequency (Fatmi and Habib, 2016), mode use frequency (Seelhorst and Liu, 2015; Rossetti et al., 2019; Saxena et al., 2019), the amount of driving (Tawfik and Rakha, 2013), and technology use (Saxena et al., 2019; Alonso-Gonzalez et al., 2020; Khan et al., 2020). Other *contextual* variables have also been employed. There are a variety of possibilities because numerous empirical contexts exist. Examples include crash characteristics in safety analysis (e.g. Eluru et al., 2012; Yasmin et al., 2014; Fatmi and Habib, 2019; Li et al., 2019; Li et al., 2020), trip-related characteristics for mode choice analysis (e.g. Bhat, 1997; Wen et al., 2012; Wang et al., 2020), choice situation (e.g. Tawfik and Rakha, 2013), and temporal/seasonal factors (e.g. Shamshiripour et al., 2019; Yu et al., 2019; Faghih-Imani and Eluru, 2020).

Table 2-6. Typology and corresponding examples of segmentation bases in travel behavior/demand applications

	Individual	**Geographical**	**Behavioral**	**Contextual**
Objective measures	Demographics; household characteristics	Regional indicator; neighborhood type; built environment characteristics	Relevant behaviors; previous records	Mode/vehicle related characteristics; temporal characteristics; seasonality
Subjective measures	Attitudes; preferences	Perceptions about region	Modality style; lifestyle	Perceptions

* Note: this table is a modification from Wedel and Kamakura (2012) by customizing it for transportation studies

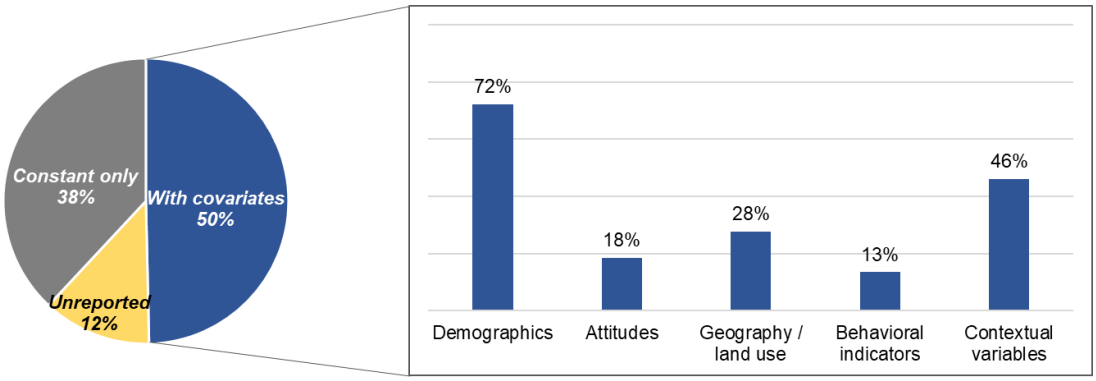


Figure 2-7. Distribution of membership variables in the literature

Although numerous variables have been used in the membership model, quite surprisingly, conceptual-level discussions about the "proper" specification of the membership model in the given context, i.e. (with reference to Figure 2-6) the definitions of $\boldsymbol{W}$ versus $\boldsymbol{X}$, are scarce. Most applications of latent class modeling follow the specification shown in Figure 2-6a (albeit without a discussion of the rationale for including a variable in one model instead of the other), whereas others (Figure 2-6 b-d) are possible in theory.[19] In statistical terms, $\boldsymbol{X}$ is a vector of explanatory variables that have *direct* effects on the outcome, whereas $\boldsymbol{W}$ acts as *moderators* of the impacts of $\boldsymbol{X}$ on the outcome ("dimmer switches" on the parameters, referring to Wu and Zumbo, 2007, although for finite mixture models the analogy is more like a knob or dial with only a finite number of settings, rather than a dial that can continuously vary the parameter).

[19] Figure 2-6b and Figure 2-6c are possible specifications, but uncommon ones. However, the mixture of experts approach in machine learning follows the specification of Figure 2-6d. For more details, please refer to CHAPTER 6.

We are aware of only a few studies that discussed (1) the rationale or hypothesis for selecting membership variables and/or (2) how such a specification is empirically beneficial in the study context. Related to the rationale for *selecting* membership variables, Walker and Li (2007) aimed to find lifestyle segments related to household residential location choice; hence they selected variables that are expected to be associated with household lifestyle (e.g. household structure, employment situations in the household). Vij and Walker (2014), to allow for preference endogeneity, proposed to formulate class membership as a function of not only demographics, but also consumer surplus (which is a function of level of services and choice sets) of each decision-maker. Kim and Mokhtarian (2018) hypothesized that attitudes influence the impact of the built environment characteristics on vehicle ownership and formulated classes with attitudinal propensities. Choi and Mokhtarian (2020) conjectured that the disutility of travel time in the choice between transit with internet access and an alternative mode (such as driving alone) would differ depending on attitudes toward working and multitasking while traveling, and specified latent class membership models accordingly.

Related to the value of *having* membership variables at all, Wen and Lai (2010) and Wen et al. (2012) compared latent class models with and without covariates (e.g. demographics and contextual factors) for choice modeling (airline choice and high-speed rail access mode, respectively), and found that adding covariates in the membership model improved the model fit. Zou et al. (2013) compared two finite mixture negative binomial models with and without membership variables; they found that the model with membership variables not only had a better performance but also that membership variables helped reveal the source of heterogeneity. Zou et al. (2014) compared 11 different

specifications of membership models and found empirical value to considering membership variables. Fountas et al. (2018) compared two models with different segmentation bases: road segment type versus accident type. They showed, for both approaches, empirical support for having membership explanatory variables compared to constant-only membership models. Hence, a common finding is that having membership variables helps to improve the model or interpretability.

As an additional observation related to formulating the membership model, in the literature we also found several studies taking a "two-step approach". This is distinct from the one described in Section 2.3.3, in which latent classes were identified first (generally using covariates) and outcome models estimated for each class in a separate second step. Here, two-step studies first build a latent class model with a *constant-only* membership model, and then construct a separate "class choice" model to characterize the latent classes (e.g. Simons-Morton et al., 2013; Ralph et al., 2016; Kim and Chung; 2016; Kim et al., 2017; Pani et al., 2020). Although this could be an alternative approach, two things need to be noted. First, to construct a post hoc membership model, individuals must be assigned into segments (often based on posterior membership probabilities). By doing so, the probabilistic nature of latent class modeling is lost. Second, incorporating covariates in the membership model in the first place is more robust and efficient for estimation.

2.3.5 Outcome model

Table 2-7. Various outcome models with selected example studies

Type of variable	Outcome model	Example studies
Binary	Logit	Wen et al. (2016), Molesworth and Koo (2016), Savolainen (2016), Lin et al. (2017), Jin et al. (2018), Ge et al. (2018), Griswold et al. (2018), Yu et al. (2019b)
Multinomial	Logit	Bhat (1997), Greene and Hensher (2003), Walker and Li (2007), Vij and Walker (2014), Kim and Mokhtarian (2018)
	Error component logit	Prato et al. (2017), Saxena et al. (2019a), Saxena et al. (2019b)
	Mixed logit	Hess et al. (2013), Razo and Gao (2013), Vij et al. (2013), Yu et al. (2019a)
	(Generalized) nested logit	Wen et al. (2012), Wen et al. (2013), Pan (2019), Tinessa et al. (2020)
Ordinal	(Generalized) ordered logit	Eluru et al. (2012), Yasmin et al. (2014), Anowar and Eluru (2018), Oliva et al. (2018), Fatmi and Habib (2019)
	Ordered probit	Erdogan et al. (2015), Fountas et al. (2018)
Count	Poisson	Simons-Morton et al. (2013), Yasmin and Eluru (2016)
	Negative binomial	Zou et al. (2013), Zou et al. (2014), Park et al. (2016)
Continuous	Gaussian	Zahabi et al. (2015); Ma et al. (2016)
	Lognormal	Van den Berg et al. (2012); Koutsopoulos and Farah (2012); Kazagli and Koutsopoulos (2013)
	Gamma	Kim and Mahmassani (2014); Elhenawy and Rakha (2015); Li et al. (2015)
	Skew-t	Zou and Zhang (2011), Zou et al. (2017)
	Tobit	Anderson and Hernandez (2017), Chand and Dixit (2018)

Due to the flexibility of the mixture modeling framework, numerous outcome models are possible (denoted as f_0 in Figure 2-6). Decisions on the functional form of outcome models are mainly dependent on particular empirical contexts. For classification or choice modeling (nominal outcomes), multinomial logit models (binary logit models, if two classes) have served as the dominant functional form. Alternative models such as ordered logit, error component logit, (generalized) nested logit, and mixed logit also have been applied. For regression (continuous or count outcomes), various outcome models have been used such as Gaussian, log-normal, Poisson, negative binomial, Tobit, Gamma, skew-

t, and so on. Given this plethora of possibilities, we do not enumerate the equations for all possible models. Readers may refer to some selected papers (Table 2-7) for details about outcome models of interest. An additional remark on the outcome model is that we can have a different type of outcome function for each class (heterogeneity in functional forms, see Section 2.3.1), by taking the confirmatory approach (Section 2.3.2).

2.3.6 Number of classes and rationale behind decisions

Identifying latent sub-segments is a critical element in applications of finite mixture modeling. This includes identification of (1) how many meaningful segments exist given the data, (2) the relative sizes of each segment (i.e. class shares), and (3) the distribution of key characteristics of interest within each segment. Thus, determining the number of classes is a key step in the modeling process.

Most papers on the subject have a subsection giving a brief background on how to determine the number of classes. Various measures, in particular some information criteria (based on the log-likelihood), are mentioned. There are numerous other measures; readers can refer to McLachlan and Peel (2001) and Vermunt and Magidson (2016). Basically, when adding more segments, log-likelihood values improve monotonically because we are adding more parameters (creating a more flexible model, which – all else equal – is more likely than a more constrained model). Then the decision on the preferred model is linked to the question of how much the model complexity (i.e. number of parameters) needs to be penalized. Among three commonly-used measures, the degree of penalization decreases in order of CAIC (consistent AIC), BIC, and AIC. There is no universally accepted measure so far. In general, minimizing the BIC is the most common decision rule in the literature.

Another semi-quantitative decision guideline is the “elbow rule”, meaning to examine the change of measures by the number of classes and find an elbow (or knee) of the curve (e.g. Keskisaari et al. 2017). Numerous studies report a table that presents model selection indices by the number of classes, optionally including the final log-likelihood and number of parameters for each number of classes. A few other measures also have been used such as entropy (Prato et al., 2019; Jahanshahi and Jin, 2020), integrated completed likelihood criterion (Zou and Zhang, 2011; Yu and MacKenzie, 2016), informational complexity (Bae et al., 2019), and bootstrap likelihood ratio test (Yu et al., 2017).

Interestingly, however, although most papers point out that these measures are common criteria for selecting the number of classes, a large fraction of studies made the final decision on the number of classes more heuristically and qualitatively (Figure 2-8a). About 25% and 40% of studies (respectively involving supervised and unsupervised learning) made more subjective final decisions. In about 25% and 8% more studies, the number of classes was pre-determined, either because a confirmatory approach was involved (see Section 2.3.2) or because the analyst chose a certain number of classes (mainly two) for convenience of comparisons, estimation, or interpretation. This lack of servitude to quantitative metrics is because it is advised that the analyst should choose the solution that is most interpretable and meaningful (cf. Scarpa and Thiene, 2005). In other words, studies look for the solution that can provide (1) conceptually valid signs/magnitudes of key parameters (e.g. Vij et al., 2017; Koresen, 2019; Thorhauge et al., 2020), (2) statistical significance of key parameters (e.g. Molesworth and Koo, 2016; Rahmani and Loueiro, 2019; Wang et al., 2020), and/or (3) “meaningful” differences across classes or healthy sizes of class shares (e.g. Liao et al., 2018; Fu, 2020). Some

studies fix the number of classes to a certain number for the sake of model comparisons, for simplifying the empirical experiment. Hence, the literature tends to select models with fewer classes than the optimal number "confessed" by the data. A limited number of studies adopted more than the "optimal" number of classes (with respect to BIC) for securing better interpretation or statistical significance of parameters (e.g. Jin et al., 2018). Fewer than 10% of studies explicitly reported that they considered both quantitative and qualitative dimensions and both criteria produced the same number of classes.[20] Some studies did not report how they made decisions.

Figure 2-8b shows the distribution of the number of classes chosen. It is worth noting that the average numbers of classes chosen vary across types of problems. On average, supervised learning problems had a smaller number of classes than that of unsupervised learning. A majority of (supervised) studies using classification/regression, in particular, adopted the two-class solution, whereas (unsupervised) studies using cluster analysis produced a more diverse number of clusters. Vij and Krueger (2017) also pointed out that most latent class (choice) models have been generally restricted to a small number of classes. We conjecture that this happens because studies focusing on cluster analysis have a primary goal of uncovering more diverse segments, whereas classification/ regression studies ultimately aim to examine meaningful outcome models across segments. In the latter instance, a larger number of classes (1) penalizes for complexity more heavily;

[20] We counted this only when the authors explicitly commented that they considered both aspects and they matched. However, we expect that additional studies implicitly considered both aspects and they produced the same optimal number of classes. For example, some studies that reported making decisions based on quantitative decisions might have been likely to choose a final solution that was conceptually valid and interpretable.

(2) can create estimation difficulties; and (3) increases the difficulty of interpretation in view of the multiplicity of parameters across all membership and outcome functions.

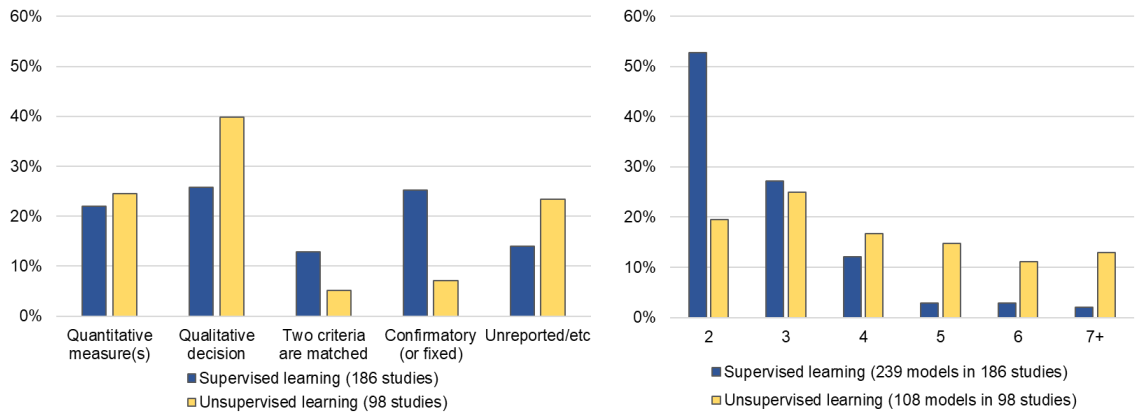


Figure 2-8. Distributions of (a) decision rationale and (b) number of classes chosen

2.3.7 Model comparisons: baseline and competing models

Model comparison has been fairly common in the literature, to show how well mixture modeling performs and thus why it is useful compared to other competing models. Model comparison is relevant to addressing important questions of whether the assumption of heterogeneity is valid and whether mixture modeling works well given the empirical context. There are multifaceted angles of possible comparisons: a certain type of mixture model can be compared with (1) the baseline model (i.e. a model with the homogeneity assumption), (2) deterministic/exogenous segmentation models, and (3) other competing models, particularly including continuous mixture models. There can also be an "internal" comparison, among models having a different number of classes (e.g. 2-class versus 3-class solution). Since we discussed the latter in Section 2.3.6, here we focus on comparisons with other types of models.

First, the majority of studies posited unobserved heterogeneity in either taste or data, and this was a major reason why they proposed to use some type(s) of mixture modeling. Indeed, numerous papers have empirically corroborated the existence of heterogeneity in some kind of analysis. A devil's advocate question is how bad the simpler model (which does not account for heterogeneity) is. Here, "how bad" could be with respect to the performance and/or bias of the results. Even if a sizable portion of studies discusses (or at least provides) a *baseline model* (i.e. a 1-class model), a non-trivial portion of studies takes heterogeneity for granted and skips this baseline model. In our literature review, about half of the studies did not report or discuss a baseline model. However, still, many studies reported better performance and more insights from a model with mixture modeling compared to the "one-size-fits-all" model. For example, Rahmani and Loueiro (2019) reported better prediction of the latent class model compared to the MNL (70.7% and 43.6%). Hackbarth and Madlener (2016) presented scenario analysis of vehicle type choice and showed that each segment produced notably different sensitivities compared to the pooled model. Kim and Mokhtarian (2018) addressed endogeneity bias via treating taste heterogeneity (using a latent class model) and compared how the biased estimates (without segmentation) and estimates based on the latent class model produced different solutions in the scenario analysis.

Deterministic/exogenous segmentation has been widely used due to its simplicity. The basic idea is to segment the sample into subgroups based on certain factor(s) that are expected to be associated with heterogeneity. We can consider a deterministic segmentation model as a special case of mixture modeling where we know the "true" segment indicator (CHAPTER 3) – i.e., class membership probabilities are 1s and 0s.

Hence, this comparison is in fact linked to the fundamental question of *whether latent segments are more helpful in explaining behaviors than observable segments*. By reflecting the probabilistic nature of segmentation, mixture models can be conceptually more appealing and their superiority has often been demonstrated with empirical data. For example, Teichert et al. (2008) compared deterministic segmentation (by flight class) and latent class approaches in the flight choice context. They found that the latent class model was able to identify more diverse customer segments, and reported value in connecting the two solutions by cross-validation of the two approaches. Wafa et al. (2015, p. 138) reported that "It is found that endogenous segmentation [i.e. the latent class model] better fits the data as compared with exogenous segmentation, allows for higher-order interaction effects, keeps the number of segments under control, and provides more intuitive results with respect to the identification of homogeneous clusters of units." Further, they showed empirical evidence of the superiority of endogenous segmentation over exogenous segmentation (by geography). Arunotayanun and Polak (2011) modeled freight shippers' mode choice behaviors; they found that models deterministically segmented by commodity retained a substantial amount of residual heterogeneity and latent class modeling helped account for those heterogeneities. Kim and Mokhtarian (2018) compared deterministic segmentation and latent class choice models of vehicle ownership with respect to attitudes; the study found that the latent class model outperformed its deterministically segmented counterpart and the authors argued for its conceptual superiority. CHAPTER 3 compares deterministic segmentation, endogenous switching regression, and latent class models in the context of modeling vehicle-miles traveled. It concludes that each approach has its own

purpose and model choice might be context-dependent, but the latent class model provided better performance than the others, and a meaningful interpretation.

As discussed in Section 1.2, finite mixture modeling treats (parameter) heterogeneity in a discrete way. A natural follow-up question is whether this assumption of discreteness is valid, or superior to that of *continuous mixture models*; hence this comparison has been particularly popular. In general, latent class models outperformed counterpart random parameter models with respect to some selected quantitative criteria. For example, Mahmud et al. (2020) model a speed choice behavior on a rural highway by comparing random parameter and latent class models; in spite of small differences due to the small sample size, the in-sample predictive performance results (e.g. RMSE, MAPE) indicated that the latent class model outperformed the random parameter model. Espino and Roman (2020) examined the transfer behavior of bus users in Gran Canaria, Spain and found that the latent class model is superior to mixed logit based on a test for non-nested choice models (while calling for more investigation into model comparisons). As well, many other studies reported that latent class models outperformed random parameter models in their empirical contexts (e.g. Wen et al., 2016; Yu and MacKenzie, 2016; Adanu and Jones, 2017; Qin et al., 2017; Faghih-Imani and Eluru, 2020).

In Guerra and Daziano (2020), mixed logit models fit the data better than two-class latent class models, and yet the study focused on a latent class solution because of its behavioral insights. A few studies concluded that there was no definitive evidence that one model was better than the other in their empirical contexts. In addition, some studies presented a more cautious standpoint related to model comparisons. A seminal work of Greene and Hensher (2003) compared MNL, mixed logit, and latent class models in the

context of route choice. Although the latent class model had a bit higher pseudo-R-squared, they emphasized that the two approaches to dealing with heterogeneity have their own merits rather than selecting a preferred one. Arunotayanun and Polak (2011, p. 145) noted that neither mixed logit nor latent class models should be expected to be unambiguously superior to the other: "…MMNL [mixed logit] and latent class models characterise the taste heterogeneity in different ways. As a result, it is not necessarily the case that fit of a latent class model is superior to that of a MMNL model, and vice versa (Provencher and Bishop, 2004). The relative performance of the models will depend on the nature of the data."

A few studies compared the latent class model with another behavioral continuous mixture model, ICLV. Tran et al. (2020) examined two ways to incorporate attitudes in mode choice modeling and thus provided behavioral interpretations of two models based on the same empirical application. Although the study did not pursue a performance comparison, their results showed that the two models presented almost the same performance (in terms of in-sample information criteria).

2.3.8 Software and estimation

Undoubtedly, commercial or open-source software has catalyzed the spread of model estimation methodologies. In the mixture modeling context, several programs that have been used in the literature include Latent GOLD (Vermunt & Magidson, 2016), NLOGIT (NLOGIT, 2016), Mplus (Muthén & Muthén, 2017), Biogeme (Bierlaire, 2018), Apollo (Hess and Palma, 2020), SAS (Lanza et al, 2007), and Stata (Lanza et al., 2015).

Some open-source libraries are available (e.g. 'MCLUST', Scrucca et al., 2016; 'Flexmix', Leisch, 2004; 'poLCA', Linzer and Lewis, 2011).

The most popular estimation method in the literature is based on the expectation-maximization (EM) algorithm, which is proposed by a seminal work of Dempster et al. (1977). The basic idea of the EM algorithm is to firstly treat class membership as known to enable writing a complete data likelihood function; then find the expected value of the complete data likelihood, conditional on current parameters and data (E-step); then maximize the expected likelihood function with respect to parameters (M-step); then repeat the E-step and M-step iteratively until there is negligible change in the log-likelihood or the estimated parameters from one iteration to the next. Direct gradient-based optimization routines (e.g. quasi-Newton Raphson) are still possible (Kamakura and Russell, 1989; Gupta and Chintagunta, 1994; NLOGIT, 2016), but some studies noted that such algorithms are less stable (cf. Bhat, 1997), for several reasons such as the possibility of an extremely flat likelihood function, and the possibility of getting stuck in regions where the function is not well approximated by a quadratic expression (Vij and Krueger, 2017). Rather, many studies and software packages have used a mixture of EM with gradient-based algorithms, thus employing the advantages of both algorithms and speeding up convergence (cf. Bhat, 1997; Vermunt and Magidson, 2016). NLOGIT (2016), however, favors direct maximum likelihood estimation (MLE) over the EM algorithm: (in theory) both algorithms do not produce different results of the log likelihood and the EM algorithm could be very slow. Regarding stability, both algorithms are subject to getting hung up on local optima, and thus in any case analysts should make an extra effort to ensure the global optimum has been found, by testing various starting points.

One promising approach to speeding up the estimation can be found in the optimization algorithms that are widely used in machine learning. For example, deep neural networks are mostly trained with stochastic gradient descent (SGD) algorithms (cf. Goodfellow et al., 2016), as opposed to the "batch" gradient methods that use all training data to estimate the gradient (like the conventional quasi-Newton algorithm). Han (2019) applied this approach to estimate latent class choice models and found that SGD (specifically "Adam" in this study) was much faster than other algorithms including direct MLE (BFGS in this study) and EM methods. A few studies have taken a Bayesian framework for mixture modeling and estimated models with Markov chain Monte Carlo (MCMC) sampling (cf. Diebolt and Robert, 1994). As this study focuses primarily on the frequentist viewpoint, details on Bayesian estimation are beyond our scope.

2.4 Conclusions

This study examined the finite mixture modeling (latent class modeling) framework with respect to how it has been used particularly in the transportation domain. Through a comprehensive and systematic review, the study aimed to provide a broader understanding of the usage landscape and also insights into detailed elements. We firstly set up the mixture modeling framework (with distinctions of finite vs. continuous and disaggregation and segmentation); outlined an arena of various relevant research fields; and explained how it is connected to transportation analyses. Then, by using the Scopus database, we explored relevant papers to investigate macroscopic trends in usage of the methodology (yearly trends and research topics). We identified six subdomains in transportation with the aid of nonnegative matrix factorization: *discrete choice modeling*, *general behavior analysis*, *crash/safety analysis*, *traffic analysis*, *travel time distribution*, and *electric vehicles*.

We examined several components of the mixture modeling framework in detail. Firstly, we examined various types of heterogeneity, with less emphasis on heterogeneity in data distributions (associated with latent cluster analysis), and more emphasis on heterogeneity in parameters, model specification, attribute processing, functional forms, decision rules, casual structure/order, and constraint/choice set. There have been two approaches to mixture modeling – exploratory and confirmatory. The confirmatory approach is suitable for testing various types of hypotheses (e.g. allowing us to adopt several types of heterogeneity in mixture modeling). The membership model is a unique component of mixture models that allows for endogenously/simultaneously segmenting the sample into more homogeneous subsamples with respect to the behavior generation process and thus it is well connected to the concept of market segmentation. We examined the use of the link function for membership models and various segmentation bases. Also, due to the flexibility of mixture modeling, we found that various outcome models have been used in the literature.

Selecting the number of classes in mixture modeling is one of the key steps. Hence, we explored the number of classes chosen in empirical studies, and the rationales for the selection of a given number. To investigate how well mixture modeling works, there have been various comparisons with other competing models. This includes comparisons with the baseline model (reflecting homogeneity), deterministic/exogenous segmentation models, and continuous mixture models. In general, latent class models outperformed other competing models, thus showing their usefulness. In addition, we briefly touched on some estimation methods and software/programs that help analysts employ mixture models.

CHAPTER 3. ALTERNATIVE APPROACHES TO TREATING PARAMETER HETEROGENEITY

Paper title: *Alternative approaches to treating finite-valued parameter heterogeneity: application to modeling vehicle-miles driven* (under peer review)

3.1 Introduction

This chapter focuses on parameter heterogeneity, taking a finite segmentation approach. Myriads of models have been developed in various fields such as statistics, economics, psychology, and data science. However, such models were developed for different purposes and contexts and, even if some models are eventually performing similar mathematical tasks, they may have different names and/or application approaches. Specifically, in the context of modeling vehicle-miles driven, the chapter highlights similarities and differences among three approaches: deterministic/exogenous segmentation, endogenous switching, and latent class models.

The remainder of this chapter is organized as follows. Section 3.2 provides an overview of some relevant literature associated with modeling vehicle-miles traveled (VMT), particularly with respect to parameter heterogeneity. Section 3.3 examines the theoretical backgrounds of each approach and delineates connections among them. Section 3.4 introduces the empirical context and describes model results. Section 3.5 discusses several considerations associated with applying the models of interest. Section 3.6 summarizes the study and then discusses some implications and limitations of the study. Appendix A provides supplementary detailed discussions about treatment effects.

3.2 Literature review

Vehicle-miles traveled (VMT) continues to serve as an important behavior/demand indicator in ground passenger transportation, mainly because it is closely related to traffic loads on roads and to transportation-based emissions. Many studies have conducted aggregate-level analyses of VMT to analyze trends (e.g. Cervero and Murakami, 2010, who modeled VMT per capita across 370 urbanized areas in the U.S.) or to investigate issues such as induced demand, the rebound effect of improved fuel economy (Hymel et al., 2010), or the impact of telecommuting on VMT (Choo et al., 2005). More recently, there have been analyses of whether VMT has "peaked" (e.g. Polzin and Chu, 2014; Circella et al., 2016), and of the impacts of ridehailing (e.g. Henao and Marshall, 2019; Tirachini and Gomez-Lobo, 2020) and (in the future) automated vehicles (e.g. Zhang et al., 2018; Harb et al., 2018) on VMT. In this study, we focus on a disaggregate-level cousin, vehicle-miles driven (VMD), to explore individuals' sensitivities to key factors such as attitudes and the built environment.

There have been several efforts to incorporate a market segmentation approach to modeling VMT. Studies did not always explicitly mention the method, but deterministic segmentation by selected indicator(s) has been popular. For example, using the 2009 National Household Travel Survey (NHTS), Akar and Guldmann (2012) modeled VMT of the pooled sample, and segmented by number of household vehicles. Ke and McMullen (2017) investigated regional differences in factors influencing household VMT. They compared an Oregon statewide (pooled) model with some segment-specific models (for selected regions and MPOs).

Several other segmentation approaches have also been taken. For example, Chen et al. (2017) modeled personal daily vehicle-kilometers traveled (VKT) in Shanghai, China. In particular, their sample selection (switching regression) model accounted for selection into transit-oriented development (TOD) areas and non-TOD areas, and allowed the VKT model to have different coefficients for each type of area. They found not only that residential self-selection existed, but also that the built environment and attitudes played crucial roles in explaining VKT. A few studies employed cluster analysis to produce particular segments of interest rather than segmenting the sample on the basis of a single variable. Ralph et al. (2016) followed a two-step approach. First, they conducted latent class cluster analysis to produce traveler types, and then applied quantile regression models of VMT for each traveler type.

Self-selection has been a major issue in analyzing travel behavior (e.g. Mokhtarian and Cao, 2008; Cao et al., 2009). It has been considered crucial to account for self-selection when modeling VMT because the conventionally estimated effects of land use or built environments can be biased when residents self-select into certain residential locations. Hence, in recent years, modeling VMT at the disaggregate level has involved correcting for self-selection, using methods such as a Heckman model (Heckman, 1979). Zhou and Kockelman (2008) employed a two-step Heckman model (in this context, a limited-information approach to estimating a switching regression model) to consider self-selection in home location choice (urban vs. suburban/rural) when modeling household VMD. Cao (2009) also used a two-step Heckman model. In the context of Northern California, he found that the self-selection effect accounted for 23.8% of the total influence of neighborhood type on weekly VMD. Salon (2015) segmented California census tracts into

five neighborhood types (central city, urban, suburban, rural-in-urban, and rural) and modeled choice of residential neighborhood types to enable incorporating neighborhood selection variables into the VMT model. This study is an application of Heckman's two-step approach where the number of selection groups is more than two. She confirmed some heterogeneous effects across neighborhood type.

Although latent class modeling has been widely used in travel behavior, particularly in choice modeling contexts (e.g. Bhat, 1997; Vij et al., 2013; Kim and Mokhtarian, 2018), we are aware of only one application of latent class modeling to VMT prediction (Zahabi et al., 2015). Based on numbers of cars, children, and persons in the household, three latent classes were identified.

As mentioned, recent studies have usually focused on self-selection effects when examining the impact of the built environment on travel behavior. In that context, living in a particular type of built environment (usually an urban, denser, mixed use neighborhood) is often viewed as a "treatment", and built environment effects are analyzed using one or more of the treatment effect measurement approaches found in the evaluation literature. Although we will continue the discussion of treatment effect measurement in Section 3.5.1 (with more details in the Appendix A), we will mainly focus on the parameter heterogeneity aspect of various approaches. In particular, we will present three different ways to address parameter heterogeneity (deterministic segmentation, switching regression, and latent class regression), apply them to modeling VMD, and compare the results. Not only is latent class modeling of VMD rare, but so also are comparisons across alternative modeling approaches to treating heterogeneity; hence it is hoped that this study will contribute to our understanding of the implications of using alternative modeling approaches in practice.

3.3 Methodology

In this section, we present several different modeling approaches for taking account of (parameter) heterogeneity, and explore how they are conceptually related and distinct. Of course, one common approach to accounting for such heterogeneity is to allow one or more parameters of the model to be a continuous-valued random variable having a pre-specified distribution (cf. McFadden and Train, 2000; Greene and Hensher, 2003). In this study, we focus on another common approach, namely allowing a given model parameter to take on a *finite* number of different values. The basic premise is that there are a finite number of population segments having different sensitivities to key factors influencing their travel behavior (here, VMD). Options for reflecting this heterogeneity include deterministic segmentation, endogenous switching, and latent class models. For this study, we focus on a regression problem (where the dependent variable is continuous) and the two-segment context. Figure 3-1 exhibits generic model specifications by approach; each specification will be described in the following subsections. Here, observed variables are portrayed by rectangles, and latent variables by ovals. The dashed rounded rectangles denote that the influence of X variables on Y (i.e. the model coefficients) differs by class membership, which is given by *z*.

In Section 3.3.1, we warm up with the simplest, "pooled", model, where we assume homogeneity of parameters. Sections 3.3.2, 3.3.3, and 3.3.4 describe how three models that deal with heterogeneity are related.

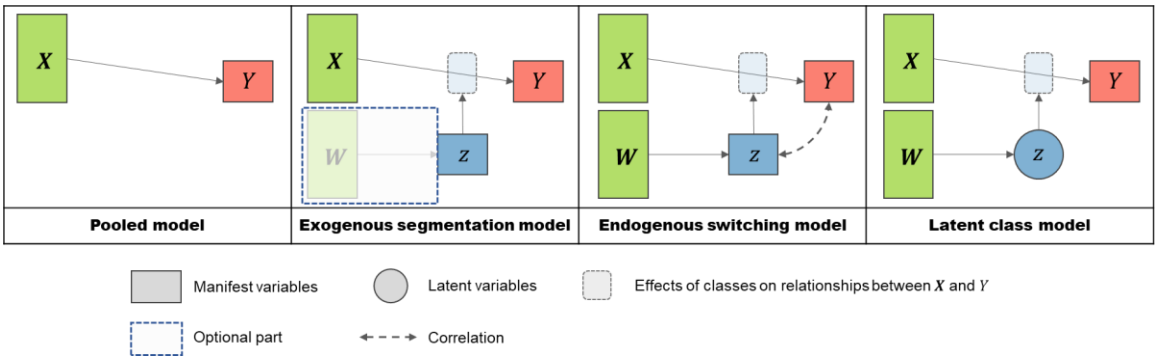


Figure 3-1. Generic model specifications by approach

3.3.1 Pooled model

Most endeavors to model an outcome use an equation that describes the presumed relationship of inputs to the outcome, including unknown parameters to be estimated and an error term capturing the net influence of unmeasured variables on the outcome. If the data are cross-sectional (as opposed to longitudinal), then conditional on the explanatory variables, the observations on the dependent variable are conventionally assumed to be *independent and identically distributed* (*i.i.d.*) and thus the model assumes *homogeneity* of the sample. Homogeneity has multifaceted meanings, including that individuals are drawn from the same distribution and same decision process (i.e. decision structure/model, variables, parameters). Hence, the usual linear regression for a continuous dependent variable $(Y_i|\boldsymbol{X}_i)$ assumes an i.i.d. normal error term and can be formulated as follows:

$$Y_i = \boldsymbol{X}_i\boldsymbol{\beta} + \varepsilon_i \qquad \varepsilon_i \sim N(0, \sigma^2)\ , \tag{3.1}$$

where i indexes the individual, Y is a (continuous) dependent variable, $\boldsymbol{X}$ is a vector of explanatory variables, $\boldsymbol{\beta}$ is a vector of parameters, and ε is the error term.

3.3.2 Deterministic (and exogenous) segmentation model

If the homogeneity assumption does not hold, the estimated model may not adequately represent the decision process of individuals, groups of individuals, or even the population, and could thus lead to a misunderstanding of behavioral implications and/or inferior predictive ability. Suppose we have two groups of interest and we have *a priori* knowledge (or conjecture) that the two groups have distinctive decision processes (meaning, in this context, parameter heterogeneity, i.e. differing $\boldsymbol{\beta}$ coefficients). Let z be a binary indicator of group membership where for any individual i, $z_i = 1$ or $z_i = 0$. Then, the most practical approach might be to deterministically and exogenously segment the groups and model them independently (or model the pooled sample by interacting all variables with the group indicator, a special case of using interaction terms). Note that individuals belong to only one of the groups (i.e. the groups are mutually exclusive). Because of that, deterministically segmented models have different conditional distributions, meaning that Y is distributed *given* that the group is 1 or 0: i.e. we can distinguish the outcomes by segment, and speak of $Y_{i1}|X_{i,z_i=1}$ and $Y_{i0}|X_{i,z_i=0}$. Hence, any estimates of the two models provide expectations conditional on the group indicator. Specifically, we have:

$$Y_{i1} = \boldsymbol{X}_i\boldsymbol{\beta}_1 + \varepsilon_{i1} \qquad \varepsilon_{i1} \sim N(0, \sigma_1^2) \tag{3.2}$$

$$Y_{i0} = \boldsymbol{X}_i\boldsymbol{\beta}_0 + \varepsilon_{i0} \qquad \varepsilon_{i0} \sim N(0, \sigma_0^2)\ , \tag{3.3}$$

and conditional class-specific expectations are:

$$E(Y_{i1}|\boldsymbol{X}_i, z_i = 1) = \boldsymbol{X}_i\boldsymbol{\beta}_1 \tag{3.4}$$

$$E(Y_{i0}|\boldsymbol{X}_i, z_i = 0) = \boldsymbol{X}_i\boldsymbol{\beta}_0 \, . \tag{3.5}$$

Then, the unconditional outcome can be expressed as:

$$y_i = z_i Y_{i1} + (1 - z_i)Y_{i0} \, .^{21} \tag{3.6}$$

Since we assume normality, the conditional densities can be written as follows:

$$f(y_i|\boldsymbol{X}_i, z_i = 1) = f(Y_{i1}|\boldsymbol{X}_i) = \frac{\exp\left[-\frac{1}{2\sigma_1^2}(Y_{i1} - \boldsymbol{X}_i\boldsymbol{\beta}_1)^2\right]}{\sqrt{2\pi\sigma_1^2}} = \frac{1}{\sigma_1}\phi\left(\frac{Y_{i1} - \boldsymbol{X}_i\boldsymbol{\beta}_1}{\sigma_1}\right) \tag{3.7}$$

$$f(y_i|\boldsymbol{X}_i, z_i = 0) = f(Y_{i0}|\boldsymbol{X}_i) = \frac{\exp\left[-\frac{1}{2\sigma_0^2}(Y_{i0} - \boldsymbol{X}_i\boldsymbol{\beta}_0)^2\right]}{\sqrt{2\pi\sigma_0^2}} = \frac{1}{\sigma_0}\phi\left(\frac{Y_{i0} - \boldsymbol{X}_i\boldsymbol{\beta}_0}{\sigma_0}\right) . \tag{3.8}$$

The log-likelihood (LL) can be expressed as:

$$LL = \sum_{i=1}^{N} \ln f(y_i|\boldsymbol{X}_i) = \sum_{i=1}^{N} \ln[z_i f(Y_{i1}|\boldsymbol{X}_i) + (1 - z_i)f(Y_{i0}|\boldsymbol{X}_i)] \tag{3.9}$$

---

[21] Alternatively, it can be expressed as $y_i = {Y_{i1}}^{z_i} {Y_{i0}}^{(1-z_i)}$. For the deterministic segmentation model, the two expressions produce identical log-likelihood functions and thence identical parameter estimates:

$$\begin{aligned} LL &= \sum_{i=1}^{N} \ln f(y_i) = \sum_{i=1}^{N} \ln\{f(Y_{i1})^{z_i} f(Y_{i0})^{(1-z_i)}\} = \sum_{i=1}^{N}\{z_i \ln f(Y_{i1}) + (1 - z_i)\ln f(Y_{i0})\} \\ &= \sum_{i=1}^{N}\left\{z_i \ln\left[\frac{1}{\sigma_1}\phi\left(\frac{Y_{i1} - \boldsymbol{X}_i\boldsymbol{\beta}_1}{\sigma_1}\right)\right] + (1 - z_i)\ln\left[\frac{1}{\sigma_0}\phi\left(\frac{Y_{i0} - \boldsymbol{X}_i\boldsymbol{\beta}_0}{\sigma_0}\right)\right]\right\} \\ &= \sum_{z_i=1} \ln\left[\frac{1}{\sigma_1}\phi\left(\frac{Y_{i1} - \boldsymbol{X}_i\boldsymbol{\beta}_1}{\sigma_1}\right)\right] + \sum_{z_i=0} \ln\left[\frac{1}{\sigma_0}\phi\left(\frac{Y_{i0} - \boldsymbol{X}_i\boldsymbol{\beta}_0}{\sigma_0}\right)\right] . \end{aligned}$$

$$= \sum_{i=1}^{N} \ln \left[z_i \frac{1}{\sigma_1} \phi \left(\frac{Y_{i1} - \boldsymbol{X}_i \boldsymbol{\beta}_1}{\sigma_1} \right) + (1 - z_i) \frac{1}{\sigma_0} \phi \left(\frac{Y_{i0} - \boldsymbol{X}_i \boldsymbol{\beta}_0}{\sigma_0} \right) \right]$$

$$= \sum_{z=1} \ln \frac{1}{\sigma_1} \phi \left(\frac{Y_{i1} - \boldsymbol{X}_i \boldsymbol{\beta}_1}{\sigma_1} \right) + \sum_{z=0} \ln \frac{1}{\sigma_0} \phi \left(\frac{Y_{i0} - \boldsymbol{X}_i \boldsymbol{\beta}_0}{\sigma_0} \right)$$

Note that this deterministic segmentation model is closely related to *decision tree regression*. The basic idea of decision tree regression is to partition the input space into a set of rectangles (i.e. into segments) and then fit a model for each segment (Breiman et al., 1984; Hastie et al., 2009). Hence, both the deterministic segmentation and decision tree regression models aim to build segment-specific models. However, in the usual applications of decision tree regression, *all* input variables are used to split individuals into branches (i.e. segments) and constant outcomes are predicted for each branch (usually the average outcome over cases in that branch) rather than having models per se.[22] By contrast, in the usual applications of deterministic segmentation models, researchers pick one or a few segmentation variable(s) and then build segment-specific models with the remaining input variables. The deterministic segmentation model can easily be expanded to multiple classes and/or different types of outcome variables (e.g. binary, ordered, nominal); in fact, decision tree regression usually produces more than two segments.

[22] We take quite a different approach from Brathwaite et al. (2017) for interpreting decision tree models. In a superb connection of machine learning and microeconomics, Brathwaite et al. (2017) linked decision tree models to non-compensatory decision rules (one of various decision protocols) in the discrete choice modeling context. In that case, input variables are considered explanatory variables in the model, but the linkage between explanatory variables and outcome function does not take a tidy linear compensatory form. Here, we interpret input variables in typical tree-based models as segmentation variables, so as to link them to latent class structure. Although in this study we focus on a regression model, our interpretation can also be applied to the context of discrete choice modeling.

3.3.3 Switching regression model

The original idea of switching regression assumes two regimes, each generating an outcome, based on different equations. Switching could be either endogenous or exogenous (Maddala, 1986), respectively depending on whether the error terms of the outcome equations are or are not correlated with that of the switching (segmentation) model. The endogenous switching regression model, a type of sample selection model, is also known as the Tobit type 5 model (Amemiya, 1985), the mover-stayer model, or Roy's model. This model is particularly related to some others [original Tobit (Tobin, 1958), two-part (e.g. Cragg, 1971), and Heckman's original sample selection models (Heckman, 1979)].[23]

To relate the switching regression models to mixture modeling, we need to consider similarities and differences. Both approaches start from the same key assumption – they posit that there are *subpopulations* in the population, which exhibit different behavioral processes. Strictly speaking, finite mixture modeling generally refers to cases where the mixing segments are *latent* (i.e. true segment membership is unknown; hence so-called latent class models), whereas in switching models, we know the true group

[23] Both two-part and sample selection models appeared in the 1970s in the econometrics literature. The two approaches are interrelated, but the initial motivations were different. The two-part model focused on predicting *actual* (conditional) outcomes (Leung and Yu, 1996), and aimed to address having an "excessive" number of zeros in the sample distribution of a continuous variable. Cragg (1971) proposed some formulations of *two-part* models, following a precedent work of Tobin (1958). The earlier *Tobit* model introduced a latent variable (which is a sort of proxy for the outcome variable) to tackle a censor problem, but simply formulated probability-of-zero and outcome-if-not-zero models with the same explanatory variables. Cragg's (1971) model allowed the so-called "participation" (zero versus non-zero) and "intensity/amount" (positive values) parts to have different specifications (so it has "two parts"). On the other hand, the sample selection model focused on predicting *potential* (unconditional) outcomes, in the context of endogenous sampling. In the original *Heckman (1979) sample selection* model, the outcome is only observed if self-selected (e.g. wage is not observed for the unemployed). The endogenous switching regression model can be considered as a variation of the sample selection model in which each segment (or regime) has its own outcome model (i.e. outcomes are observed in all regimes). As two-part and sample selection models gained attention, several studies compared the two approaches (Manning et al., 1987; Leung and Yu, 1996; Dow and Norton, 2003; Madden, 2008).

indicator. In this study, we want to generalize the concept of the finite mixture modeling framework by recognizing that knowledge of the true membership results in a special type of mixture model, one in which the ordinarily probabilistic class membership in fact consists of probabilities 1 and 0. In this regard, the (exogenous) switching model can be embraced under the mixture modeling framework. *Endogenous* switching, however, deviates from the mixture modeling approach in a seemingly slight, but substantively meaningful, way.

An (endogenous) switching regression system of equations consists of two parts: the membership model (known in that literature as the selection model) and the outcome models. The membership model can be expressed as:

$$z_i^* = \boldsymbol{W}_i\boldsymbol{\alpha} + u_i \,, \tag{3.10}$$

where i indexes the individual, z_i^* is a latent continuous variable determining class membership, $\boldsymbol{W}$ is a vector of membership variables, $\boldsymbol{\alpha}$ is a vector of membership parameters, and u_i is an error term. If $z_i^* > 0$ the individual belongs to class 1 ($z_i = 1$), and otherwise the individual belongs to class 0 ($z_i = 0$). If we assume $u_i \sim N(0,1)$, we have a binary probit membership model, with probabilities given by:

$$P(z_i = 1) = \Phi(\boldsymbol{W}_i\boldsymbol{\alpha}) \,, \tag{3.11}$$

$$P(z_i = 0) = 1 - P(z_i = 1) = 1 - \Phi(\boldsymbol{W}_i\boldsymbol{\alpha}) = \Phi(-\boldsymbol{W}_i\boldsymbol{\alpha}) \,. \tag{3.12}$$

The two class-specific outcome models are:

$$Y_{i1} = \boldsymbol{X}_i\boldsymbol{\beta}_1 + \varepsilon_{i1} \quad \varepsilon_{i1} \sim N(0, \sigma_1^2) \tag{3.13}$$

$$Y_{i0} = \boldsymbol{X}_i\boldsymbol{\beta}_0 + \varepsilon_{i0} \quad \varepsilon_{i0} \sim N(0, \sigma_0^2)\ , \tag{3.14}$$

where $\boldsymbol{X}$ is a vector of outcome-model variables, $\boldsymbol{\beta}$ is a vector of outcome-model parameters, and ε_i is an error term. Y_1 is observed if $z^* > 0$ and Y_2 is observed if $z^* < 0$.

The error terms in the system follow the trivariate normal distribution:

$$\begin{pmatrix} u_i \\ \varepsilon_{i1} \\ \varepsilon_{i0} \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho_1\sigma_1 & \rho_0\sigma_0 \\ \rho_1\sigma_1 & \sigma_1^2 & 0 \\ \rho_0\sigma_0 & 0 & \sigma_0^2 \end{pmatrix}\right] \tag{3.15}$$

The variance of u_i is fixed as 1 for convenience and identification, while ρ_1 and ρ_0 represent the correlations of the unobserved variables influencing class membership with those influencing the respective outcomes for class 1 and class 0 members. Note that the covariance between ε_{i1} and ε_{i0} is fixed at zero since everyone can belong to only one of the two classes and thus the correlation is unidentifiable. As noted by Greene (2012), the choice of zero is merely for convenience and it does not play a role in the estimation of the model coefficients.

Here, we need to be clear about a difference between exogenous and endogenous switching. This distinction is important because it affects how we characterize the model type. A critical question is, how do we define the measurement spaces of ε_1 and ε_0? As will be described later, in mixture modeling, ε_1 and ε_0 are *defined over the subpopulations* of class 1 and class 0, respectively. What about for the switching model? If it is exogenous

switching, ε_1 and ε_0 can be defined over either the population or the subpopulation. In the latter case, exogenous switching can be considered a special type of mixture modeling where the class indicator is known (which is the model in Section 3.3.2). In the case of endogenous switching, however, ε_1 and ε_0 must be *defined over the population* since we are allowing correlations with u (which is defined over the population)[24].

For the switching regression model, the conditional class-specific expectations can be expressed as:

$$E(Y_{i1}|z_i = 1, \boldsymbol{X}_i, \boldsymbol{W}_i) = \boldsymbol{X}_i\boldsymbol{\beta}_1 + E[\varepsilon_{i1}|z_i = 1, \boldsymbol{X}_i, \boldsymbol{W}_i] = \boldsymbol{X}_i\boldsymbol{\beta}_1 + \rho_1\sigma_1 \frac{\phi(\boldsymbol{W}_i\boldsymbol{\alpha})}{\Phi(\boldsymbol{W}_i\boldsymbol{\alpha})} \quad (3.16)$$

$$E(Y_{i0}|z_i = 0, \boldsymbol{X}_i, \boldsymbol{W}_i) = \boldsymbol{X}_i\boldsymbol{\beta}_0 + E[\varepsilon_{i0}|z_i = 0, \boldsymbol{X}_i, \boldsymbol{W}_i] = \boldsymbol{X}_i\boldsymbol{\beta}_0 + \rho_0\sigma_0 \left[\frac{-\phi(\boldsymbol{W}_i\boldsymbol{\alpha})}{1-\Phi(\boldsymbol{W}_i\boldsymbol{\alpha})}\right] \quad (3.17)$$

where the second terms in each equation (compare Eqs. (3.4) and (3.5) for the deterministic segmentation model) reflect the selection bias in the conditional expected outcome when there is a non-zero correlation between unobserved influences on class membership and outcome. The factor multiplying ρ and σ in each equation is known as the inverse Mills ratio.

The overall (i.e. unconditional with respect to class) density is:

[24] The confusion can arise because the statement "Y_1 is observed if $z^* > 0$" sounds like it is a conditional distribution. However, as noted by Maddala (1986), such a statement does not necessarily mean that the disturbance term should be specified only for that subpopulation ($z^* > 0$). In the endogenous switching model, both disturbance terms are defined over the population and this approach will be at the heart of selection modeling.

$$f(y_i|\boldsymbol{X}_i, \boldsymbol{W}_i) = P(z_i = 1|\boldsymbol{W}_i) \times f(y_i|z_i = 1, \boldsymbol{X}_i, \boldsymbol{W}_i) + P(z_i = 0|\boldsymbol{W}_i) \times f(y_i|z_i = 0, \boldsymbol{X}_i, \boldsymbol{W}_i) \tag{3.18}$$

The log-likelihood (LL) can be expressed as follows:

$$LL = \sum_{z=1} \ln \left[\Phi \left(\frac{\boldsymbol{W}_i\boldsymbol{\alpha} + \rho_1 (y_i - \boldsymbol{X}_i\boldsymbol{\beta}_1)/\sigma_1}{\sqrt{1-\rho_1^2}} \right) \times \left[\frac{1}{\sigma_1} \phi \left(\frac{y_i - \boldsymbol{X}_i\boldsymbol{\beta}_1}{\sigma_1} \right) \right] \right] + \sum_{z=0} \ln \left[\Phi \left(- \frac{\boldsymbol{W}_i\boldsymbol{\alpha} + \rho_0 (y_i - \boldsymbol{X}_i\boldsymbol{\beta}_0)/\sigma_0}{\sqrt{1-\rho_0^2}} \right) \times \left[\frac{1}{\sigma_0} \phi \left(\frac{y_i - \boldsymbol{X}_i\boldsymbol{\beta}_0}{\sigma_0} \right) \right] \right]. \tag{3.19}$$

Several comments are in order. First, deterministic segmentation and switching regression models share the feature that we *know*, in the sample, the group to which each individual belongs (i.e. the group indicator of interest). Second, if ρ_0 and ρ_1 are both 0, i.e. if we have exogenous rather than endogenous switching, then the conditional expectations (Eqs. (3.16) and (3.17)) are identical to those of the deterministic segmentation model (Eqs. (3.4) and (3.5)). In that case, the two models would differ only in their unconditional expectations and densities, where the deterministic segmentation model replaces the segment membership probability weights of the exogenous switching model with 0s and 1s (as can be seen by comparing the argument of the natural log function in the first line of Eq. (3.9) with Eq. (3.18)), reflecting the certainty of segment membership.

As generally applied, switching regression aims to be able to predict the *expected* outcome for a *randomly-selected individual*, and therefore needs to incorporate the probability of belonging to one group or the other, whereas analysts using the deterministic

(exogenous) segmentation model typically content themselves with explaining outcomes *conditional on segment membership*. But *if* unconditional as well as conditional outcomes are a key interest, then the foregoing discussion raises the question: when ρ_0 and ρ_1 are both 0, when should deterministic segmentation be used, as opposed to exogenous switching? More precisely (since, again, the conditional equations would be the same for both models), when is it useful or necessary to estimate a segment membership model as well as segment-specific outcome models? Note that a segment membership model *could* be estimated in a deterministic segmentation context (and when ρ_0 and ρ_1 are both 0, it would be identical to its exogenous switching counterpart) – it is just that it usually is not needed and therefore not estimated when segmenting deterministically. The answer to the question lies in how an estimated model is intended to be used. Clearly, if the model is to be used in a predictive, out-of-sample capacity, where for new cases $\boldsymbol{X}_i$ and $\boldsymbol{W}_i$ are known (e.g. for a synthetically-generated population) but not class membership z_i, then the segment membership model (Eqs. (3.11) and (3.12)) is a "must", to enable prediction of class membership. For in-sample applications, by contrast (e.g. to scenario analysis), if $\boldsymbol{X}_i$ variables are changed but not $\boldsymbol{W}_i$ variables (i.e. if there is no reason to expect segment membership to change), then it would seem appropriate to continue to reflect the certain knowledge of segment membership by using the deterministic segmentation model without a membership model[25]. On the other hand, even for in-sample applications, if desired scenarios are likely to involve shifts among segments, then the segment membership model is again indispensable.

[25] Of course, any application to population-level analysis presumes that the sample is either representative of the population – *particularly with respect to the segment membership shares* (which is often not the case) – or else weighted to be so.

Furthermore, if ρ_0 and ρ_1 are *not* 0, then when conditioning on segment membership, the error distributions of the outcome equations are *truncated* (the related equations will be delineated in Section 3.5.1 and Appendix A). In this case, *even if the analyst only cares about outcomes conditional on segment membership*, the estimators of $\boldsymbol{\beta}$ that are obtained from the deterministic segmentation model are inconsistent, because (by assuming *untruncated* error distributions) they are absorbing the last terms of Eqs. (3.16) and (3.17). Put more plainly, the point (well-known in sample selection settings but not routinely taken into account in deterministic segmentation contexts) is that if unobserved factors associated with individuals' selection into segments are correlated with those influencing the outcome of interest, then effects of the explanatory variables $\boldsymbol{X}_i$, *obtained when conditioning on segment*, will be improperly estimated if not corrected for the presence of that correlation. This suggests that it *should* perhaps become routine, before using deterministic segmentation, to estimate an endogenous switching model to test whether ρ_0 and ρ_1 are both 0.

3.3.4 Latent class model

As noted, in the aforementioned models we know each individual's group membership. Here, we posit that there are some underlying sub-populations having different decision processes and/or distinctive distributions, but we *do not know* who belongs to each sub-population (portrayed by the membership indicator *z* becoming an oval instead of a square in Figure 3-1). Hence, we aim to uncover the population segments themselves, as well as their different behavioral models. Because we never know the ground truth of segment membership, we treat such membership as a (discrete) latent variable. In essence, group membership can be considered to be *completely missing data*

and thus unobservable (Little and Rubin, 2019). In addition, since we rely on the data to identify the latent segments, this latent class approach can be considered as more "data-driven" than the other two models[26]. The way we functionally formulate latent class models is to use finite mixture modeling. The membership model is as follows:

$$z_i^* = \boldsymbol{W}_i\boldsymbol{\alpha} + u_i \quad u_i \sim \text{EV}(0,1)\ ^{27} \tag{3.20}$$

$$P(z_i = 1) = 1/[1 + \exp(-\boldsymbol{W}_i\boldsymbol{\alpha})]\ ^{28} \tag{3.21}$$

$$P(z_i = 0) = 1 - P(z_i = 1)\ = 1/[1 + \exp(\boldsymbol{W}_i\boldsymbol{\alpha})]\ . \tag{3.22}$$

The two class-specific outcome models are:

$$Y_{i1} = \boldsymbol{X}_i\boldsymbol{\beta}_1 + \varepsilon_{i1} \quad \varepsilon_{i1} \sim N(0, \sigma_1^2) \tag{3.23}$$

$$Y_{i0} = \boldsymbol{X}_i\boldsymbol{\beta}_0 + \varepsilon_{i0} \quad \varepsilon_{i0} \sim N(0, \sigma_0^2)\ . \tag{3.24}$$

[26] In other words, deterministic segmentation and endogenous switching regression could be considered as more theory-driven in that, *a priori*, we speculate that certain identifiable segments (e.g. age cohorts) have different behavioral models, or aim to test the "treatment effect" of a certain identifiable treatment (e.g. living in an urban versus non-urban area in the residential self-selection context). By contrast, latent class models aim to find *optimal* (but previously unknown) segmentations *with respect to* the target variable, because we maximize an objective function that includes membership-function terms. However, the distinction between being theory- and data-driven is continuous rather than binary, because theory generally suggests the variables to be included in the membership function of a latent class model, while data (i.e. empirical results) generally influence the final specifications of the other two models.

[27] As discussed further in Section 3.5.2, most latent class modeling applications employ the logit link function for the membership model and thus assume that the error term follows the extreme value distribution with location parameter equal to zero (without loss of generality as long as a constant term is being estimated) and scale parameter fixed to unity (for convenience and identifiability). However, in theory, it can be formulated with the probit model and thus have a normally distributed error term.

[28] This study focuses on two classes. Membership probabilities can be specified with the multinomial logit model if there are more than two classes.

The conditional class-specific expectations are:

$$E(Y_{i1}|z_i = 1, \boldsymbol{X}_i) = \boldsymbol{X}_i\boldsymbol{\beta}_1 \tag{3.25}$$

$$E(Y_{i0}|z_i = 0, \boldsymbol{X}_i) = \boldsymbol{X}_i\boldsymbol{\beta}_0 \ . \tag{3.26}$$

The overall density is:

$$\begin{aligned} f(y_i|\boldsymbol{X}_i, \boldsymbol{W}_i) = P(z_i = 1|\boldsymbol{W}_i) \times f(y_i|\boldsymbol{X}_i, z_i = 1) \\ + P(z_i = 0|\boldsymbol{W}_i) \times f(y_i|\boldsymbol{X}_i, z_i = 0) \end{aligned} \tag{3.27}$$

The log-likelihood (LL) can be expressed as follows:

$$\begin{aligned} LL &= \sum_{i=1}^{N} \ln\left[\sum_{z_i=0,1} P(z_i|\boldsymbol{W}_i) \times f(y_i|\boldsymbol{X}_i, z_i)\right] \\ &= \sum_{i=1}^{N} \ln\left[\frac{1}{1+\exp(-\boldsymbol{W}_i\boldsymbol{\alpha})} \times \frac{1}{\sigma_1}\phi\left(\frac{Y_i - \boldsymbol{X}_i\boldsymbol{\beta}_1}{\sigma_1}\right) + \frac{1}{1+\exp(\boldsymbol{W}_i\boldsymbol{\alpha})} \times \frac{1}{\sigma_0}\phi\left(\frac{Y_i - \boldsymbol{X}_i\boldsymbol{\beta}_0}{\sigma_0}\right)\right]. \end{aligned} \tag{3.28}$$

It is instructive to examine closely the difference in the log-likelihoods of the switching regression model (Eq. 3.19) and the latent class model (Eq. 3.28) (setting aside the variant functional forms of the segment membership probabilities – which, as indicated in footnote 8, differ only by convention, not by necessity – and the ability of the former to account for correlated error terms). Each term in both equations involves outcome densities multiplied by segment membership probabilities, reflecting the contribution of both elements to parameter estimation and evaluation of the model's performance. But the log-likelihood of the switching regression model can be split into two sums, corresponding to

the known members of each of the two groups, and accordingly only the membership probability and outcome density associated with the segment to which the individual belongs need to be included in each term. In essence, each term is the natural log of the *joint density* of belonging to the observed group *z* and experiencing the observed outcome given belonging to *z*. The log-likelihood of the latent class model, on the other hand, cannot be split in that way because group membership is not known, and accordingly each term must include the membership probability and outcome density of both possible segments. In essence, each term is the natural log of the "*total probability*" (actually the *expected density*, for continuous outcomes) of experiencing the observed outcome. Thus, the former is the log-likelihood of a *joint* event (membership and outcome) while the latter is the (*marginal*) log-likelihood of a *single* event (outcome, summed over the membership probability distribution), and consequently the two log-likelihoods are not directly comparable. We will discuss the practical issue caused by this difference in Section 3.4.4.

The deterministic segmentation model, on the other hand, can be considered as a special case of both models, where we know the binary indicator of membership (i.e. the probability of group membership, $P(z_i)$, is either 1 or 0). In that case, all the probabilities of Eq. (3.19) for switching regression are 1, simplifying to Eq. (3.9), whereas in Eq. (3.28) for the latent class model, one of the probabilities in each term is 1 while the other one is 0, again simplifying to Eq. (3.9).

3.4 Empirical application

3.4.1 Data

This study employs the GDOT data (Section 1.2.3). This study focuses on Georgia residents who have a driver's license and drive. The dependent variable is self-reported weekly vehicle-miles driven (VMD), which is often studied as a major travel behavior indicator. As usual, VMD has a skewed distribution; hence we log-transform it (first adding 1, to avoid taking the log of zero) to more closely achieve normality. Specifically for the deterministic segmentation and endogenous switching models, we divide the sample into urban and non-urban residents based on the population density of their residential Census block group[29]. This is a common segmentation variable in the literature associated with investigating the effect of neighborhood type on travel behavior, as described in Section 3.2. For the sake of comparison, we employ the same set of variables (some attitudes, demographics, and population density) for the membership function in the latent class model. The original dataset consists of more than 3,200 cases, but it is reduced to 3,022 after excluding non-drivers and cases with missing values on key variables (Table 3-1). Based on geocodes of the home location, we collected additional land use information from the American Community Survey (ACS), the Longitudinal Employer-Household Dynamics (LEHD) database, Alltransit, and Google Place API. The descriptive statistics

[29] Since there is no universally-accepted definition of "urban", we segment the sample into urban and non-urban residents based on population density (people per acre; census block-group level American Community Survey 2017 estimates): urban if the population density is 4 people per acre (2,560 per square mile) or higher, and non-urban otherwise. We have a subjective measure of neighborhood type in the survey and an urbanized area designation in the census data, but they are too broad to effectively capture the effects being sought. Our definition of urban area is more conservative than the census designation – i.e. we draw the boundary at a higher density than the census does.

of the key variables are reported in Table 3-4 rather than in this section because it is more useful to compare the sample characteristics with other information.

Table 3-1. Variable descriptions

Variable	Description	Source
Socio-demographics		
Gender	Female dummy	GDOT survey
Race	White dummy	GDOT survey
Age	18-34 dummy	GDOT survey
	65+ dummy	GDOT survey
Income	Middle income ($50,000 - $99,999) dummy	GDOT survey
	Higher income ($100,000 or more) dummy	GDOT survey
Telecommute	Dummies: weekly-based telecommuting, no or infrequent telecommuting (base: non-worker)	GDOT survey
Household	Number of household members	GDOT survey
	Number of vehicles	GDOT survey
MPO type	Atlanta region dummy	GDOT survey
Attitudes [a]		
	Pro-environmental	GDOT survey
	Urbanite	GDOT survey
	Travel-liking	GDOT survey
	Pro-car-owning	GDOT survey
Land use		
Density	Population density (people per acre)	ACS
	Job density (jobs per acre)	LEHD
Transit score	Level of transit service [0-1]	Alltransit
Local accessibility [b]	PCA score of local amenities	Google Place API

a. Selected factor scores from a factor analysis. Selected attitudinal factors and highly-loading statements are reported in Table 1-1. The full factor analysis solution is reported in Kim et al. (2019b) as well.
b. Principal component analysis (PCA) score of the first dimension which captures the largest portion of variation in number of amenities near home location. Amenities include 23 types of places such as restaurant, bar, store, and café. Pattern loadings are reported in Table 1-2.

3.4.2 Estimation results

For modeling VMD, based on the literature and empirical experimentation, we consider three sets of explanatory variables: demographics, geographic characteristics, and work-related characteristics. In Table 3-2, the pooled model shows mostly statistically significant parameters and those are consistent with conceptual expectations. Males, whites, higher income people, and those in a middle age group (35-64) tend to drive more. Atlanta residents, on average, drive more. Three geographic characteristics of the

residential vicinity (job density, transit score, and a proxy for local amenities) present negative parameters, indicating that accessibility and availability of transit reduce VMD in general. Compared to non-workers, workers generate more VMD and we can observe, on average, that workers who telecommute at least once a week generate less VMD than non- or infrequent telecommuters.

Turning to the deterministic segmentation model, it shows some notable parameter heterogeneity across urban and non-urban segments (Table 3-2). For example, among urban residents younger people (18-34) tend to drive less than others, whereas among non-urban residents they do not. Oppositely, older people (65 or higher) tend to drive less than others among *non-urban* residents, while those in *urban* areas do not. The impact of income on VMD is greater in urban than in non-urban areas; on average, *ceteris paribus*, switching from being lower income to being medium or higher income respectively leads to 49% and 64% increases $[(\exp 0.40 - 1) * 100\%$ and $(\exp 0.50 - 1) * 100\%)]$ in (VMD+1) for urban residents, compared to 25% and 42% for non-urban residents. Urban residents are more sensitive to the availability of transit than non-urban residents are, whereas non-urban residents are more sensitive to local amenities than urban residents are. For the latter, we speculate the reason to be that urban residents tend to have more local amenities as a baseline, and thus marginal differences in the number of amenities in their activity radius may not significantly affect their overall VMD. However, for non-urban residents, if there are not enough local amenities within their activity radius, they may need to drive farther and thus increase average VMD.[30]

[30] Note that there are some non-negligible correlations among the geographic variables (job density, transit score, local accessibility, and Atlanta MPO resident; rows/columns are in this order). The respective

Table 3-3 exhibits the estimation results of the endogenous switching and latent class models. Turning first to the endogenous switching model, the membership (or sample selection) model, which we do not have in the deterministic segmentation approach, shows how randomly-selected individuals choose their residence between urban and non-urban areas. Those with stronger pro-environmental and urbanite attitudes are more likely than others to choose to live in urban areas, whereas those who like traveling and owning/driving cars are more likely than others to choose to live in non-urban areas. In addition, workers and those having fewer vehicles and smaller households are more likely than others to live in urban areas. All of these results are plausible and, for the most part, expected. The outcome models, in general, are consistent with the results of the deterministic segmentation model.

As we simultaneously estimate "membership" in neighborhood type and VMD, the key merits over the deterministic segmentation model are twofold. First, the endogenous switching model enables us to explain what kinds of people are more likely to belong to each neighborhood type. For the deterministic segmentation model we can only explain VMD generation *given that we know which neighborhood type a person lives in*, whereas for the endogenous switching model we can predict VMD for a *randomly-selected* person. This might be meaningful for demand forecasting in that analysts do not know which person will live in which neighborhood type. Additionally, error correlation estimates corroborate that the selection of neighborhood type and the generation of VMD share

---

correlation matrices for the pooled sample, urban area, and non-urban area segments are:

$$\begin{bmatrix} 1.00 & 0.43 & 0.50 & 0.15 \\ & 1.00 & 0.62 & 0.28 \\ & & 1.00 & 0.16 \\ & & & 1.00 \end{bmatrix}, \begin{bmatrix} 1.00 & 0.36 & 0.47 & 0.12 \\ & 1.00 & 0.55 & 0.24 \\ & & 1.00 & 0.03 \\ & & & 1.00 \end{bmatrix}, \begin{bmatrix} 1.00 & 0.43 & 0.46 & 0.10 \\ & 1.00 & 0.45 & 0.16 \\ & & 1.00 & 0.07 \\ & & & 1.00 \end{bmatrix}.$$

substantial unobserved characteristics. Hence, failing to account for those correlations among unobserved variables leads to inconsistent estimates of the coefficients of the deterministic segmentation model. Indeed, comparing the inconsistent coefficient estimates of the deterministic segmentation model (Table 3-2) to the consistent ones of the endogenous switching model (Table 3-3) indicates that while many coefficients are quite similar, there are also some substantial differences. For example, the deterministic segmentation model appears to exaggerate the importance of transit accessibility for urban residents and being white for non-urban residents, and to understate the importance of being a (tele)worker for both groups.

The latent class model also involves a membership model and two outcome models (Table 3-3). For the purposes of comparison, we estimated only a two-class model. We specified the class membership model to have the same explanatory variables as those in the selection equation of the endogenous switching model, plus a population density variable that (since neighborhood types were defined on the basis of density) would have been essentially tautological to include in the selection equation. For the most part the same variables were significant with the same signs in the two models, with the exception that workers were more likely than others to live in urban areas for the endogenous segmentation model, but less likely than others to belong to "class 1" for the latent class model. Otherwise, however, those who are more urbanite and pro-environmental, who like travel and owning cars less, and who tend to own fewer vehicles and live in smaller households – traits that are also largely identified with urban residents – are more likely to be latent class 1 members.

Accordingly, it might be expected that the two latent classes respectively match the urban and non-urban segments of the previous two models. However, neither the membership model nor the outcome models fully support that presumption. First, as will be shown in Section 3.4.3, the latent class membership is somewhat different than the membership in the deterministic urban and non-urban classes (as can already be seen in Table 3-3 from the fact that 23% of the sample live in an urban area, whereas 52% belong to latent class 1), although latent class 1 *tends* to be more urban than class 0 as expected. On the basis of results shown in Section 3.4.3, we label class 1 as "lower VMD-inclined" and class 0 as "higher VMD-inclined".

In addition, when investigating the parameters in the two outcome models, we can observe that patterns of sensitivities to factors for the two latent classes are different from those for the deterministic groups in the previous two models. For example, relative to having a lower income, being of middle income has a substantially weaker (positive) impact on VMD for class 1 (lower VMD) of the latent class model than for the urban classes of the previous two models, and similarly for the (negative) impact of the transit score. The latter result (as corroborated by Table 3-4 in Section 3.4.3) is likely because members of latent class 1 are more scattered between urban and non-urban areas, and there may be little variability in transit scores for non-urban areas, which will dilute the impact of that variable across the segment. On the other hand, relative to non-workers, non- or infrequent telecommuting has a much smaller (positive) impact on VMD for latent class 0 (higher VMD) than for the non-urban segments of the previous two models, while weekly

telecommuting has a smaller and insignificant impact for latent class 1 compared to its impact for the urban segments of the other two models[31].

In essence, given the variables proposed for the membership model, the latent class model identifies two distinct groups on the basis of their VMD model coefficients (finding the optimum balance between within-group homogeneity and between-group heterogeneity in that respect), without direct regard to whether they are urban or non-urban residents. Put another way, urban residents can "behave" like (stereotypical) non-urban residents with respect to influences on their VMD, and conversely; this model can more flexibly group the "non-urban-like" urban residents with their likeminded non-urban counterparts, and conversely for the "urban-like" non-urban residents.

Table 3-2. Estimation results for the pooled and deterministic segmentation models (N=3,022)

	Pooled		**Deterministic segmentation**			
			Urban (Class 1)	(23%)	Non-urban (Class 0)	(77%)
Variable	Parameter	t-value	Parameter	t-value	Parameter	t-value
Intercept	**4.154**	**71.99**	**3.993**	**34.67**	**4.190**	**61.98**
Female	**-0.241**	**-7.36**	**-0.270**	**-3.98**	**-0.232**	**-6.21**
White	**0.152**	**3.84**	0.138	1.89	**0.153**	**3.20**
Age:18-34	-0.058	-0.98	**-0.280**	**-2.68**	0.037	0.52
Age:65+	-0.069	-1.72	0.011	0.13	**-0.095**	**-2.09**
Middle income	**0.277**	**6.91**	**0.402**	**4.80**	**0.228**	**5.01**
Higher income	**0.400**	**9.16**	**0.498**	**5.45**	**0.356**	**7.14**
Atlanta residence	**0.108**	**2.98**	**0.294**	**4.11**	0.054	1.28
Job density (jobs per acre)	**-0.012**	**-2.59**	**-0.014**	**-2.44**	-0.017	-1.86
Transit score	**-0.480**	**-5.15**	**-0.675**	**-4.93**	-0.217	-1.53
Local accessibility	**-0.072**	**-3.30**	-0.021	-0.65	**-0.085**	**-2.66**
Weekly telecommuting	**0.274**	**4.78**	**0.233**	**2.04**	**0.287**	**4.32**
No or infrequent telecommuting	**0.524**	**12.64**	**0.534**	**6.18**	**0.526**	**11.15**

Note: Coefficients statistically significant at the 0.05 level are bolded. The goodness-of-fit measures of these models are shown in Table 3-5 and discussed in Section 3.4.4.

[31] We note in passing that for all models and segments, frequent telecommuters are associated with lower VMD than non-frequent or non-telecommuting workers, all else equal, consistent with typical findings of longitudinal studies (e.g. Mokhtarian et al., 1995) that telecommuting reduces an individual's travel, on net.

Table 3-3. Estimation results for the endogenous switching and latent class models (N=3,022)

	Endogenous switching model				Latent class model			
	Urban (Class 1)	(23%)	Non-urban (Class 0)	(77%)	Lower VMD (Class 1)	(52%)	Higher VMD (Class 0)	(48%)
Variable	Parameter	t-value	Parameter	t-value	Parameter	t-value	Parameter	t-value
Outcome models								
Intercept	**3.000**	**13.16**	**4.495**	**63.31**	**3.595**	**25.72**	**4.936**	**36.98**
Female	**-0.243**	**-3.60**	**-0.213**	**-5.88**	**-0.191**	**-3.07**	**-0.270**	**-5.59**
White	0.123	1.69	0.084	1.78	**0.244**	**3.47**	-0.047	-0.70
Age:18-34	**-0.278**	**-2.67**	0.037	0.55	-0.150	-1.44	0.005	0.07
Age:65+	0.022	0.26	**-0.096**	**-2.12**	0.117	1.59	**-0.272**	**-3.94**
Middle income	**0.399**	**4.81**	**0.232**	**5.24**	**0.267**	**3.85**	**0.243**	**4.12**
Higher income	**0.481**	**5.29**	**0.321**	**6.60**	**0.448**	**5.75**	**0.296**	**4.90**
Atlanta residence	**0.300**	**4.23**	0.078	1.91	0.108	1.62	**0.132**	**2.66**
Job density (jobs per acre)	**-0.013**	**-2.34**	-0.012	-1.50	-0.013	-1.91	-0.006	-0.52
Transit score	**-0.555**	**-4.03**	-0.163	-1.20	**-0.356**	**-2.31**	-0.193	-1.14
Local accessibility	-0.013	-0.42	**-0.074**	**-2.42**	-0.031	-0.82	-0.060	-1.54
Weekly telecommuting	**0.392**	**3.28**	**0.419**	**6.23**	0.104	0.82	**0.249**	**2.68**
No or infrequent telecommuting	**0.677**	**7.17**	**0.630**	**12.47**	**0.498**	**4.52**	**0.286**	**3.66**
Membership model								
Intercept	**-0.677**	**-11.44**	-	-	0.523	1.37	-	-
Pro-environmental	**0.043**	**1.97**	-	-	**0.308**	**3.50**	-	-
Urbanite	**0.143**	**6.14**	-	-	**0.169**	**2.05**	-	-
Travel-liking	**-0.055**	**-2.62**	-	-	**-0.195**	**-2.31**	-	-
Pro-car-owning	**-0.127**	**-5.91**	-	-	**-0.425**	**-4.38**	-	-
Worker	**0.309**	**5.96**	-	-	-0.649	-1.85	-	-
Number of household vehicles	**-0.078**	**-3.80**	-	-	**-0.151**	**-2.02**	-	-
Household size	-0.035	-1.77	-	-	-0.008	-0.11	-	-
Population density	-	-	-	-	**0.113**	**1.99**	-	-
Additional parameters								
Sigma	**1.033**	**14.83**	**1.022**	**45.29**	**0.880**	**34.46**	**0.605**	**21.57**
Error correlation	**0.660**	**5.22**	**0.844**	**13.22**	-	-	-	-

Note: Coefficients statistically significant at the 0.05 level are bolded. The goodness-of-fit measures of these models are shown in Table 3-5 and discussed in Section 3.4.4.

3.4.3 How do segments differ across models?

Given the differences that have already been alluded to in class membership across models, it is natural to examine these differences more closely. As we touched on in the previous section, the two groups in the latent class model are distinct with respect to their average weekly VMD levels. Specifically, the average VMDs (based on prior probabilities) are 132 and 165 miles, respectively, for the latent class model, and are 118 and 160 miles respectively in urban and non-urban areas. Thus, as expected, in both models urban-oriented residents have lower VMD. On average, both within-group averages are higher for the latent class model, which is also as expected. Latent class 1 (with 52% of the cases) must have collected a sizable fraction of (residentially mismatched, i.e. "attitudinally urban") non-urban residents (since non-urban residents as a whole comprise 77% of the cases). The weekly VMD for those non-urban, latent class 1 cases will tend to be higher than that of matched *urban* residents (because of the pull that their *non-urban built environment* exerts on their travel behavior) and lower than that of matched *non-urban* residents (because of the pull that their *urban attitudes* exert on their travel behavior). The net result is that the average VMD of latent class 1 is higher than that of urban residents as a whole. At the same time, latent class 0 has "lost" (to class 1) a number of non-urban residents with lower-than-(*non-urban*-)average VMD, and presumably gained some urban residents with higher-than-(*urban*-)average VMD, with the net result that its average VMD is higher than that of non-urban residents as a whole[32] (Schwanen and Mokhtarian, 2005). Thus, the latent class model is better able to group similarly-minded individuals regardless of their residential location type.

[32] This discussion unavoidably reminds us of statistician Frederick Mosteller's (possibly apocryphal).

The greater within-group homogeneity of the latent class model is also attested by the sigma parameters (i.e. square roots of error variances). The estimates are a bit greater than one (1.03 and 1.02) in the endogenous switching model, while they are much lower (0.88 and 0.61) in the latent class model. In other words, the latent class model uncovers segments having less variability with respect to unobserved influences on VMD. This is because the latent class model identifies underlying groups specifically based on their VMD distributions, whereas the endogenous switching model segments individuals based on a certain indicator (here, neighborhood type) that is not necessarily directly tied to VMD.

Table 3-4 shows profiles of the pooled, urban, non-urban, and two latent classes. Let us first compare the two latent classes. Compared to class 1 (lower VMD-inclined), class 0 (higher VMD-inclined) has the following characteristics, on average: higher income, more often white, living in smaller MPO areas with lower population density and lower accessibility, less urbanite, less pro-environmental, owning a higher number of vehicles, and more favorable toward owning vehicles. These tendencies are generally consistent with the contrasts between urban and non-urban segments. When it comes to built environment characteristics, however, the contrasts between latent classes are (understandably, given the previous discussion) less sharp than those between groups based purely on residential location. Specifically, the gaps between urban and non-urban segments on population density, transit scores, number of stores, and local accessibility are markedly wider than those between the lower- and higher-VMD classes. In addition, there

remark that "by leaving the Princeton math department to join the math department at Harvard he succeeded in raising the average IQ in both places" (Wainer, 1999, p. 44).

are more drastic differences in age, race, and MPO type between urban and non-urban than between lower- and higher-VMD classes.

The latent classes also help explain some otherwise puzzling patterns with respect to two of the four attitudes. Urban dwellers are apparently less pro-environmental than non-urban residents, which may be counter to stereotype, but the latent class model confirms that it is the higher-VMD individuals – regardless of where they live – who tend to be less pro-environmental. Similarly, there is a slight tendency for urban residents to like traveling (despite the significant negative coefficient of that variable in the segment membership model of the switching regression system; Table 3-3), but higher-VMD individuals like traveling considerably more, on average. Thus, at least for these two variables, there are enough non-stereotypical people in urban and non-urban neighborhoods to yield potentially non-intuitive "average" attitudes, but when sorting people by VMD proclivities irrespective of residential location, the stereotypes hold true.

Table 3-4. Profiles of segments (N=3,022)

Variable		Pooled	Urban	Non-urban	Class 1: lower VMD	Class 0: higher VMD
Distribution						
Gender	Female	51%	55%	50%	53%	49%
Age	18-34	23%	34%	18%	23%	22%
	35-44	17%	17%	17%	16%	18%
	45-64	39%	33%	42%	37%	42%
	65+	21%	16%	23%	24%	18%
Income	Lower	39%	38%	39%	43%	35%
	Medium	33%	35%	33%	31%	36%
	Higher	27%	27%	28%	25%	30%
Race	White	63%	52%	68%	60%	67%
	Black	29%	38%	25%	31%	26%
	Else	8%	10%	7%	9%	7%
Telecommute	Non-worker	35%	28%	38%	41%	29%
	No or infrequent telecommuting	51%	57%	48%	46%	57%
	Weekly-based telecommuting	14%	15%	13%	13%	15%
MPO tier	ATL	53%	77%	43%	56%	49%
	Mid-sized MPO	17%	15%	17%	16%	17%
	Small-sized MPO	13%	7%	16%	12%	14%
	Rural areas	17%	1%	24%	15%	19%
Mean						
	VMD	148.06	118.58	160.90	132.35	165.22
	ln(VMD+1)	4.55	4.31	4.65	4.41	4.70
	Number of vehicles	2.10	1.81	2.23	1.90	2.32
	Household size	2.34	2.21	2.39	2.23	2.45
	Population density	3.32	7.74	1.39	4.08	2.50
	Urbanite	0.11	0.42	-0.03	0.22	-0.01
	Pro-environmental	-0.07	-0.12	-0.05	0.07	-0.22
	Travel-liking	0.03	0.08	0.01	-0.05	0.11
	Pro-car-owning	0.04	-0.18	0.14	-0.18	0.28
	Transit score	0.19	0.42	0.09	0.23	0.14
	Number of stores	10.64	16.91	7.90	11.77	9.42
	Local accessibility	0.08	0.90	-0.27	0.25	-0.10

Note: all statistics are case-weighted to correct for sampling biases with respect to MPO size, income, household size, vehicle ownership, gender, education, race, age, and work status.

3.4.4 Model performance

In addition to the behavioral insights provided by the models, their goodness of fit is also of interest. First, we examine the final log-likelihood values, together with some information criteria that are most commonly considered in applications of latent class

modeling (Table 3-5). Due to having two classes, the deterministic segmentation model contains more information than the pooled model, at the cost of doubling the number of parameters; the latent class model contains even more information (requiring even more parameters) because of its membership model. Three information criteria penalize the model complexity, with the degree of penalization highest for the Consistent Akaike Information Criterion (CAIC), followed by the Bayesian Information Criterion (BIC) and then AIC. Hence, although the deterministic segmentation model is better than the pooled model with respect to AIC, it is not better with respect to BIC and CAIC. In other words, with a stronger penalty for complexity, the increment of model improvement is not sufficient to compensate for its additional complexity. On the other hand, the more lax AIC and BIC support the parameter-heavy latent class model, whereas with respect to the stricter CAIC, the much simpler pooled model barely edges it out.

Following the discussion in Section 3.3.4, note that, with respect to the log-likelihood and information criteria, we are not able to compare the switching regression model to the others. This is because the switching model estimation is maximizing the *joint likelihood of class membership and outcome*, whereas for the latent class model, class membership is unknown (latent) and so the model estimation is maximizing the *likelihood of the outcome* (expressed as a marginal likelihood obtained by "summing out" over the class membership probability distribution). The deterministic segmentation model, on the other hand, is estimated conditional on class, but since class membership is assumed to be independent of the outcome, the conditional likelihood of the outcome is equal to its marginal likelihood in that model.

Figure 3-2 provides a graphical comparison of model fits. The first panel shows the histogram of observed log-transformed VMD. The remaining panels present the estimated conditional densities for each observation in the sample (i.e. the individual-specific contributions to the likelihood function, conditional on class) [$f(y_i|\boldsymbol{X}_i, z_i = 1)$ or $f(y_i|\boldsymbol{X}_i, z_i = 0)$][33]. Accordingly, for each data point the higher the density (i.e. the higher the model-implied likelihood of observing this outcome y_i given those $\boldsymbol{X}_i, z_i$ values), the better the fit of the model for that observation. On that basis, we can readily see that the latent class model has many data points with higher densities than the other models ever achieve. In particular, compared to the endogenous switching model, *both* classes of the latent class model tend to have higher densities and tighter spreads, and to be more sharply distinct from each other (signifying the greater within-group homogeneity that we have already seen in Section 3.4.3), graphically illustrating the superior fit achievable when the classes are flexibly identified to be best suited for the outcome variable at hand, rather than arbitrarily defined a priori. Collectively, the individual densities for the latent class model well approximate the bi-modal distribution of the observed data, unlike any of the other three models.

Prediction accuracy[34] is also an important measure by which to evaluate model performance. Given that the outcome variable is continuous, we compare R-squared and

[33] For the latent class models, each case is plotted twice – once for $z_i = 1$ and once for $z_i = 0$ – whereas for the deterministic segmentation and switching regression models, each case is plotted once, for its known group indicator. As can be seen in earlier equations, the case-specific densities are a function of $y_i - \boldsymbol{X}_i\boldsymbol{\beta}$ and σ. The density is maximized when $y_i = \boldsymbol{X}_i\boldsymbol{\beta}$ (so the flat tops on the plots in the figure represent the cases where y_i is close or equal to $\boldsymbol{X}_i\boldsymbol{\beta}$), and the height of that maximum density is determined by σ (the narrower the spread of the density function, the taller and more peaked it is).

[34] Here, prediction accuracy is an in-sample measure. A more rigorous comparison would be based on out-of-sample accuracy. Most transportation papers involving latent class modeling have used in-sample measures (with a sizable fraction of papers not even reporting performance). For simplicity we focus on excellence in making in-sample predictions, which is at least an expected precondition for excellence in

root mean squared error (RMSE) measures across models. In terms of R-squared (the higher the better) and RMSE (the lower the better), the models in descending order of goodness of fit are latent class, endogenous switching, deterministic segmentation, and pooled. One thing to note is that the endogenous switching and latent class models each have two different ways of calculating performance. In the case of the endogenous switching model, we can obtain predicted values (or residuals) either (1) conditional on the known urban indicator, or (2) as the membership-probability-weighted average of the predicted values for each class (van Herick and Mokhtarian, 2020a). In the case of the latent class model, we can obtain probability-weighted predicted values using either (1) prior membership or (2) posterior membership probabilities that take the outcomes into account. In a latent class discrete choice modeling context, Kim and Mokhtarian (2018) argued that, although using the posterior probability gives better performance measures, using the prior probability is more appropriate in many cases (in forecasting applications we will know neither class membership nor outcome, in which case using posterior information gives an inappropriate boost to the model's calculated performance). Thus, it is not surprising to see that the *posterior*-weighted R-squared and RMSE for the latent class model are substantially better than the R-squareds and RMSEs for the other three models (regardless of whether the conditional or probability-weighted method is used for the endogenous switching model). It is a bit more surprising to note that the *prior*-weighted measures for the latent class model are essentially equivalent to the conditional measures for the endogenous switching model (which, in turn, are only slightly better than the

making out-of-sample predictions (Kim and Mokhtarian, 2018).

probability-weighted measures). Hence, on the basis of these measures alone, the latent class model is not markedly superior unless posterior probability weights are used.

Table 3-5. Model performance

	Pooled model	Deterministic segmentation	Endogenous switching	Latent class model
Parameters	13	26	38	37
Log-likelihood	-3872.12	-3856.26	-5370.54	-3766.77
AIC	7770.25	7764.51	10817.09	7607.55
BIC	7848.43	7898.40	11045.61	7830.05
CAIC	7861.43	7924.40	11083.61	7865.73
R-squared	0.196	0.204	0.215 (conditional)	0.509 (posterior-weighted)
			0.208 (probability-weighted)	0.216 (prior-weighted)
RMSE	0.871	0.867	0.861 (conditional)	0.681 (posterior-weighted)
			0.865 (probability-weighted)	0.860 (prior-weighted)

Note: In comparing AIC, BIC, CAIC, and RMSE across models, lower values are better. For log-likelihood and R-squared, higher values are better.

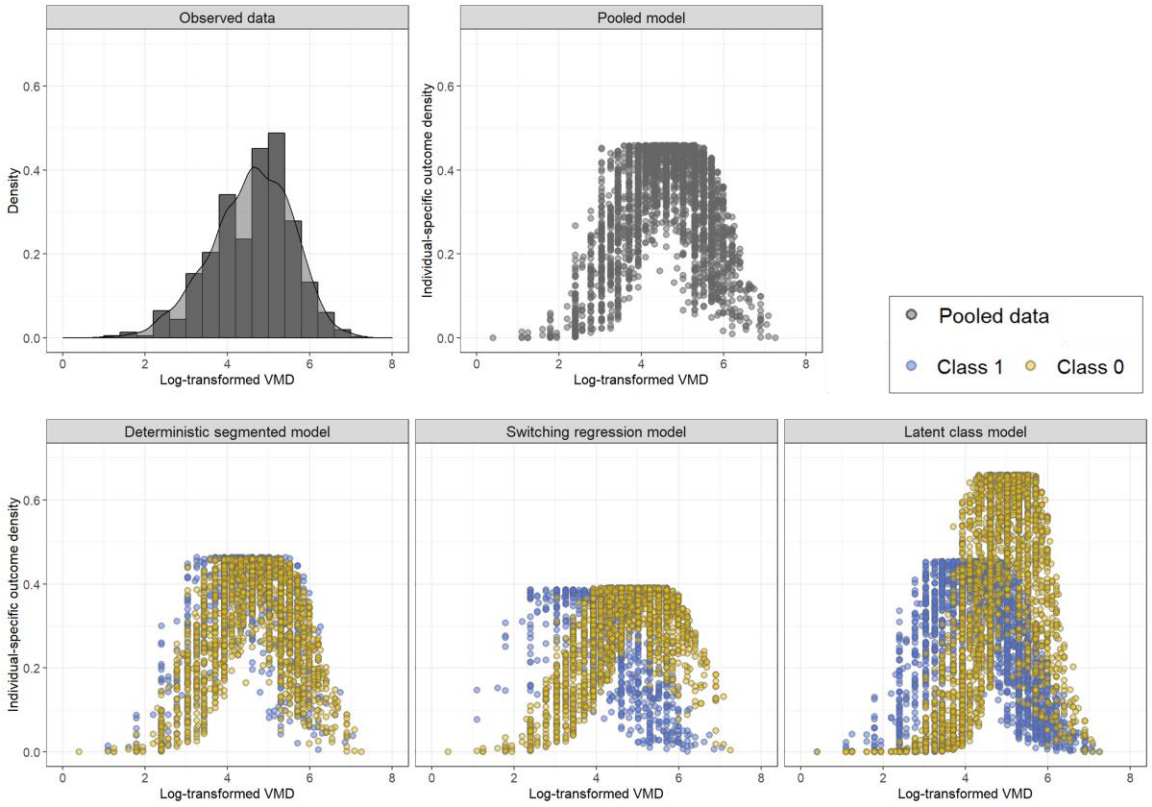


Figure 3-2. Plots of estimated likelihood contributions of each case by model

3.5 Further discussion

This section examines several implications/issues that arise in applications of endogenous switching and mixture models. Section 3.5.1 discusses an additional benefit of the endogenous switching model compared to other competing models, namely its ability to provide a consistent estimate of the effect of a discrete treatment or intervention. Section 3.5.2 addresses how membership functions can be specified (particularly related to the latent class model) and how they behave for the different models. Section 3.5.3 briefly connects the models presented in this study with mixture modeling in a machine learning context (specifically under a neural network structure).

3.5.1 Treatment effects

Although our main focus in this study is on parameter heterogeneity, in view of the different focus normally given to the endogenous switching model in the literature, we touch on this issue in our application. One of the main reasons for using an endogenous switching model is to estimate the effects of a discrete factor, or "treatment", of interest after correcting for non-random selection (often self-selection) into either the treated or non-treated condition. Simply put, selection bias occurs when the selection and behavior mechanisms are not independent. In our context, we may be interested in the average effect on VMD of *formerly non-urban residents who have already moved to urban areas*, referred to as the "average treatment effect on the treated (TT)". Alternatively, we may be interested in a second effect. A prominent policy discussion in urban planning has been: given the observation that people who have already chosen to live in urban areas tend to drive less than those *currently living in non-urban areas*, would the latter similarly reduce the amount of driving they do if they *were to move to urban areas* (or if their neighborhood becomes more urbanized)? The average change in VMD for such individuals is the "average treatment effect on the untreated (TUT)". To the extent that those who have already moved to urban areas have "opted in" due to attitudinal predispositions, whereas those who move to (or end up in) such areas in the future do so as a consequence of policy-driven incentives or supply constraints, we would expect TT to be larger in magnitude than TUT. Finally, we could be interested in a third effect: that of *randomly-selected* individuals *moving from non-urban to urban areas*. This is simply referred to as the "average treatment effect (ATE)" (Heckman et al., 2001; Mokhtarian and van Herick, 2016). Across the population, the average treatment effect is the weighted average of the treatment effects on the treated

and the treatment effects on the untreated, where the weights are the shares of each group of people in the population (van Herick and Mokhtarian, 2020b; Wooldridge, 2015, Eq. 15).

In sum, if people are not *randomly* assigned to treatment and non-treatment groups (i.e. if they are *self-selected*), and if unobserved variables associated with self-selection are also associated with the outcome, then an estimate of the average treatment effect that is based on a simple, static comparison across groups is biased. It is important to correct for the fact that those who "opt in" to the treatment can differ in meaningful ways from those who do not.

To express the issue more formally, seemingly the simplest approach to estimating the average treatment effect of neighborhood type on a person of characteristics $\boldsymbol{X}_i$ would be to compute the difference between the expected outcomes indicated by the *deterministic segmentation* model:

$$\int E(Y_{i1}|\boldsymbol{X}_{i1}, z_i = 1)dF(\boldsymbol{X}_1) - \int E(Y_{i0}|\boldsymbol{X}_{i0}, z_i = 0)dF(\boldsymbol{X}_0)$$

$$= \int \boldsymbol{X}_{i1}\boldsymbol{\beta}_1 \, dF(\boldsymbol{X}_1) - \int \boldsymbol{X}_{i0}\boldsymbol{\beta}_0 \, dF(\boldsymbol{X}_0) \approx \frac{1}{n_1}\sum_{z=1} \boldsymbol{X}_{i1} \, \widehat{\boldsymbol{\beta}}_1 - \frac{1}{n_0}\sum_{z=0} \boldsymbol{X}_{i0} \, \widehat{\boldsymbol{\beta}}_0 \quad (3.29)$$

where n_1 and n_0 respectively equal the number of cases for which $z = 1$ and $z = 0$, and the right-hand side indicates the sample-estimated versions of the quantities on the left-hand side.

The main problems with this approach are twofold. First, the two terms comprising this measure refer to different people, because no one can belong to both groups

simultaneously. Second, even if we compare expected values averaged over *all* individuals for both terms, the *coefficients* estimated for each equation are representative of the (samples of) people who self-select into their respective conditions, but not necessarily representative of the population as a whole. In other words, we would not expect to have the same (statistically equivalent) estimates of $\boldsymbol{\beta}_1$ if *everyone* were subject to the treatment as we do when only a non-random (self-selected) segment of the population is subject to it, and similarly for $\boldsymbol{\beta}_0$.

Hence, to tackle these issues, we employ the endogenous switching model, which *corrects the estimates of* $\boldsymbol{\beta}_1$ *and* $\boldsymbol{\beta}_0$ *for the selection bias* associated with the deterministically segmented model. The average treatment effect (ATE) informally described above is defined as the expected gain (change) from treatment for a *randomly chosen* individual as opposed to a *self-selected* one (Heckman et al., 2001). The ATE for *Y*, conditional on characteristics $\boldsymbol{X}_i$, is:

$$ATE_Y(\boldsymbol{X}_i) = E(Y_{i1} - Y_{i0}|\boldsymbol{X}_i) = \boldsymbol{X}_i(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0)\ . \tag{3.30}$$

Note that the estimates of $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_0$ obtained from this equation (i.e. from the endogenous switching model) will differ from those appearing in Eq. (3.29). Then, the unconditional estimate of the ATE can be obtained by integrating the conditional (on characteristics) effect over the distribution of $\boldsymbol{X}$:

$$\begin{aligned} ATE_Y &= E(Y_{i1} - Y_{i0}) = \int ATE_Y(\boldsymbol{X}) dF(\boldsymbol{X}) \\ &= \int \boldsymbol{X}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0)\, dF(\boldsymbol{X}) \approx \tfrac{1}{n}\textstyle\sum_{i=1}^{n} \boldsymbol{X}_i\big(\widehat{\boldsymbol{\beta}}_1 - \widehat{\boldsymbol{\beta}}_0\big)\ , \end{aligned} \tag{3.31}$$

where the final expression indicates that the sample average treatment effect can be taken as a consistent estimate of the true average, provided of course that the pooled sample is representative of the population (or that the cases are weighted to achieve representativeness).

As in a similar application of Cao (2009), we employed a log-transformation of VMD, and thus to see ATE in its original scale, ATE$_V$, we need to back-transform it. Since, following the standard situation for linear regression, we assume that the transformed VMD is normally-distributed, the conditional (on characteristics) mean of VMD in its original scale obeys the lognormal distribution. For individual i, we have

$$\begin{aligned} &E[V_{1i}+1|\boldsymbol{X}_i]-E[V_{0i}+1|\boldsymbol{X}_i] \\ &\quad=E[\exp(\boldsymbol{X}_i\boldsymbol{\beta}_1+\varepsilon_{1i})|\boldsymbol{X}_i]-E[\exp(\boldsymbol{X}_i\boldsymbol{\beta}_0+\varepsilon_{0i})|\boldsymbol{X}_i] \\ &\quad=\exp(\boldsymbol{X}_i\boldsymbol{\beta}_1+\sigma_1^2/2)-\exp(\boldsymbol{X}_i\boldsymbol{\beta}_0+\sigma_0^2/2)\ , \end{aligned} \tag{3.32}$$

where V_{1i} and V_{0i} refer to the VMD for person i if she lived in an urban and non-urban area, respectively[35]. We *estimate* the *population-wide* ATE$_V$ with

[35] Recall that the transformation is actually $Y_z=\ln(V_z+1)$ for z = 1, 0 (to avoid taking ln(0) and to map V_z = 0 to Y_z = 0), where $Y_z \sim N[\boldsymbol{X}\boldsymbol{\beta}_z,\sigma_z^2]$. This yields $V_z+1=f(Y_z)=\exp(Y_z)$, where $V_z+1\sim LN[\boldsymbol{X}\boldsymbol{\beta}_z,\sigma_z^2]$. From known properties of the lognormal distribution, the conditional expectation of V_z + 1 is $E[\exp(\boldsymbol{X}\boldsymbol{\beta}_z+\varepsilon_z)\mid\boldsymbol{X}]=\exp(\boldsymbol{X}\boldsymbol{\beta}_z+\sigma_z^2/2)$, and its conditional median is $\exp(\boldsymbol{X}\boldsymbol{\beta}_z)$. If we approximate $E[f(Y_z)\mid\boldsymbol{X}]$ with $f(E[(Y_z)\mid\boldsymbol{X}])=f([\boldsymbol{X}\boldsymbol{\beta}_z])$ (e.g. Cao, 2009), Eq. (3.33) would simplify to $[\exp(\boldsymbol{X}\boldsymbol{\beta}_1)-\exp(\boldsymbol{X}\boldsymbol{\beta}_0)]$. But the equivalence is exact only for linear functions, whereas here, $f(Y_z)$ is not linear. Since $f(Y_z)$ is a strictly convex function of Y_z, Jensen's inequality holds that $E[f(Y_z)\mid\boldsymbol{X}]>f(E[(Y_z)\mid\boldsymbol{X}])$, and thus the simpler approximation on the right-hand side of the inequality will underestimate the true expected value on the left-hand side. In fact, the approximation $f([\boldsymbol{X}\boldsymbol{\beta}_z])=\exp(\boldsymbol{X}\boldsymbol{\beta}_z)$ is an estimate of the *median* of the lognormal distribution, as noted above. Because $\exp(\boldsymbol{X}\boldsymbol{\beta}+\sigma^2/2)>\exp(\boldsymbol{X}\boldsymbol{\beta})$, $E[V+1]>median[V+1]$. That is, the mean of VMD is greater than its median, consistent with the long right tail of the VMD distribution.

$$\mathrm{ATE}_V \approx \frac{\sum_{i=1}^{N} wt_i[\exp(\boldsymbol{X}_i\widehat{\boldsymbol{\beta}}_1+\hat{\sigma}_1^2/2) - \exp(\boldsymbol{X}_i\widehat{\boldsymbol{\beta}}_0+\hat{\sigma}_0^2/2)]}{\sum_{i=1}^{N} wt_i}, \quad (3.33)$$

where wt_i is the case weight for person *i*, correcting for our oversampling of less-urban (non-Atlanta) cases (among other sampling and non-response biases[36]; see Kim et al., 2019b for further details).

In this study, the estimated ATE is -190 miles based on the endogenous switching model. In other words, when a *randomly-selected* person moves from a non-urban area to an urban area, the average reduction in weekly VMD would be 190 miles (with the average VMDs of random people living in urban and non-urban areas estimated at 71 and 261 miles, respectively). This estimate seems fairly substantial (more than 27 miles a day) and perhaps counterintuitive. For example, the deterministic segmentation model tells us that the difference in expected VMD between urban and non-urban residents is about -42 miles per week. If the stereotypical residential self-selection effects were valid (i.e. "urbanists", who are likely to prefer other travel modes over driving, self-select into urban areas), then we might expect negative correlations between selection into an urban area and VMD generation (the ρ_1 and ρ_0 of Eqs. (3.16) and (3.17)): unobserved traits increasing the propensity to live in urban areas would also tend to dampen VMD. In that case, we would expect the uncorrected treatment effect to overestimate the true treatment effect, i.e. we would expect the true treatment effect in this application to be smaller than 42 miles.

[36] Specifically, the original sample was weighted to represent the population of Georgia with respect to gender, age, race, education, work status, MPO type (of residential location), income, household size, and vehicle ownership. We did not recompute the weights after excluding some cases as described in Section 3.4.1.

Instead, however, we found $\hat{\rho}_1 = 0.66$ (and $\hat{\rho}_0 = 0.84$), leading to the true treatment effect being much larger than 42 miles. We explain these results from the perspective of conceptual plausibility here. We also discuss this issue from the mathematical perspective, but do so in the Appendix A to avoid a lengthier digression here; readers who are interested in the technical details are referred to the Appendix A.

Although it may be counterintuitive, other studies have reported similar situations. Van Herick (2018) applied mover-stayer models (in both two-step and full information estimations) to weekly drive-alone commute days and reported positive correlations in both urban and suburban regimes. Singh et al. (2018) also reported that unobserved attributes of living in a medium (high) density area contribute to an increase in household VMT after accounting for other exogenous covariates. They commented (p. 34) that "Although this may appear counter-intuitive at first glance, it is not necessarily so. The very unobserved attributes that contribute to seeking residential location in higher density neighborhoods may very well contribute to higher VMT production. After controlling for built environment attributes and household socio-economic and demographic characteristics, households that favor active lifestyles and seek a variety of activity opportunities (latent unobserved traits) [leading them to locate in urban environments with a vibrant street life] are likely to undertake more travel and hence produce more VMT than observationally equivalent households that have different (more sedentary) lifestyle preferences." Furthermore, in our application it is unclear what the unobserved variables actually represent. In the stereotypical explanation, unobserved variables are likely to include some attitudes such as environmental consciousness and a "car-lite" orientation, which would support a negative correlation between the error terms of the two equations. In our

application, however, the residential location choice model *observes* several attitudes that are typically unobserved, including pro-environmental, urbanite, travel-liking, and pro-car-owning attitudes[37]. The logarithmic transformation of VMD may also be a source of sensitivity: a small difference in log terms becomes much larger when exponentiated to the original units.

Another thing to note, particularly for transportation analysis, is that it is often ambiguous how to define the treatment (Mokhtarian and van Herick, 2016, especially footnote 10). Unlike in Heckman's classical example, the treatment of interest is often continuous rather than binary, and there are multifaceted dimensions to be considered in defining what constitutes a treatment. For example, in the application of this study, treatment was defined as neighborhood type (urban and non-urban area). However, aside from the fact that every place has its own geophysical or cultural context, the level of urbanization ranges across a spectrum, and thus defining neighborhood type by a single cutpoint on that spectrum is rather arbitrary.

In our application, we define "urban area" using census block-group-level population density, with the cutpoint being a population density of 4 people per acre in the residential block group. Although, as indicated in our footnote 10, this is already a higher density than the US Census uses to define "urbanized area", it is still a rather low threshold, with the result being that "urban" includes many low-density areas, and "non-urban" is not

[37] These attitudes, if unobserved, could be expected to contribute to ρ_1 being "more negative", and indeed, when the model was re-estimated without these attitudes (and keeping everything else the same), $\hat{\rho}_1$ became considerably smaller (more closely approaching negativity) at 0.43, and insignificant (t = 1.32). On the other hand, $\hat{\rho}_0$ remained roughly unchanged at 0.86. In general, of course, the values of ρ_1 and ρ_0 are very dependent on what is observed versus unobserved, i.e. on the specifications of the membership and outcome models.

far from “rural” (in our sample, the median population densities are 6.09 and 1.17 people per acre, or about 3,900 and 750 people per square mile, for urban and non-urban areas, respectively). From that perspective, it may not be surprising that, at least in our study (in line with the observation by Singh et al., 2018), unobserved characteristics increasing the propensity to live in less sparsely populated areas would also be associated with a propensity for *more* travel (propensities to take advantage of the greater number and diversity of activity opportunities found in more urban areas compared to very small towns and rural areas) rather than less. Conversely, unobserved traits *decreasing* the propensity to live in (somewhat) higher-density areas would tend to lead to *less* travel. To be sure, the built environment still exerts an important influence on travel, which is why the observed average VMD is higher for non-urban areas, given the longer distances required for essential travel in very low-density environments. The unobserved attitudinal predispositions, however, are (in our case) apparently acting counter to, rather than in concert with, the built environment. This has the following implications:

(1) When a person self-selects into an urban area, she tends to be predisposed to travel more, in the sense that she desires a more active lifestyle with more, more diverse, and possibly more dispersed activity opportunities, but at the same time the built environment allows her to enact that predisposition with (perhaps much) less travel than would be required to accomplish that desired lifestyle in a non-urban area. The net effect of these counteracting forces is that she travels less than before (Eqs. (A3) versus (A6) in the Appendix A).

(2) If a non-urban resident were counterfactually to be “dropped into” an “urban” area (despite having a lower propensity to live in such a place), the built environment would then support her predisposition to travel within a smaller activity space (because “everything needed” would be nearby), and her average VMD would be *lower* than that of a *self-selected* “urban” resident (Eqs. (A5) versus (A3) in the

Appendix A). Conversely, if an urban resident were to "be moved" to a non-urban area (despite having a lower propensity to do so), her inherent predisposition to travel more would be amplified by the longer distances between activities in her new lower-density built environment, and her average VMD would be *higher* than that of a *self-selected* non-urban resident (Eqs. (A6) versus (A4) in the Appendix A).

(3) These counteracting effects also help explain why the average VMD for urban residents is not much lower than that for non-urban residents: urban residents are inclined to travel more but the built environment makes that less "necessary", while the converse is true for non-urban residents (Eqs. A3 versus A4 in the Appendix A).

These relationships are exactly what we see in our sample, as shown in Table A3 and Figure A1 in the Appendix A.

In addition, the range of the spectrum itself could be subject to the empirical context. For example, in this study, we modeled statewide residential locations, requiring that we include a wider spectrum of residential types (from urban to rural). However, many studies focused on treatment effects where the "treatment" indicates being moved from a suburban to an urban neighborhood in a specific metropolitan area (e.g. Cao, 2009; van Herick and Mokhtarian, 2020); Pinjari et al. (2008) and Bhat and Eluru (2009) divided 1099 zones in the San Francisco Bay Area into neo-urbanist and conventional neighborhoods by applying factor/cluster analyses on several measures related to urbanicity. Due to the described incongruencies, treatment effects across studies are not necessarily comparable.

Because it is out of scope, this study does not expand the discussion about ATE further, but for future studies it would be worthwhile to investigate how ATE behaves depending on model specifications and definitions of "treatment".

Turning to the latent class model, we could mechanically apply Eq. (3.33) for ATE_V to two classes, which yields a (weighted) ATE of -107 miles, suggesting that when a random person switches from class 1 to class 0, the average VMD reduction is 107 miles. However, unlike the endogenous switching model, it is unclear what this result really implies, in that belonging to a certain latent class is not associated with a particular physical treatment – as we have seen, individuals in either class can live in either type of area. Furthermore, it should also not be assumed that the same correction terms that apply to the sample selection model automatically apply here as well. Therefore, estimating a treatment effect is more appropriate for the endogenous switching model.

3.5.2 Membership model: link function, specification, and type of probability

The membership model plays a crucial role in either the endogenous switching or latent class formulation, in that we interpret the model as explaining how likely a given individual is to belong to each class. In this section we address several issues associated with the membership model: its functional form, specification, and the type of probability to use in downstream computations (prior versus posterior).

There is little discussion in the literature about the optimal link function (and related "best" functional form of the membership model), and indeed, as stand-alone models, when the impacts of the unobserved variables associated with belonging to each class are independently distributed there is little empirical difference between the top contenders of

logit and probit. Here, we simply want to highlight that various fields seem to have different traditions. As is well known, it is most common to construct the membership model with a logit link function for latent class modeling, while the endogenous switching model typically employs a probit link function (for the purpose of formulating a bivariate normal density). In the early stages of latent class modeling (e.g. Kamakura and Russell, 1989), it seems that the logit link function was utilized mainly because of the simplicity of satisfying probability constraints when maximizing the log-likelihood function. The basic implicit constraints are $0 \leq \pi_z \leq 1$ and $\sum_{z=1}^{Z} \pi_z = 1$, where z is a class indicator and π_z is the membership probability (also known as the mixing coefficient or mixing proportion, cf. DeSarbo and Cron, 1988) of class z. These constraints are required because we assume that classes are collectively exhaustive and mutually exclusive given the sample. By using the logit link function, we do not need to directly estimate mixing coefficients (which requires imposing the constraints above); rather we estimate unconstrained constants, which the logit formula (e.g. Eq. 3.21), for the two-class case) ensures will meet the necessary constraints. In other words, the choice of link function for the membership model often stems more from a mechanical reason than from an interpretation purpose. Hence, for example, many applications of latent class modeling do not necessarily contain variables in the membership model (e.g. Chiou et al., 2013; Anderson and Hernandez, 2016).

As mixture modeling gained popularity, many studies parameterized the membership model as a function of some information (e.g. demographics), allowing a more meaningful characterization of the classes (this enhancement dates back at least to Swait (1994) and Bhat (1997)). This approach is at least partly due to the influence of market segmentation concepts from the field of marketing research (please refer to Wedel and

Kamakura (2012) for background on the application of latent class modeling in marketing research; additional discussion about membership variables and models can be found in Section 2.3.4. Parameterizing the membership model, therefore, means that there are two (possibly overlapping) sets of explanatory variables: the $\boldsymbol{X}$ variables of the outcome model, and the $\boldsymbol{W}$ variables of the membership model. The first three specifications in Figure 3-3 respectively depict the roles of $\boldsymbol{X}$ and $\boldsymbol{W}$ for (a) typical single-class models, (b) latent class models, and (c) saturated models.

However, in this case an important dilemma arises with respect to model specifications, one that applies to endogenous switching models as much as to latent class models: which variables belong to $\boldsymbol{X}$, and which to $\boldsymbol{W}$ (Figure 3-3)? Put another way, does a particular variable *directly* affect the outcome (in which case it belongs to $\boldsymbol{X}$), or does it affect the *weight (coefficient) another variable has on the outcome* (in which case it belongs to $\boldsymbol{W}$)? This decision is usually based on theory or knowledge. For example, Swait (1994) provided a conceptual framework for latent class modeling, in which the membership likelihood is a function of general perceptions/attitudes and socio-demographics. However, problems could be twofold. First, for latent class modeling, it may not be clear how to characterize the latent classes with respect to the target outcome. Second, in behavioral modeling, situations will very likely arise where it is conceptually valid to model membership and outcome equations with the same variables (for example, income is likely to affect choice of residence as well as VMD). In theory, membership and outcome equations can have *entirely* the same set of variables (as shown by specification (c) in Figure 3-3); however, in practice we rarely find papers using such specifications (we

are unaware of any). We speculate that this is partly because doing so is apt to create estimation issues and/or to make interpretation difficult.

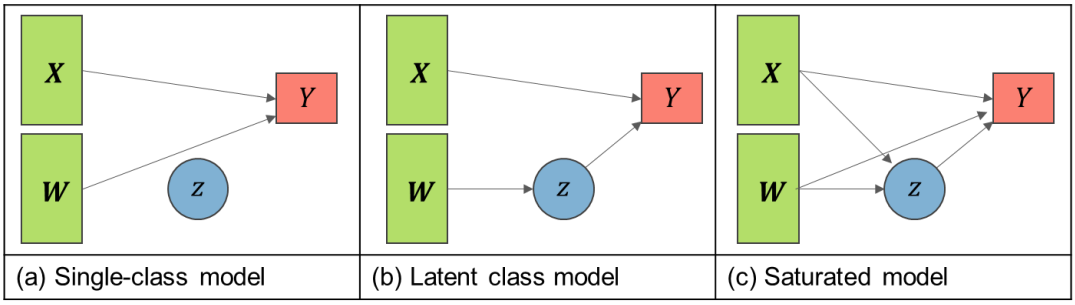


Figure 3-3. Prototypical model specifications

Turning now to the type of probability to use in downstream computations, we note that the ways membership probabilities can be calculated and how they behave could be of interest. Figure 3-4 shows the membership probabilities associated with our endogenous switching and latent class models. The first panel presents the membership probabilities (selection into urban or non-urban area) for endogenous switching. Since the membership model is estimated with respect to the selection of residential type, it is not necessarily correlated with the outcome variable (log-transformed VMD). The second and third panels exhibit the two types of membership probabilities of the latent class model: prior and posterior probabilities. The prior membership probability is as defined in Eqs. (3.21) and (3.22) in Section 3.3.4, $P(z_i = 1|\boldsymbol{W}_i)$ or $P(z_i = 0|\boldsymbol{W}_i)$. The posterior membership probability ($\check{P}$) considers the information provided by the outcome, and updates the probability using Bayes' Rule:

$$\check{P}(z_i = 1|y_i, \boldsymbol{X}_i, \boldsymbol{W}_i) = \frac{P(z_i = 1|\boldsymbol{W}_i) \times f(y_i|\boldsymbol{X}_i, z_i = 1)}{f(y_i|\boldsymbol{X}_i, \boldsymbol{W}_i)}$$

$$\check{P}(z_i = 0|y_i, \boldsymbol{X}_i, \boldsymbol{W}_i) = \frac{P(z_i = 0|\boldsymbol{W}_i) \times f(y_i|\boldsymbol{X}_i, z_i = 0)}{f(y_i|\boldsymbol{X}_i, \boldsymbol{W}_i)} \quad (3.34)$$

where we make the conventional assumption that y_i is independent of $\boldsymbol{W}_i$ given z_i, and thus that $f(y_i|\boldsymbol{X}_i, \boldsymbol{W}_i, z_i = 1)$ simplifies to $f(y_i|\boldsymbol{X}_i, z_i = 1)$ (and similarly for $z_i = 0$).

There are two observations from Figure 3-4. First, the membership probabilities of the latent class model are more closely associated with VMD than those of the endogenous switching model. In other words, membership in class 1 (the lower-VMD segment) is negatively associated with VMD, while for class 0 it is positively associated. Again, this is because the latent class model finds the solution that optimally fits the distribution of the outcome variable, whereas, for the endogenous switching model, segmenting on neighborhood type does not necessarily align with how much they travel. Second, after taking into account the outcome, posterior probabilities show a much stronger association with VMD. Of course, it is not particularly surprising that class memberships are more distinctly sorted in this case, since the individual's VMD outcome is accounted for in estimating (posterior) class memberships. As mentioned in Section 3.4.4, Kim and Mokhtarian (2018) argued that using prior probabilities is more appropriate in most discrete choice transportation applications where one-time prediction is the ultimate goal, since in such cases the outcome is not known in advance.

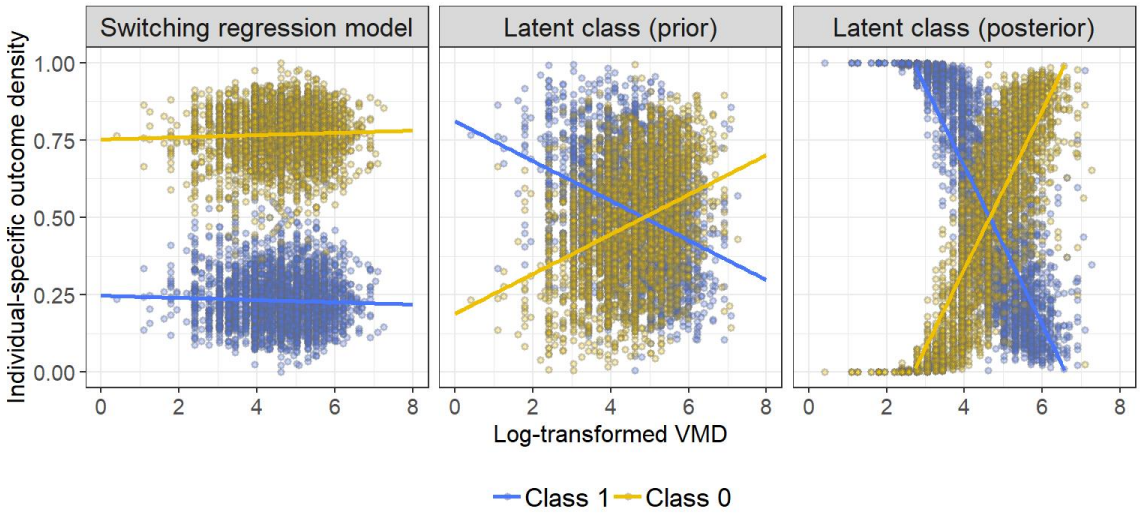


Figure 3-4. Membership probabilities for the endogenous switching and latent class models

3.5.3 Mixture modeling in machine learning: Mixture density networks

So far, we have examined various members of the family of statistical models that fall under the mixture modeling framework. It is worth touching on another family member that has been proposed in the machine learning field. Bishop (1994) proposed *mixture density networks* (MDN), which combine mixture modeling and neural network approaches. Some applications include speech analysis (e.g. Richmond, 2007; Zen and Senior, 2014) and touchscreen interaction locations in space and time (e.g. Martin and Torresen, 2018).

Figure 3-5 shows a conceptual illustration of MDN. In usual neural network applications, the neural network (regression) aims to minimize the squared loss function that approximates the *conditional mean* of the outcome (which is analogous to linear

regression). In contrast, MDN aims to map input variables to the *parameters of a Gaussian mixture model* (means μ_z, variances σ_z, and mixing weights π_z). Hence, MDN provides a probability density function of an outcome conditional on the input variables. Densities of related models are:

Linear regression: $f(Y_i|\boldsymbol{X}_i) \sim N(\boldsymbol{X}_i\boldsymbol{\beta}, \sigma^2)$,

Standard latent class model: $f(y_i|\boldsymbol{X}_i, \boldsymbol{W}_i) \sim \sum_{z=1}^{Z} \pi_z(\boldsymbol{W}_i) \times N(\boldsymbol{X}_i\boldsymbol{\beta}_z, \sigma_z^2)$,

Mixture density networks: $f(y_i|\boldsymbol{X}_i, \boldsymbol{W}_i) \sim \sum_{z=1}^{Z} \pi_z(\boldsymbol{X}_i, \boldsymbol{W}_i) \times N(\mu_z(\boldsymbol{X}_i, \boldsymbol{W}_i), \sigma_z^2(\boldsymbol{X}_i, \boldsymbol{W}_i))$,

where $\pi_z(\cdot)$ is a membership function and all other notation is defined in Section 3.3.

Figure 3-6 presents an application of MDN to our data. Since this study does not focus on MDN per se, we simplify the example. To model log-transformed VMD, we employ the four attitude measures which were used in our previous models (namely pro-environmental, urbanite, travel-liking, and pro-car-owning) to estimate a mixture of two Gaussian densities. In addition, we assume a single hidden layer with two nodes, as shown in Figure 3-5. The three histogram panels in Figure 3-6 present how the estimated μ, σ, and π are distributed for each mixture class, while the plots below the histograms represent individual density functions based on specific draws from each of the histograms.

When comparing MDN and the usual latent class model, MDN has several potential advantages:

1. It automatically captures nonlinearity in segmentation (Bishop, 2006) or in effects on the conditional mean.

2. We assume a homogeneous error variance within segment for the latent class model, whereas error variances are estimated as a function of input variables in MDN.
3. We do not need to determine how to allocate variables between membership and outcome models.

Of course, all advantages are possible at the expense of reducing interpretability. In addition, MDN is subject to multiple decisions of the analyst due to the embedded natures of the mixture modeling and neural networks. Specifically, the analyst must choose not only the number of mixtures (analogous to the choice of number of latent classes), but also the structure of hidden layers (i.e. how many layers and nodes and how each node is connected; cf. Bishop, 2006; Hastie et al., 2009). Put another way, the standard latent class model requires us to decide which variables belong to $\boldsymbol{X}$ and $\boldsymbol{W}$ (as described in Section 3.5.2), but neural networks structure requires us to decide the network structure per se instead of decisions on $\boldsymbol{X}$ and $\boldsymbol{W}$. The network structure will influence how input variables interact with each other (potentially transforming the input space) in the "black box". A key difference between the two decisions is that the decision in standard latent class models heavily relies on the analyst's knowledge/theory (but is not entirely free from the data since the final decision is also based on model fitness), whereas the decision in neural networks is purely dependent on the data – i.e. how well the structure fits the data. For example, there are no conceptual considerations to influence the choice of a certain number of nodes and layers, no policy implications for having a 3-node-1-layer structure instead of a 2-node-2-layer structure.

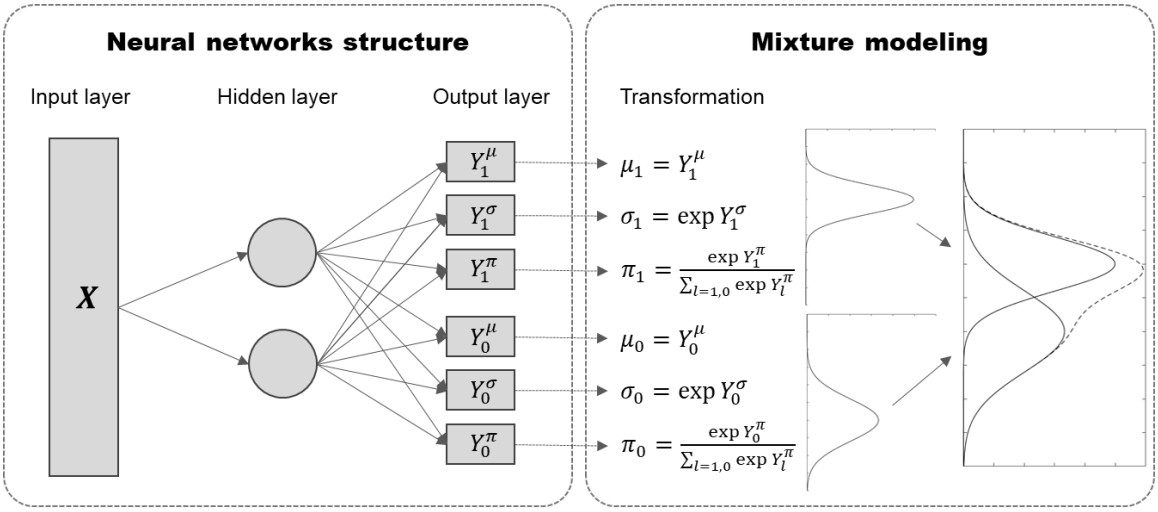


Figure 3-5. Conceptual diagram of mixture density networks

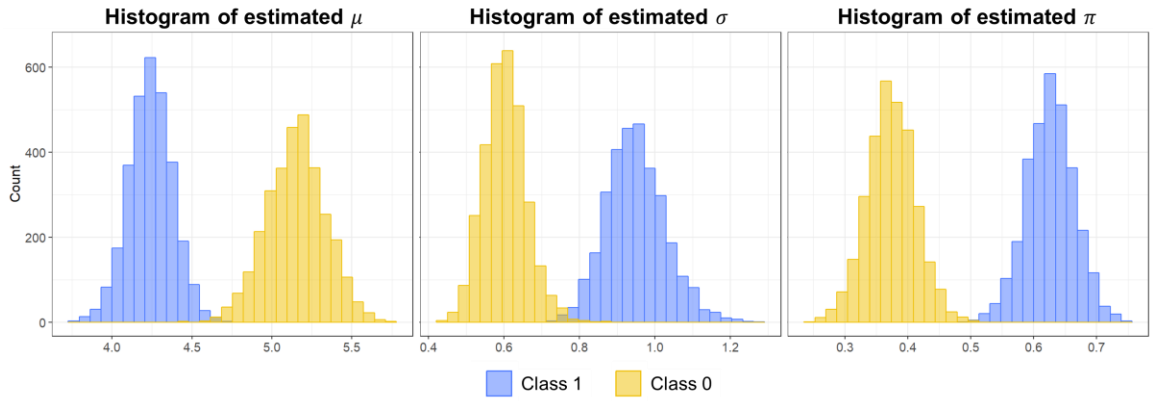


Figure 3-6. Application of mixture density networks

3.6 Conclusions

This study examined various modeling approaches that are closely related to each other. Focusing on a regression problem (i.e. with a continuous-valued outcome), the study explored the theoretical backgrounds of, and connections between, pooled, deterministic

segmentation, (endogenous) switching, and latent class models. In particular, the study highlighted similarities and differences among the models from the standpoint of finite mixture or market segmentation approaches as a way of dealing with (parameter) heterogeneity.

Models were applied to empirical data obtained from more than 3,000 Georgia residents. In particular, the study focused on weekly vehicle-miles driven (VMD), which is a key travel behavior indicator, and identified key explanatory variables as well as the different sensitivities to those variables exhibited by various population segments. Consistent with prior research, we found that, on average, not only do key demographic traits (e.g. gender, race, income) affect personal VMD, but so also do residential land use characteristics (job density, transit service level, and local accessibility) and telecommuting. Among the models of interest, the latent class model outperformed the competition. This implies that (1) there is notable heterogeneity in the population (when compared to the pooled model), and (2) uncovering latent segments can have benefits compared to segmenting on a certain predetermined factor. However, all models were able to provide meaningful insight for understanding the behavior of interest.

The study posited that people would generate different VMD and have different sensitivities by residence type (urban versus non-urban), and supporting evidence was found in the deterministic segmentation and endogenous switching models. An endogenous switching model was able to find those who are more likely to live in urban areas (e.g. workers, pro-environmental, and less favorable to driving). In addition, the estimated error correlations in the endogenous switching model corroborated that there are unobserved factors common to the joint decisions of residential choice and VMD

generation. After correcting for this self-selection effect, the average effect on VMD of moving from a non-urban to an urban area (where "urban" is arbitrarily defined as a population density of 4 or more people per acre in the residential block group) is estimated as -190 miles per week (further discussion of this result appears in Section 3.5 and the Appendix A). The latent class model identified lower-level and higher-level VMD segments. The two segments had different sensitivities. For example, individuals in the lower-VMD group tend to drive less as they live in more job-dense or better transit-service areas, whereas those variables do not have significant effects on VMD in the other segment.

These three segmentation models can be considered alternative ways of investigating parameter heterogeneity based on a finite segmentation framework. To help decide among them, some key questions for analysts are: (1) Is there a single variable that is reasonably suspected of introducing heterogeneity (deterministic segmentation (DS) or endogenous switching (ES)), or is segment membership conceived as being probabilistically associated with a bundle of variables (latent class (LC))? (2) Is it suspected that joint decisions (here, membership in known segments and VMD generation) share unobserved factors (ES), or not (DS)? Some models have additional merits with respect to certain purposes. For example, the endogenous switching model has the benefit of offering an estimate of the average effect of switching from one segment to another, i.e. the treatment effect, after properly accounting for non-random selection into a given segment. Hence, researchers should consider using this model when investigating the treatment effect of a specific factor, e.g. the impact of neighborhood type on VMD, as was examined in this study. The latent class model has the benefit of identifying the latent segments *best*

suited to the data and to a specific outcome variable; hence it has the potential of offering a better performance (not to mention new behavioral insights).

The study suggests some directions for future research. In the section on Further Discussions, we gave an overview of some fundamentals of the membership model in mixture modeling, which characterizes the segments. Each conceptual approach to formulating the model would have different performance and implications; hence future studies can apply each approach to empirical data, comparing their performance and analyzing how each approach leads to interpretational differences. Second, we also briefly touched on the use of mixture modeling in machine learning (mixture density networks, MDN, in particular). So far, there is a lack of discussion about heterogeneity in travel behavior studies using machine learning approaches. Hence, investigating heterogeneity with the aid of a combination of mixture modeling and a machine learning approach (e.g. MDN) might be interesting. Lastly, although this study examined the case of a continuous dependent variable (i.e. a regression problem), its conceptual discussions can also be applied to classification/choice problems.

CHAPTER 4. USEFULNESS OF THE CONFIRMATORY LATENT CLASS APPROACH

Paper title: *Who (never) makes overnight leisure trips? Disentangling structurally zero trips from usual trip generation processes* (*Travel Behaviour and Society, 28, 78-91, 2021*)

4.1 Introduction

Long-distance travel is an important pillar of the travel industry in particular and the economy in general. According to statistics from the U.S. Department of Transportation (USDOT, 2006), Americans are estimated to take 2.6 billion long-distance trips per year and 7.2 million trips per day (based on the National Household Travel Survey, NHTS, 2001 data, which defined long-distance trips as those longer than 50 miles). About nine out of ten long-distance trips are by personal vehicle, followed by air (7%) and then other modes such as bus and train. More recent statistics show the crucial economic role of long-distance travel. For example, the U.S. Travel Association (2020) estimates total long-distance travel-related output at $2.6 trillion (about 12% of the nation's gross domestic product), and 15.8 million travel-related jobs (about 10% of employed individuals). Annual growth rates of spending, employment, tax revenues, and personal trips in the U.S. domestic travel industry are estimated at 4.4, 1.8, 5.2, and 1.7% respectively.

Despite the importance of long-distance travel, data on long-distance travel behavior in the U.S. is relatively less available. For example, the NHTS had an additional long-distance component in 1977-2001, the 2009 survey did not have it, and the recent

2017 survey asked a few questions only in certain add-on regions (NHTS, 2018).[38] A few other surveys have had an emphasis on long-distance travel behavior, such as the 1995 American Travel Survey (e.g. Hwang and Fesenmaier, 2003), the Longitudinal Survey of Overnight Travel (Harvey et al., 2015), and the Utah Travel Survey (UDOT, 2013).

Long-distance travel generates discussions about social disadvantage and sustainability that differ from those associated with local travel. First, long-distance travel is more discretionary than daily travel. Long-distance travel is for those who are able (in terms of monetary expenditure and time, as well as physical and mental capacity) and desirous or willing. Second, a substantial portion of transportation-related emissions can be attributed to long-distance travel. Given that train and intercity bus have marginal market shares overall in the U.S., private vehicles and airplanes serve most long-distance travel in this country. Beyond the carbon footprint of cars, total emissions from air travel are quite substantial as well (cf. Ottelin et al., 2014; Czepkiewicz et al., 2018), with unit CO_2 emissions (per passenger per km traveled) for air exceeding those of car (BBC, 2019; Gonçalves, 2019). As a consequence, environmentalists and scholars have been concerned about the (un)sustainability of air travel for some time (cf. Becken, 2002; Åkerman, 2005), with the "flight shaming" movement representing a recent manifestation of that concern (e.g. Baron, 2019; Gossling et al., 2019; Gossling et al., 2020; Piskorz, 2019). Further, the asymmetric participation in long-distance travel implies that the carbon footprint from long-distance travel is not evenly distributed with respect to demographics and geography.

[38] However, the "NextGen NHTS" currently in advanced planning will measure long-distance travel.

Granting the importance of long-distance travel to the sustainability mission, the purpose of this study is neither to lament nor celebrate long-distance travel or those who undertake it. Rather, we wish to investigate the factors triggering long-distance travel (separately by mode) and identify those who are generating long-distance travel (or not), to inform the transportation policymaking that can address the social disadvantage and sustainability issues described (as well as informing the travel industry itself). Whereas most studies of long-distance travel focus on the trips made and the people making them, our understanding of those who "never" travel long-distance is limited. Hence, this study contributes to the literature by distinguishing between those who "structurally" do not make such trips and those who do generate long-distance trips, even if not very often (and who may therefore "incidentally" make zero trips during the study period). Based on questioning about lifetime frequencies, Graham and Metz (2017) created a similar typology of frequent, infrequent, and non-flyers for long-distance air travel in the UK. In our study, we do not have the full information required to make such a classification deterministically. Hence, we introduce a methodology that enables us to *probabilistically* classify cases into the typology. This approach is of interest in the many situations for which only imperfect information is available. We are not aware of any other studies of long-distance travel that have differentiated between structural and incidental zero-trip-makers in this way.

4.2 Literature review

There is no clear boundary between long-distance travel and other, more "usual", travel; rather, it depends on how we define "long distance". Long-distance travel has multifaceted dimensions such as distance, frequency, tour duration, purpose, mode, and destination (Table 4-1). Daily trips also have such dimensions, but there is a larger

spectrum of possibilities (e.g. distance has a wider range, destinations are more numerous and diverse) for long-distance travel. In addition, numerous constraints are likely to be involved in decisions relating to the dimensions of long-distance travel. In the literature, there is not a universally accepted definition of long-distance travel. Some studies defined it based on distance measures (e.g. 50 miles, NHTS 2001; 40 miles, UDOT, 2013; 100 miles, Berliner et al., 2018; 100 km, Czepkiewicz et al., 2020). Multiple studies employed a definition of "overnight" travel, mainly to avoid arbitrary distance thresholds (e.g. LaMondia et al., 2015; Aultman-Hall et al., 2018; Dowds et al., 2020). Some studies explored how various definitions of long-distance travel produce different results (e.g. LaMondia et al., 2014; Aultman-Hall et al., 2018). It is likely that no single measure would serve all purposes; rather, the definition chosen should depend on the research focus. In this study, we follow the definition of *trips involving an overnight stay*.

Since the travel mode is a major interest in transportation, a sizeable number of studies have examined mode choice for long-distance travel based on choice experiments. Hess et al. (2018) analyzed mode choice among train, personal car, air, and bus for selected major cities in the U.S. In particular, they applied a hybrid choice model and found a meaningful influence of attitudes on choices. van der Waerden and van der Waerden (2018) modeled medium- /long-distance travel mode choices between train and car, particularly focusing on access mode attributes. They found that travel time and cost are the most influential, whereas effects of ancillary attributes of access modes are relatively marginal. Lannoo et al. (2018) and Van Acker et al. (2020) explored the extent to which Belgian business travelers are interested in intercity coach services (compared to other modes). Monchambert (2020) focused on willingness to carpool for long-distance trips in France.

The study reported that people have a stronger willingness to travel alone rather than carpool, and carpoolers exhibited a higher average value of time compared to train or bus riders. Bergantino and Madio (2020) studied potential modal shifts under the planned high-speed rail services in Italy. They found that potential shifts are more likely from air and conventional rail services than from bus, carpooling, and private car.

Another approach taken in the literature to understanding behavior related to long-distance travel is to explore trip generation. That is, what factors stimulate long-distance travel and how much? Given the type of information available, several statistical models have been applied to modeling the frequency of long-distance travel. Frandberg and Vilhelmson (2003) modeled number of international trips in the preceding year with multiple regression models. LaMondia et al. (2014) applied ordered probit models to a four-level frequency category for each purpose (work and leisure/personal) and mode (air, intercity rail, and intercity bus). Aguilera and Proulhac (2015) modeled frequency of long-distance business trips with Poisson regression. LaMondia et al. (2015) examined inter-trip time intervals using 628 respondents to a longitudinal survey of overnight travel and employing negative binomial (NB) regression. With the same data, Aultman-Hall et al. (2018) applied NB regression to model annual tour generation for various definitions of long-distance travel. Berliner et al. (2018) modeled number of trips (total, leisure, and business purposes) with NB regression models. Czepkiewicz et al. (2020) also utilized NB regression models for modeling numbers of domestic ground trips, international leisure trips, and non-work air trips.

Various factors have been found to be significant influences on long-distance travel behavior, including gender, age, income, household composition, and geographic

characteristics. Specifically, age (e.g. Aultman-Hall et al., 2018; Berliner et al., 2018), income (e.g. Aultman-Hall et al., 2018; Berliner et al., 2018; Czepkiewicz et al., 2020), and being male (e.g. Berliner et al., 2018; Czepkiewicz et al., 2020) are positively associated with long-distance trip frequency. Attitudes are considered important influences, but relatively few studies have considered them for modeling long-distance travel (Berliner et al., 2018 and Czepkiewicz et al., 2020 being exceptions). Interestingly, long-distance travel behavior is likely to be dependent on accessibility to major airports, but a fairly limited number of studies have accounted for it (e.g. Enzler, 2017; Aultman-Hall et al., 2018) – we speculate that this is because information on the airports that are relevant to a given individual's trip, and the distance/travel time to those airports, is not readily available. These findings in previous studies are the foundations of our key hypotheses, which will be described in Section 4.3.3.

Graham and Metz (2017) is an important paper that shares a similar aim with this study. They employed UK data collected in 2014-2015 and described profiles of three types of people related to air travel: *frequent flyers* are those who have flown in the 12 months preceding the survey, *infrequent flyers* are those who have not flown in the preceding 12 months, and *non-flyers* are those who have never flown. Although this typology overlaps with the one we adopt (which will be described in Section 4.4.4), there are several important differences between Graham and Metz (2017) and this study. First, in addition to the two studies having different geographical contexts (UK vs. US), the two studies also have different focuses on travel mode. Graham and Metz (2017) particularly focused on air travel (no distinction between domestic and international), whereas we explore both air and car travel separately (particularly for domestic trips). In other words, Graham and Metz

(2017) specialized in air travel, whereas this study focuses on general long-distance travel. We believe car travel is a worthwhile form of long-distance travel to be explored, because the US is relatively more car-dependent than European countries[39] and thus the personal vehicle is a major source of long-distance trips in the US[40]. We start the study with the hypothesis that the motivations and factors respectively associated with air and car travel could be somewhat different and thus that the profiles of the three types of travelers may be distinctive across modes.

Second, Graham and Metz (2017) focus on examining profiles of air-travel market segments by using descriptive statistics. On the other hand, this study aims to *model* as well as describe, and to address *both* the segmentation and frequency of long-distance travel. Descriptive statistics are useful, but they generally do not indicate the effects of some factors while controlling for other factors, and thus they offer limited information when classifying new persons into segments. As well, our study simultaneously models the frequency of long-distance trips, whereas the trip frequency itself was not the interest of Graham and Metz (2017).

Lastly and importantly, the survey used in Graham and Metz (2017) explicitly asked respondents whether they had *ever* flown or not, whereas our study only measured the frequency in the past 12 months. In our case, we cannot explicitly distinguish the two types of zeros (infrequent and never) since the survey did not include such a direct question. This

[39] There are clear differences in the distributions of household car ownership in 2017 (in order of zero, one, two, three or more vehicles): US (9, 33, 37, and 21%), Georgia (7, 33, 38, and 22%), and UK (21, 43, 29, and 7%). (Source: the U.S. Census and the office for National Statistics, UK)

[40] For example, according to the 2001 NHTS in the US, 90 percent of long-distance trips were by personal vehicle, with the caveats that this share depends on the definition of long-distance trips and the actual distance traveled, where air would likely dominate for the longer trips.

is not an uncommon situation in real-world data analysis, where a general-purpose survey (like the NHTS) does not drill down into any specific issue in great detail, and thus where there is imperfect information. Hence, it is useful to present and apply appropriate methods for treating such cases. This study aims to tackle such a situation where only imperfect information is available; specifically, with the aid of the *latent segmentation approach* (which will be delineated in Section 4.3), the study will separate out the two different types of zero trips (by identifying two different behavioral mechanisms) and profile people in each segment.

Table 4-1. Scopes of recent long-distance travel studies in the literature

Study	Definition	Mode	Destination	Purpose	Time period
Jou et al. (2013)	NA	High-speed rail	Domestic cities (in Taiwan)	Business	Previous year
LaMondia et al. (2014)	Overnight / mode / international trip	Air, rail, bus	Any	Business / leisure	Previous year
LaMondia et al. (2015)	Overnight travel	Any	Any	Any	Past 12 months
Aguilera and Proulhac (2015)	Over 80 km		Any	Business	Previous 3 months
Reichert et al. (2016)	Involving overnight stay	Any	Any	Any	Last 3 months
Davis et al. (2018)	Over 50 miles	Any	Any	Non-commute	8 weeks
Aultman-Hall et al. (2018)	Overnight stop at least 50 miles from home; multiple distance thresholds	Any	Regional/inter-regional/contin-ental/global	Work / personal	Per year / last month
Berliner et al. (2018)	A trip longer than 100 miles (one way)	Total / air	Any	Business / leisure	Past 12 months
Czepkiewicz et al. (2020)	100 km (one way)	Ground / air	Domestic (Iceland) / international	Leisure	Previous 12 months / last month
Dowds et al. (2020)	Overnight travel	Total / air	Domestic (US) / international	Business / leisure	Last month

4.3 Methodology

4.3.1 Confirmatory latent class modeling

The main methodological issue posed by this study is the fact that we are missing the information needed to distinguish two types of processes generating zero domestic long-distance trips in the past 12 months. As noted in Sections 4.1 and 4.2, the ideal scenario is to measure life-long experiences of long-distance travel and thus explicitly separate those who have "never flown" from those who "have flown, but not in the past 12 months" (cf. Graham and Metz, 2017). In the absence of such an explicit measure, this study aims to separate the two types of zeros by probabilistic modeling. One way of making an inference on completely missing data is to use mixture modeling to identify underlying latent classes. As opposed to standard latent class models, which have an "exploratory" nature (cf. Hoijtink, 2001; Laudy et al., 2005), we employ a *confirmatory* approach in that we design latent classes with specific but differing assumptions about the behavioral models associated with each class.

To elaborate: as will be indicated in Section 4.4.1, our data are characterized by having a disproportionate share of zero-trip counts and this hints that there could be possibly more than one type of underlying behavioral mechanism generating trip counts (whether zero or not). Specifically, we posit that there are two behavioral processes in this context: some people are governed by a typical trip generation process (including zeros), while others are governed by a deterministic process of systematically producing zeros (i.e. they do not make long-distance trips at all). This is a type of heterogeneity in behavioral mechanisms and functional forms (Section 2.3.1). If people make a non-zero number of trips in the past 12 months we know they belong to the first class, but if they make zero

trips, we do not know to which class they belong, and we characterize their class membership with a binary-alternative probabilistic model. I.e., we assume the existence of two latent classes of zero tripmakers, each with a substantively different behavioral outcome process. Since we are imposing, *a priori*, these differing assumptions about the outcome processes for each class, such a model is considered confirmatory rather than exploratory, where an exploratory model would typically assume that outcome processes are similar in their essential nature but differing in parameters across latent classes (cf. Hoijtink, 2001; Finch and Bronk, 2011; Hess, 2014).

4.3.2 Formulation

As noted in Section 4.2, generation of long-distance travel has been modeled mainly with Poisson or negative binomial (NB) regression models. In view of our large shares of zeros, we consider the zero-inflated (ZI) versions of these models (Lambert, 1992). Note that another model cousin has also been proposed to capture cases involving a disproportionate share of zeros: hurdle models (Mullahy, 1986). Figure 4-1 illustrates these potential models. They resemble each other in that they each consist of two parts: a certain type of segmentation and a model for the outcome of interest (trip count in this case). However, the two models differ with respect to how they operationalize behavior. In brief, the ZI model assumes two different outcome regimes, where one regime always produces a zero outcome and the other follows a typical trip generation process (including zeros). For zero-outcome cases, regime membership is *unknown*, and selection between the two regimes is governed by a probabilistic model. From this perspective, the ZI model can be viewed as a particular type of *latent class model* as described above, in that we probabilistically segment zero-trip cases into two regimes whose membership is unknown

to researchers: the "structural zero-trip" regime and the "trip-making" regime. The hurdle model, by contrast, focuses on explaining *non-zero* outcomes with a *truncated* count model, together with a probabilistic participation model governing whether the outcome is observed (i.e., non-zero) or not – a distinction that is *known*. We now elaborate on each model in turn.

The ZI model can be expressed as follows:

$$P[Y = 0] = P[z = 0] + P[z = 1] \times P[Y = 0|z = 1] \text{ , and} \quad (4.1)$$

$$P[Y = y_i > 0] = P[z = 1] \times P[Y = y_i|z = 1] \text{ ,} \quad (4.2)$$

where $P[\cdot]$ is a probability, Y is the number of long-distance trips made by an individual [$y_i = 0, 1, 2, \ldots$ for individual i] and z is a regime indicator (0 for structural zero; 1 for trip generation). Eq (4.1) reflects the probability of zero trips occurring via either of the two regimes, while Eq. (4.2) captures the probability of being in the trip generation regime and $y_i > 0$ trips occurring.

Here, we employ the NB model for the trip generation process (i.e., given z = 1); hence, the "ZINB" model overall. We also experimented with Poisson models, but NB regressions outperformed them in every case. This is because our data exhibit overdispersion (greater variability than the Poisson model would predict) and thus an additional parameter (δ) controls for such dispersion (Cameron and Trivedi, 2005). The NB probability density can be obtained by incorporating a random component u_i into the conditional mean of the Poisson distribution, such that

$$P[Y = y_i|\boldsymbol{X}_i, u_i] = \frac{e^{-\lambda_i u_i}(\lambda_i u_i)^{y_i}}{y_i!} \text{ and} \tag{4.3}$$

$$E[Y|\boldsymbol{X}_i, u_i] = \lambda_i u_i \,, \tag{4.4}$$

where E is the expectation operator, λ_i, the mean of the Poisson distribution, is parameterized as $e^{\boldsymbol{X}_i\boldsymbol{\beta}}$ for mathematical convenience and to ensure a non-negative mean, $\boldsymbol{X}_i$ is a vector of explanatory variables, and $\boldsymbol{\beta}$ is a vector of parameters. Assuming the gamma distribution with mean 1 and variance $1/\theta$ for u_i, and then integrating out Eq. (4.5) over that distribution, yields a relatively tractable solution for the conditional density of Y (for details, please to refer to Greene, 2012 and Cameron and Trivedi, 2005):

$$\begin{aligned} P[Y = y_i|\boldsymbol{X}_i] &= \int_{u_i=0}^{\infty} P[Y = y_i|\boldsymbol{X}_i, u_i] g(u_i) du_i \\ &= \frac{\Gamma(\theta + y_i)}{\Gamma(y_i+1)\Gamma(\theta)} r_i^{y_i}(1 - r_i)^{\theta} \,, \end{aligned} \tag{4.5}$$

where $g(u)$ is the gamma probability density function, $r_i = \frac{\lambda_i}{\lambda_i+\theta}$, θ is reparametrized by δ (=$1/\theta$), which is the dispersion parameter, $\Gamma(\cdot)$ is the gamma function, and the other notation is defined as above. The conditional mean of Y is still λ_i (as for the Poisson distribution), but its conditional variance is now $\lambda_i(1 + \delta\lambda_i)$ instead of λ_i. Therefore, when $\delta = Var(u_i) = 0$ (signifying that the random variable u_i is actually a constant, namely its mean of 1), the conditional variance of Y becomes λ_i, Eq. (4.5) becomes Eq. (4.3) with u_i = 1, and the NB collapses into Poisson.

Also, the regime membership model is expressed as follows:

$$P[z = 0] = \frac{\exp(\boldsymbol{W_i \alpha})}{1+\exp(\boldsymbol{W_i \alpha})}, \tag{4.6}$$

where $\boldsymbol{W_i}$ is a vector of explanatory variables and $\boldsymbol{\alpha}$ is a vector of parameters. The log-likelihood function to be maximized (with respect to the unknown parameters $\boldsymbol{\beta}$, δ, and $\boldsymbol{\alpha}$) is as follows: $LL = \sum_{Y=0} \ln\{P[z = 0] + P[z = 1] \times P[Y = 0|z = 1]\} + \sum_{Y>0} \ln\{P[z = 1] \times P[Y = y_i|z = 1]\}$.

The hurdle model, as mentioned, also consists of two models. One governs the probability of generating a zero versus non-zero outcome, and the second is a count model that is *truncated* at (i.e. does not include) zero:

$$P[Y = 0] = P[non - participation] = P[z = 0] \tag{4.7}$$

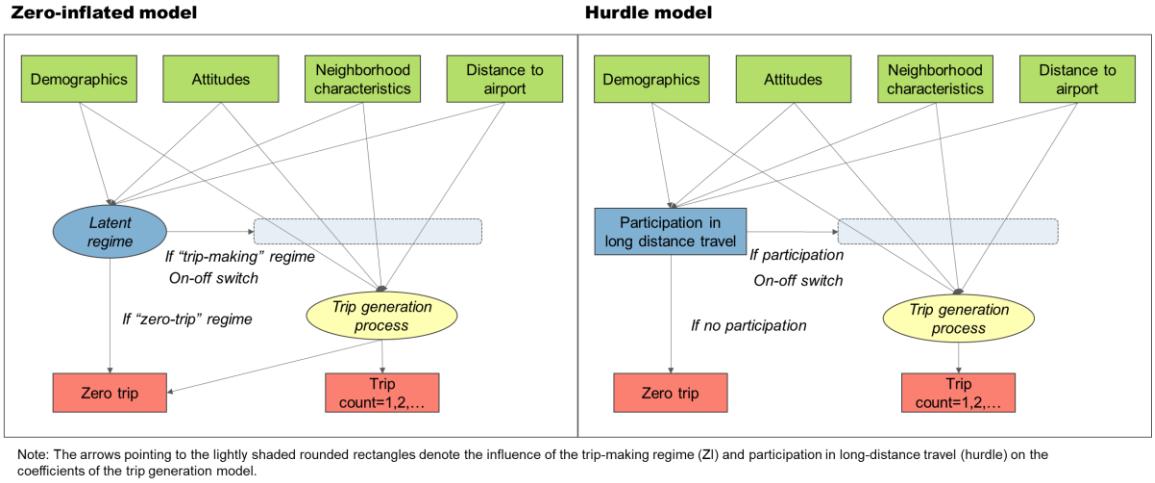
$$\begin{aligned} &P[participation\ and\ Y = y_i > 0] \\ &\quad = P[Y = y_i|participation] \times P[participation] \\ &\quad = \frac{P[Y=y_i]}{1-P[Y=0]} P[z = 1] \end{aligned} \tag{4.8}$$

where now $z = 0$ signifies non-participation (making 0 trips, for any reason) while $z = 1$ indicates participation (making a non-zero number of trips); $P[Y]$ denotes the untruncated count density (probability); $\frac{P[Y=j]}{1-P[Y=0]}$ is the truncated density of observing $j > 0$ trips given participation (i.e. the untruncated density rescaled so that the new probabilities will sum to 1 across $j > 0$); and $P[z = 1]$ is the selection or participation probability.

For reasons explained momentarily, we consider the ZI approach to be more suitable for this application. However, we also experimented with hurdle models and found that, in practice, the two approaches produced similar parameter estimates. The conceptual appeal of the ZI approach is that it allows us to differentiate between two types of zero-trip cases: (1) systematic or structural zeros, arising in our context because not everybody makes long-distance trips in the general population; and (2) random or incidental zeros, here arising because long-distance travel can be relatively infrequent and thus the respondent may simply not have made any such trips during the time period in question. Thus, we can separate the effects of factors on the *quantity* of trips generated (which could include zero trips generated for that period) from the effects of factors on the *participation* in trip-making altogether. For the hurdle model, by contrast, the "zero trips generated during this period" cases are confounded with the "non-participation altogether" cases. Accordingly, the basic idea of the ZI approach is potentially useful for many behavioral studies where there are heterogeneous reasons for a given value such as zero, and the data do not explicitly classify cases according to those reasons.

What kinds of people might fall into each regime? Conceptually, the non-trip-making regime captures those who do not make long-distance leisure trips because of possible *structural* constraints (e.g. affordability, mental/physical limitations, or attitudinal indifference or resistance with respect to long-distance travel, leisure travel, and/or adventure/exploration). By contrast, zero trips being made by members of the trip-making regime are less likely to be owing to such structural constraints. Rather, they are likely to be accounted for by temporary factors (e.g. some important life events, and/or changes in

income or free time) prohibiting (or failing to stimulate) long-distance travel in the past 12 months.



Note: The arrows pointing to the lightly shaded rounded rectangles denote the influence of the trip-making regime (ZI) and participation in long-distance travel (hurdle) on the coefficients of the trip generation model.

Figure 4-1. Illustration of two modeling approaches

4.3.3 Hypotheses

As aforementioned, in this study, we posit that there are two population segments related to long-distance travel. The *structural zeros regime* always produces zero long-distance trips because of structural/ permanent constraints or lack of motivation, whereas the *trip-making regime* can generate long-distance travel (where the number of trips is modeled with a count model). First, we generate a list of several hypotheses based on the literature (H1, H2, H3, and H5) and additional hypotheses (H4 and H6) based on informed speculation. The key hypotheses include:

H1. Income is a major factor that influences participation in long-distance travel and (if participating) the number of trips (e.g. LaMondia et al., 2014; Berliner et al., 2018);

H2. Presence of children reduces the likelihood of both participation in long-distance travel and (if participating) the number of trips (e.g. Berliner et al., 2018);

H3. Distance to a major airport reduces the likelihood of participation in air travel and (if participating) the number of air trips;

H4. Distance to a major airport increases the likelihood of participation in car travel and (if participating) the number of car trips;

H5. Urbanites make long-distance trips more often than others (potentially due to several reasons, such as a greater geographic dispersion of social networks, a "rebound effect" due to lower expenditures on car travel, or "compensation" for lower local access to green spaces; Holz-Rau et al., 2014; Czepkiewicz et al., 2018);

H6. Overall, factors more strongly affect air travel than car travel.

H1, H2, H3, and H5 aim to check whether our data return results related to the key drivers of long-distance travel that are consistent with those in the literature. For example, LaMondia et al. (2014) and Berliner et al. (2018) reported positive impacts of income on the frequency of long-distance travel, while Holz-Rau et al. (2014) and Czepkiewicz et al. (2018) discussed the role of urban form in explaining long-distance travel behavior. H3, H4, and H6 focus on the possibility of heterogeneous behaviors across travel modes. Specifically, the way people respond to airport accessibility may differ by mode. After incorporating our key hypotheses, we tested various model specifications by adding a number of other commonly-included variables, such as education, employment, and car ownership. Although H1-H6 are the key hypotheses of the study, this second step allows us to investigate additional potential influences on long-distance trip generation and latent class membership and to control for the effects of other factors. Our final models were chosen based on goodness of fit and conceptual validity; some tested variables were

dropped due to multicollinearity or insignificant impact in this sample. Note also that we postulate that attitudes cause behavior, as is conceptually plausible, conjectured in related studies (Berliner et al., 2018; Czepkiewicz et al., 2020), and in keeping with numerous enduring psychological behavioral theories such as the Theory of Planned Behavior, Technology Acceptance Model, and Extended Model of Goal-Directed Behavior. Nevertheless, the opposite direction of causality is also plausible, and this could be considered a limitation of the study.

4.4 Empirical application

4.4.1 Data

This study employs the GDOT data (Section 1.2.3). In this study, we define long-distance trips as those that involve an *overnight stay*. We collected information on respondents' self-reported number of long-distance trips over the past 12 months. For better understanding, we decomposed long-distance travel by purpose (business/work/school-related and leisure/recreational/social), mode (car, bus, plane, and other), and destination (within Georgia, states adjacent to Georgia, elsewhere in the U.S., Canada/Mexico/ Caribbean, and elsewhere in the world). We confine our interest in this study to certain types of long-distance travel. First, we limit the analysis to leisure/recreational/social (henceforth, "leisure") trips, for two reasons: work/business travel is closer to mandatory travel and thus the generation of long-distance business travel is *less* relevant (albeit not completely *ir*relevant) to the individual's willingness or constraints; and work trips are only taken by employed individuals while leisure trips can be taken by anyone. In terms of modes, we focus on car and plane because the shares of bus and other modes are marginal (in our sample, 0.8% and 2.2%, respectively); this is not only true of Georgia, but is also

generally characteristic of the U.S. context. In addition, we narrow the scope to domestic trips (77.8% of the air trips reported) to enable us to compare models of car and air travel.

We employ several sets of key variables, which have been found relevant in the literature as described in Section 4.2. In modeling, explanatory variables include demographics (both individual and household characteristics), attitudes, and geographical characteristics. The current study employs some attitudinal constructs that were identified through factor analysis on attitudinal statements. Geographical characteristics based on home locations are appended using external sources. Population density (per acre; Census block-group level) is calculated by using the 2017 American Community Survey (ACS) 5-year estimates. As a key variable, we measured the distance from home to airports via the Google Map API[41].

Table 4-2 exhibits descriptive statistics for key variables of the study. After excluding a few cases having missing values, the study analyzes 3,230 observations. Due to typical non-response biases, the sample is older and higher-income than the Georgia population as a whole (ACS estimates). We apply sample weights to help correct for sampling biases with respect to MPO size, income, household size, vehicle ownership, gender, education, race, age, and work status. The two dependent variables of interest are the numbers of long-distance domestic leisure air and car trips made over the past 12 months. On average, people made 0.76 and 4.25 such trips by air and car respectively. As

[41] In Georgia, there are nine commercial service airports. In particular, Atlanta Hartsfield International Airport (ATL) is one of the busiest airports in the world, serving as an international hub and thus having a distinct level/size of services. The other airports are rather small and have limited routes and frequencies (see Table B1 in Appendix B). We did not collect distances to airports outside Georgia. Although this is a limitation, given the few cases that would be closer to such airports than to ones in Georgia (with no assurance that such airports would actually be chosen by those few Georgia residents), we do not expect this to materially affect the results.

expected, in general, people traveled more often by car than by air. In addition, large standard deviations indicate that the number of long-distance trips is fairly dispersed in the sample.

It is worth noting that 74% (air) and 35% (car) of the sample reported zero for the number of long-distance trips within the period. As in many count data contexts, the numbers of zeros appear to be, so to speak, "inflated" (relative to what would be expected from a reasonable distribution, such as Poisson, describing the other counts; see Figure 4-2). The large shares of people not making a long-distance trip suggests that there could be heterogeneous reasons why they did not travel. In the next section, we will posit possible reasons and outline the methodology we used to address this issue. Figure 4-3 presents the geographical distribution of the sample and airports in Georgia.

Table 4-2. Descriptive statistics of key variables (N=3,230)

Variable	Category	Unweighted count	Unweighted Share	Weighted count	Weighted share	Share in population
Gender	Female	1574	48.7%	1682	52.0%	52.0%
Age (yrs)	18-34	285	8.8%	715	22.1%	31.5%
	35-64	1631	50.5%	1840	57.0%	52.1%
	65+	1314	40.7%	675	20.9%	16.4%
Annual household income	Lower income (below $50,000)	1025	31.7%	1365	42.2%	49.0%
	Medium income ($50,000 - $99,999)	1173	36.3%	1019	31.6%	29.7%
	Higher income ($100,000 or more)	1032	32.0%	846	26.2%	21.3%
MPO size	Atlanta MPO	1043	32.3%	1665	51.6%	52.1%
	Mid-sized MPOs	1171	36.3%	588	18.2%	13.8%
	Small-sized MPOs	814	25.2%	425	13.2%	11.0%
	Non-MPO areas	202	6.3%	552	17.1%	23.1%
Household composition	Presence of children age 14 and under	505	15.6%	698	21.6%	-
		Mean	Std. deviation	Mean	Std. deviation	
Long-distance travel (domestic, leisure)	Number of air trips over the past 12 months	0.71	1.84	0.76	2.01	-
	Number of car trips over the past 12 months	4.91	7.84	4.25	6.69	-
Geographical characteristics	Population density (per acre)	2.83	3.21	3.41	3.92	-
	Distance to Atlanta airport (miles)	104.74	78.60	82.50	73.20	-
	Distance to nearest major airport (miles) [a]	68.89	44.60	60.05	44.68	-
	Distance to nearest commercial service airport	26.24	17.51	29.16	17.65	-
Attitudinal constructs [b]	Tech-savvy	0.00	1.00	0.17	1.01	-
	Pro-car-owning	0.00	1.00	-0.02	1.09	-
	Travel-liking	0.00	1.00	0.00	1.02	-
	Polychronic	0.00	1.00	-0.01	1.05	-

a. In this study, Atlanta (ATL) and Savannah (SAV) are considered "major" airports. They are the only international airports in Georgia, and although SAV's passenger count is dwarfed by ATL's, SAV still has more than three times as many annual flights and four times as many passenger enplanements as the next largest airport in the state (Table B1, Appendix B).

b. Attitudes are estimated factor scores (standardized) obtained by applying factor analysis to attitudinal statements. Statements highly loading on each factor, with their loadings, are reported in Table 1-1.

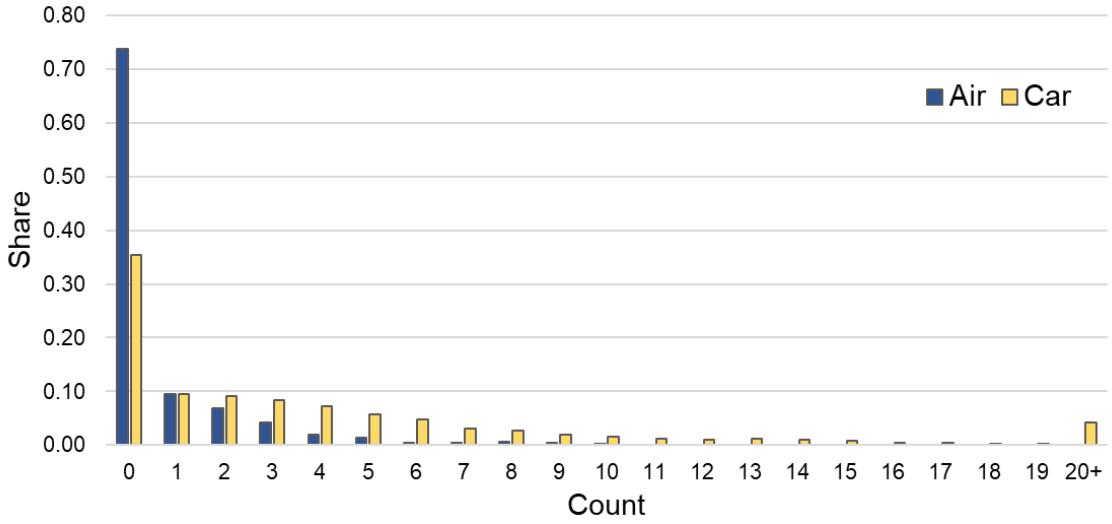


Figure 4-2. Distribution of the number of overnight domestic leisure air/car trips in the past 12 months (N=3,230)

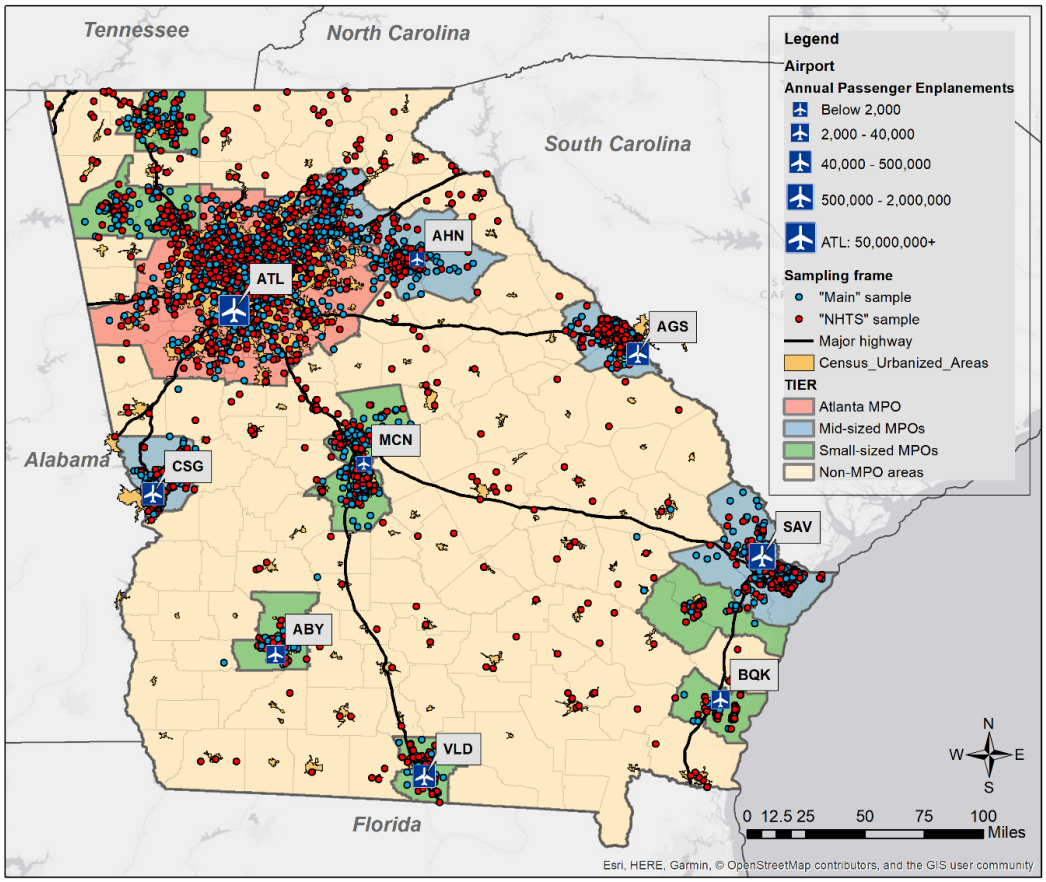


Figure 4-3. Geographic distribution of the sample and commercial service airports in Georgia

4.4.2 Estimation results

4.4.2.1 Models of membership in the zero-trip regime

Table 4-3 and Table 4-4 show the final ZINB models for air and car travel. First, we investigate the zero-inflation component – i.e. the selection model that explains membership in the structural zero-trip regime. For this model, a negative coefficient means that an increase in the associated variable decreases the likelihood of belonging to the structural zeros regime and thus increases the likelihood that domestic leisure long-distance trips will be generated. Not surprisingly, lower income people are more likely to belong to the structural zeros regime (i.e. no domestic leisure long-distance travel) for both modes. The same is true for women, a result for which we had no prior expectation. With respect to long-distance trip *count* models, there are mixed results in the literature: our finding is consistent with that of LaMondia et al. (2014) and Berliner et al. (2018), but the opposite directionality is also found (e.g. Aultman-Hall et al., 2018). With respect to total distance traveled for long-distance entertainment/recreation/ social purposes, Mokhtarian et al. (2001) also found the opposite directionality, with women traveling farther than men, *ceteris paribus*. Interestingly, the presence of children age 14 and under has opposite effects across modes. Having children increases the likelihood of never traveling by air, whereas it decreases the likelihood of never traveling by car. This, too, is not particularly surprising in view of the relative financial and logistical burden associated with flying with children compared to traveling by car with them, perhaps together with a lingering tradition of the "road trip" for family vacations or visiting relatives.

Population density is used as a proxy for the urbanicity of the residential location. It is negatively associated with selection into the structural zeros regime. That is, people living in more urban areas are more likely to generate long-distance travel for both air and car travel (but the magnitude and significance level are weak for car travel), in support of H5. This finding is consistent with the compensation hypothesis and similar arguments found in the literature, namely that people living in dense areas travel farther or more frequently for leisure, to compensate for a scarcity of open and/or green space in their living environment (Holz-Rau et al., 2014; Czepkiewicz et al., 2018). However, as mentioned, there are other possible explanations for why urbanites could make more long-distance trips, such as a rebound effect, access to transport infrastructure, socio-demographics, and greater dispersion of the social networks of urban residents (cf. Czepkiewicz et al., 2018). To test the compensation hypothesis more rigorously, we would need to control for some relevant factors (e.g. access to a private garden; cf. Strandell and Hall, 2015), which are not available in our data.

As distance to the Atlanta Hartsfield-Jackson airport increases, people are more likely to produce structural zeros for air travel, whereas it is the opposite for car travel. It is not surprising that the impedance of accessing the airport greatly affects whether the person can or will make long-distance trips by air. The opposite effect for car travel indicates that ground transportation, which is comparatively less burdensome when the airport is far away, is used to meet some needs for long-distance travel. Attitudes are also significant factors, confirming that structural zeros are not necessarily owing to monetary or physical barriers. Rather, attitudes or mental willingness are also involved in decisions on long-distance travel. Those who are tech-savvy, favorable toward travel, and

polychronic (enjoy multitasking) are more likely to generate long-distance travel. Polychronicity might be involved because the expectation of working or amusing oneself while traveling (particularly at the airport or on the plane) or at the destination could lower a mental barrier to long-distance travel. In addition, polychronicity is associated with a personality that thrives on multiple synchronous sources of stimulation, such as may be found on a long-distance leisure trip.

4.4.2.2 Count models for domestic leisure long-distance travel

Now, we turn to the second part of the model, which explains the amount of travel. First, we confirm that both dispersion parameters are significantly different from zero (otherwise, the NB model would collapse into the Poisson), indicating significant overdispersion in frequency. It is worth mentioning that when we used conventional (non-zero-inflated) NB regression models, both the air and car models presented much larger dispersion parameters (in our experiments, they were about 2.5). That is, when using ZI models, the magnitudes of the dispersion parameters are reduced. This implies that the ZI model is disentangling "ordinary" dispersion from the effects of heterogeneity in the trip generation process that would otherwise have been absorbed into the effects of the usual dispersion.

In both air and car travel, younger people tend to generate more long-distance trips. Income is an important factor that increases the amount of travel. However, the magnitudes are greater for air travel. This implies that income has a greater role in increasing the amount of travel for air than for car, which makes sense in view of the economies of scale associated with multiple people traveling together by car. The presence of children

decreases the number of trips in general for both modes. Hence, recalling the selection model for car travel, the presence of children has different roles between its influence on the selection into the zero-trip regime and its influence on the number of trips made. That is, the presence of children increases the likelihood of generating car-based long-distance trips in the first place, but it has a negative effect on the number of trips generated (given that the person belongs to the trip-making regime). Both results are logical, and illustrate again the value of separating the structural zero regime from the trip-making regime.

Population density is not a significant factor for either mode. Hence, living in a denser area increases the likelihood of generating trips at all (from the regime membership model), but it does not necessarily affect how many trips are generated. Distances to the closest major airport (either Atlanta or Savannah; see note on Table 4-2) significantly affect the amount of travel for both modes, but having opposite effects. Such distances reduce the number of air trips, but they increase the number of car trips (although the magnitudes are smaller). As described above, this may signify that people living farther away from the major airports tend to meet their needs for travel by car. Note that as described in Section 4.4.1, we collected distance measures to all commercial service airports in Georgia. In modeling, we tried various combinations of variables for distance measures in both the selection and count parts of the model. The current specification produces the best fit as well as more meaningful results. In particular, the shortest distance considering all commercial service airports in Georgia (Figure 4-3) is not generally significant. There are huge differences in airport sizes in Georgia (Table B1 in Appendix B), so this result implies that the amount of air travel (and, secondarily, the amount of car travel) is affected by accessibility primarily to the *major* airports, which provide more numerous and diverse

flight options. Lastly, although other key attitudes are not significant for the count models, travel-liking propensity generally increases the number of air and car trips. We tested whether a pro-environmental predisposition affects decisions on leisure long-distance travel [cf. cognitive dissonance between environmental attitudes and long-distance travel in Hares et al. (2010) and Davison et al. (2014)]. We could not find meaningful results in this application, indicating that there is no strong support in our data for such a presumption. This result is consistent with other findings in the literature – for example, Hares et al. (2010) found that none of the tourists in their four focus groups considered climate change when planning holiday trips (despite the fact that climate change was the explicit subject of the focus group and flying was the third-most-commonly cited influence of their lifestyles on climate change); Chen et al. (2011) surveyed Taiwanese travelers and found notable gaps between their general pro-environmental behaviors and pro-environmental air travel behaviors.

Table 4-3. Zero-inflated negative binomial model (air travel, N=3,230)

Membership component			
	Variable	Estimate	t-value
Intercept	Intercept	-0.572	-1.72
Gender	Female	**-0.440**	**-3.97**
Annual household income (ref: less than $50,000)	$50,000 - $99,999	**-0.684**	**-4.82**
	$100,000 or more	**-1.324**	**-8.17**
Household composition	Presence of children age 14 and under	**0.540**	**3.24**
Geographical characteristics	Log-transformed population density (per acre)	**-0.177**	**-4.09**
	Log-transformed distance to ATL airport (mi)	**0.330**	**5.14**
Attitudes	Tech-savvy	**-0.236**	**-3.97**
	Pro-car-owning	**0.383**	**5.98**
	Travel-liking	-0.111	-1.83
	Polychronic	**-0.245**	**-4.49**

Count component			
	Variable	Estimate	t-value
Intercept	Intercept	**0.790**	**3.79**
Age (ref: 35-64)	18-34	0.212	1.77
	65+	-0.113	-1.48
Annual household income (ref: less than $50,000)	$50,000 - $99,999	**0.463**	**4.48**
	$100,000 or more	**0.782**	**7.80**
Household composition	Presence of children age 14 and under	**-0.219**	**-1.97**
Geographical characteristics	Log-transformed population density (per acre)	0.019	0.59
	Log-transformed distance to the nearest major airport (mi)	**-0.275**	**-5.82**
Attitudes	Travel-liking	**0.130**	**3.60**
Dispersion parameter	Delta	**0.908**	**6.42**

Summary		
	Final log-likelihood (LL_F)	-3121.96
	Pseudo-R^2, $1 - \frac{LL_F}{LL_C}$ (where "C" means the constant-only Poisson model)	0.30
	Vuong statistic	5.99

Note: The bolded numbers indicate coefficients that are statistically significant at the 0.05 level.

Table 4-4. Zero-inflated negative binomial model (car travel, N=3,230)

Membership component			
	Variable	Estimate	t-value
Intercept	Intercept	0.149	0.47
Gender	Female	**-0.324**	**-2.96**
Annual household income (ref: less than $50,000)	$50,000 - $99,999	**-1.272**	**-9.80**
	$100,000 or more	**-1.657**	**-9.38**
Household composition	Presence of children age 14 and under	-0.389	-1.93
Geographical characteristics	Log-transformed population density (per acre)	-0.065	-1.69
	Log-transformed distance to ATL airport (mi)	**-0.168**	**-2.46**
Attitudes	Tech-savvy	**-0.366**	**-6.20**
	Pro-car-owning	-0.066	-1.20
	Travel-liking	**-0.312**	**-5.98**
	Polychronic	**-0.202**	**-3.55**

Count component			
	Variable	Estimate	t-value
Intercept	Intercept	**1.316**	**11.02**
Age (ref: 35-64)	18-34	**0.176**	**2.58**
	65+	**-0.094**	**-2.12**
Annual household income (ref: less than $50,000)	$50,000 - $99,999	0.063	1.28
	$100,000 or more	**0.277**	**5.37**
Household composition	Presence of children age 14 and under	**-0.156**	**-2.52**
Geographical characteristics	Log-transformed population density (per acre)	-0.016	-1.03
	Log-transformed distance to the nearest major airport (mi)	**0.077**	**2.80**
Attitudes	Travel-liking	**0.113**	**5.74**
Dispersion parameter	Delta	**0.855**	**25.31**

Summary		
	Final log-likelihood, LL_F	-8129.49
	Pseudo-R^2, $1 - \frac{LL_F}{LL_C}$ (where "C" means the constant-only Poisson model)	0.42
	Vuong statistic	8.27

Note: The bolded numbers indicate coefficients that are statistically significant at the 0.05 level.

4.4.3 Investigation of the zero-trip shares

As argued in Section 4.3, a major benefit of the zero-inflated methodological approach compared to the usual ones is that we can (probabilistically) identify two distinct behavioral groups and two different types of zeros. When testing typical NB model

specifications, we observed that their coefficients were inflated relative to those of the final models presented here. In the former models, effects that would otherwise appear in the regime membership models are absorbed into the count models, thereby biasing the parameter estimates of the typical models.

From our final models, two important estimates for zero trips are possible. First, we can estimate the fraction of those who structurally generate zeros, as 0.40 (air) and 0.10 (car) respectively. Put the other way, it implies that "all" domestic leisure long-distance trips are generated by about 60% and 90% of the sample, respectively. An additional takeaway is that the share of the zero-trip regime for air travel is substantially higher than that of car travel. This is not surprising in that there are more constraints for air travel (probably mainly monetary, but also psychological, e.g. a fear of flying). Hence, in fact, only half the population is expected to be able/inclined to generate long-distance travel. Some may meet their needs for long-distance travel by car, others' needs may remain unmet, and some portion of people may not have such needs at all.

A second estimate we can make from our models is the probability of generating zero trips for those who make trips. Under the modeling assumptions, those in the trip-making regime can generate long-distance travel, but some may not make any long-distance trips in the given period. It is estimated that 43.1% (air) and 13.3% (car), respectively, of those in the trip-making regime did not make long-distance trips. This might be due to temporal constraints (as opposed to systematic or permanent constraints) or simply a low average demand (i.e. the inter-trip time is greater than a year). Air travel has a higher fraction of these incidental zeros than car travel does. This is also reasonable in that air travel is relatively less frequent and thus zeros can be encountered more easily.

Combining these two estimates, we can calculate that the shares of structural zeros, incidental zeros, and non-zero trip-making in the sample are respectively 41%, 33%, and 26% for air, and 10%, 25%, and 65% for car. Referring back to Figure 4-2, these numbers mean that structural zeros constitute about two-thirds (69%) of the zero bar for air, and 66% of the zero bar for car. It is of interest to compare our results with those obtained by Graham and Metz (2017) for the UK. Their sample contained 8% non-flyers, 43% infrequent flyers, and 49% frequent flyers. Given that the typology of segments is similar, it is quite surprising that the two studies report such different shares of segments (correspondingly 41%, 33%, and 26% in our study). Seemingly, overall, the fraction of zero-trip individuals in our study is greater than that of the UK study. However, this simple comparison requires some caveats. First, the numbers in Graham and Metz (2017) include any trip purposes and any destinations (i.e. both domestic and international), whereas our study focused on leisure-purpose domestic trips. Hence, we expect a larger share of zero trips in our study. Second, due to geographical and economic factors, driving is more feasible/attractive for those who make long-distance trips in the contiguous US than those in the UK (which is an island about $1/40^{th}$ of the land area of the US, and having higher gas prices in general). Third, the UK has more airports (about 40 airports, whereas there are 9 airports in Georgia) and greater urbanization; hence, UK residents may have greater accessibility to airports than Georgians. These second and third factors make it much more likely that Georgians will drive instead of fly for many of their long-distance leisure trips, relative to UK residents. Lastly, the distribution of demographics might be at play. Graham and Metz (2017) reported that their sample quotas were set to be representative of the UK with respect to gender and age; however, it is not clear how geography, income, and vehicle

ownership (which would affect air travel patterns) are distributed in the sample. On the other hand, our study employed sample weights, but the sample is still biased to some extent. All these discrepancies doubtless contributed to the differences in the results (which are in the expected direction).

Regarding one important factor, Figure 4-4 exhibits the estimated share of cases in the structural zeros regime as distance to ATL airport is hypothetically varied for everyone in the sample, from "no impedance" to several hundred miles away (leaving other factors at their sample values). As everyone moves from the place of "no impedance" to 100 miles away from the ATL airport, the share falling into the structural zeros regime changes from 0.2 to 0.52 for air travel. This means, on the one hand, that 20% of people would not make long-distance leisure trips by air even if they have *no* accessibility impedance to the airport – i.e., about a fifth of the population may have hard-core constraints or disinclinations to traveling by air for leisure. It means, on the other hand, that if everyone lived 100 miles away from the ATL airport, slightly under half (about 48%) of them would still generate domestic leisure air travel. By contrast, the change for the same scenario is from 0.3 to 0.18 for car travel. The share of structural zeros decreases because greater distance from a major airport motivates some to choose car for their long-distance leisure travel instead of flight. However, about 15% of people would not generate car-based long-distance leisure trips regardless of their accessibility to the airport (i.e., no matter how far away an airport is) – again indicating the presence of hard-core constraints or disinclinations, this time associated with the car mode.

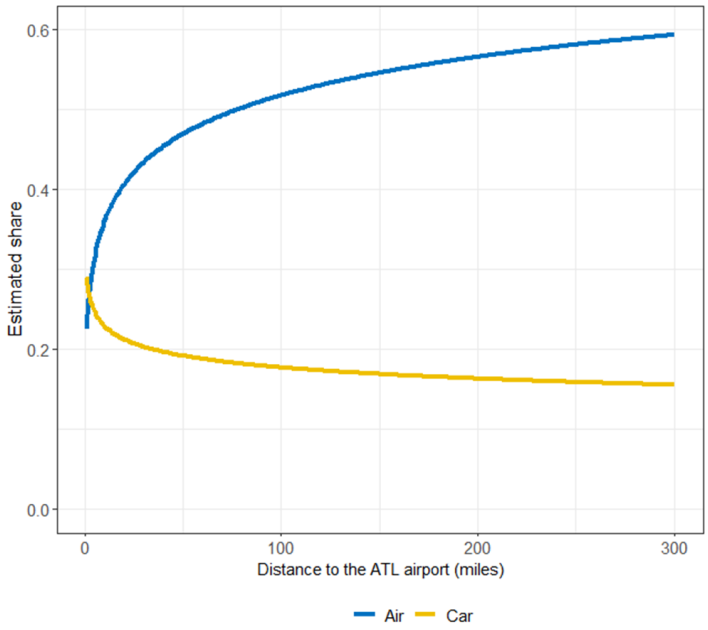


Figure 4-4. Estimated share of those in the structural zeros regime by hypothetical distance to the ATL airport

4.4.4 Profiles of each group

In this study, we built two ZINB models for air and car travel. A natural follow-up question might be, "who never makes long-distance trips?" and "who did not make long-distance trips all year, even if they do so less often?" Table 4-3 and Table 4-4 already hinted at answers to this question. To provide more concrete answers, Table 4-5 presents the profiles of specifically three types of people: structural zero trip-makers, incidental zero trip-makers, and non-zero trip-makers. Profiles for the two types of zeros are calculated based on membership probability-weighted (and sample-weighted) characteristics among zero-trip individuals, whereas those of non-zero trip-makers are determined by their non-zero-trip observations.

On average, the non-zero trip-makers are the youngest and have the highest income, whereas the structural zero trip-makers are the oldest and have the lowest household income for both air and car travel. There is a difference in the share of cases having children age 14 and under with respect to car travel – the share of those in the structural zero regime is particularly lower. Non-zero air trip-makers live in the densest areas on average, whereas structural zero regime members for air travel live in the lowest density areas. As expected, there are notable differences in average distance to ATL or nearest major airport among the three air travel groups. However, the differences seem marginal among the three car travel groups. As expected, average tech-savvy, travel-liking, and polychronic propensities score in descending order for non-zeros, incidental zeros, and structural zeros for both air and car trips. In terms of the pro-car-owning propensity, the air and car groups exhibit opposite directions of progression. In sum, we can clearly see the role of instrumental factors such as income, age, and presence of children, as well as attitudinal factors such as travel-liking and pro-car-owning, in distinguishing the structural zero cases from the trip-making cases in general and the incidental zero cases in particular.

We looked at two additional attitudinal constructs that are seemingly relevant but not included in the final models: the pro-exercise and family-oriented predispositions also differ by group. For both air and car modes, the non-zeros group has the highest average proclivity for exercise, followed by the incidental zeros and then structural zeros. It seems that the enjoyment of physical activity is one motivation for long-distance leisure travel (e.g. for camping, hiking, and other outdoor physical activities). Similarly, the average family-oriented attitude scores follow the same ordering, for both air and car trips, suggesting that visiting family and/or traveling with family are additional motivations for

long-distance leisure travel. The differences in average propensities are greater for car travel than for air travel. In particular, the structural zero regime members for car travel have the lowest average family-oriented predisposition.

Table 4-5. Characteristics of the three groups

	Mode	***Air***			***Car***		
	Group	Structural zeros	Incidental zeros	Non-zeros	Structural zeros	Incidental zeros	Non-zeros
	Share	0.41	0.33	0.26	0.10	0.25	0.65
Share							
Gender	Female	0.49	0.54	0.54	0.49	0.51	0.53
Age	18-34	0.17	0.22	0.31	0.09	0.19	0.25
	35-44	0.16	0.16	0.20	0.15	0.16	0.18
	45-64	0.40	0.41	0.37	0.40	0.41	0.39
	65+	0.27	0.21	0.11	0.36	0.24	0.17
Income	Below $50,000	0.60	0.39	0.19	0.84	0.53	0.32
	$50,000 - $99,999	0.28	0.33	0.34	0.13	0.31	0.35
	$100,000 or more	0.12	0.27	0.47	0.04	0.16	0.34
Household composition	Presence of children age 14 and under	0.24	0.21	0.19	0.12	0.20	0.24
Mean							
Geographical characteristics	Population density (per acre)	2.46	3.60	4.67	2.98	3.62	3.40
	Distance to ATL airport (mi)	104.42	77.71	54.37	87.28	82.28	81.86
	Distance to nearest major airport (mi)	73.83	58.07	41.07	67.30	61.68	58.30
Attitudes	Tech-savvy	-0.13	0.25	0.53	-0.51	0.08	0.31
	Pro-car-owning	0.20	-0.12	-0.26	-0.21	0.04	-0.02
	Travel-liking	-0.16	0.04	0.20	-0.54	-0.08	0.12
	Polychronic	-0.28	0.10	0.26	-0.57	-0.09	0.10
	Pro-exercise	-0.33	-0.11	0.32	-0.51	-0.19	0.02
	Family-oriented	-0.16	-0.06	-0.04	-0.36	-0.28	0.02

4.5 Conclusions

4.5.1 Summary and relevance of findings

This study examined long-distance travel behavior of residents in the state of Georgia. Based on a survey conducted in 2017-2018, we modeled the number of domestic leisure long-distance trips (involving an overnight stay) over the past 12 months by air and

car modes. We observed that zero trips comprised more than half the responses, and posited that there could be two possible types of zeros: structural and incidental. The former is generated by those who either cannot make long-distance trips because of more permanent reasons such as monetary/physical/mental constraints or who simply lack the motivation to do so; the latter occurs among those who do make such trips but did not do so over the past 12 months for incidental reasons.

Based on that supposition, we used zero-inflated negative binomial (ZINB) models, which are a special type of latent class model, to simultaneously explain the selection of people into a *structural zero-trip regime* or a *trip-making regime*, and the number of trips (possibly including zero) made by the latter group. To our knowledge, this is the first study of long-distance travel to decompose the zeros using this methodological approach to treating an issue of imperfect information (namely, ignorance of the regime to which a zero-trip person belongs). The models produced meaningful insights. Selected demographics, attitudes, and geographical characteristics played important roles in explaining the segmentation of people into the two regimes, and the amount of long-distance leisure travel.

The study presented separate models by mode and they showed different sensitivities to the explanatory variables. In particular, the presence of children and distance to a major airport had different roles in the models. For example, the presence of children acted as a barrier to membership in the trip-making regime for air travel, but it was a facilitator for car travel. On the other hand, it was negatively associated with the number of trips in both modes. Not surprisingly, accessibility to airports does matter. As distance increased, both the probability of membership in the trip-making regime and the

count of trips were diminished for air travel, but car travel exhibited the opposite effects. In addition, it is not about accessibility merely to any nearest airport (a variable that was tested and found insignificant). Rather, accessibility to major airports, which provide greater flight frequencies and more destination options, appears to be more important. A profile analysis of the two zero groups and the non-zero group showed clear differences across all three groups with respect to both instrumental factors (e.g. income, age, and presence of children) and attitudinal factors (e.g. travel-liking and pro-car-owning), in expected but still informative ways.

The two types of zeros identified in this study could be of interest to several types of actors, including planners/policymakers, environmental groups, and the tourism industry, as well as researchers specializing in each of those areas. The structural zeros call for deeper investigation into their sources, specifically the extent to which people never make domestic leisure long-distance trips due to constraints (which points to an association with social disadvantage) versus lack of interest/motivation (which is a matter of personal preference). In particular, those with constraint-based structural zeros for *air* travel would tend to have a limited range for their long-distance leisure travel. For the incidental zeros, on the other hand, it is desirable to understand the factors contributing to making more or fewer trips. But while the questions may be similar, the uses to which the answers would be put may differ by entity. Environmentalists may wish to discourage long-distance travel (especially by air); the tourism industry wishes to encourage it (both by increasing the trip frequency of trip-makers and by nudging some people out of the zero-trip regime); and policymakers may wish to balance well-being, social disadvantage, environmental, and economic considerations.

4.5.2 Limitations and directions for future research

The decomposition of zeros into permanent or structural zeros versus occasional or incidental zeros is suitable for many contexts in behavioral studies. Specifically, analysts are often required to explain both whether people participate in a particular activity at all, and how much they do so, with different processes governing each of those decisions and with an inability to observe whether a zero is permanent or incidental. For example, we can expect that some kinds of people will "never" use (micro-) shared mobility (e.g. Uber/Lyft, e-scooter, e-bike), e-shopping, teleworking, etc., while others may only "incidentally" not use it during the study period. As with other studies taking a similar approach, such as Alemi et al. (2019), it will be worthwhile to apply the method to those empirical contexts. As noted in Section 4.3, zero-inflated and hurdle models are both candidates for addressing "excessive" zero issues, but they work on different assumptions about the behavioral mechanisms. Hence, researchers should exercise care in choosing the method and interpreting the results, to adduce proper implications from the empirical context.

The present study has several limitations, which point to fruitful directions for future research. It focused on the number of *any* domestic leisure overnight trips by each mode. However, decisions on the number of trips and travel mode are expected to heavily depend on a particular origin-destination pair (for example, decisions for Atlanta-Orlando and Atlanta-Los Angeles could be different). Due to the limited information in the survey, the study could not account for such specificity, and thus is not capable of addressing how people make decisions on travel mode given the particular choice situation. Further studies may delve into those issues.

Along the same lines, multiple purposes are often involved in long-distance travel (e.g. a leisure purpose can be added to a business trip). The survey did not capture such mixed purposes, and indeed the present study focused on trips whose primary purpose was leisure. In this regard, it would be desirable for future research to analyze mixed-purpose long-distance trips or interactions among multiple purposes in the long-distance travel context. Another limitation of our survey is that it did not measure trip durations. There could well be a tradeoff between frequency and duration, with some people – perhaps especially families with children (as a reviewer pointed out) – making fewer but longer trips. It would be useful to measure both frequency/counts and duration, and model them jointly.

With respect to modeling, we estimated the share of those who systematically generate zero trips. Although we were able to profile the members of the structural zero-trip regime to some extent, our general-purpose survey did not measure life-long experiences of long-distance travel, nor explicit reasons for not making long-distance trips (such as fear of flying, or geographic extent of one's social network). Ideally, those key variables should be measured by designing a special-purpose survey. For example, Graham and Metz (2017) presented the shares of the self-reported reasons for being non-flyers (e.g. budget constraints, flying is not an option, preferring other modes, etc.). Tackling the issue of the missing information, the present study relied on modeling, resulting in the probabilistic classification of people, instead of being able to exploit an explicit indicator that could deterministically classify them. There is no formal way to demonstrate that our (probabilistic) classifications are valid, without an indicator of the ground truth. At best, we can check the goodness of fit of the model and confirm that it explains the data

significantly better than a simpler model. Thus, this caveat remains in the method (in fact, this issue is embedded in any type of latent class modeling).

Two remarkable "interventions" are prominent in transportation research nowadays, and they have implications for long-distance travel. Autonomous vehicles (AVs) enable "hands-free" travel, thus rendering greater convenience and comfort for long-distance trips and particularly allowing the passenger to conduct other activities while traveling (cf. LaMondia et al., 2016; Kim et al., 2019a; Perrine et al., 2020). Hence AVs can provide the benefits of both air travel (e.g. comfort, no need to drive, activities while traveling) and car travel (e.g. privacy, less expensive for groups, availability of one's own car at the destination). We speculate that AVs will ultimately reduce the share of people belonging to the structural zero-trip regime. For example, some of those who do not make long-distance leisure trips by car because of their high value of time and resistance to long-distance driving may benefit from hands-free car travel. In addition, some of those who cannot travel by air because of the price of airfare, living far from the airport, and having mental/physical barriers to flying will take advantage of AVs. Another intervention is the COVID-19 pandemic, which is significantly affecting long-distance travel-related industries. We expect that COVID-19 creates both types of zero trips for long-distance travel. A first-order impact is to generate temporal or incidental zeros. This is mainly because of travel advisories/ prohibitions/ flight-and-event cancellations as well as fear of contact with others. Making zero trips due to these reasons may not last as conditions eventually return to a new normal, and hence, those zero trips are expected to gradually change to trips. In the near-term recovery period, some people are likely to travel at even higher frequencies than before, due to the release of suppressed demand and as postponed

events are rescheduled. This study is unable to explore the effects of these two interventions, but they are fertile areas for future research.

Lastly, the study context is specifically the U.S. state of Georgia. We can conjecture that long-distance travel behavioral processes will vary across states and countries as we already hinted in Section 4.4.3 in comparing our results with those of Graham and Metz (2017). Aside from the fact that demographics and economies differ across region, we particularly expect that the geographical locations of major airports, the size of the airports, and the destination options in adjacent regions would matter. Recall that Georgia has the ATL airport, which has the largest passenger flows in the U.S. (and among the highest in the world), and thus other airports may have smaller catchment areas (meaning that the effect of accessibility may be smaller). In addition, European or Asian contexts, which are geographically more condensed and have substantial rail passenger flows, may produce different stories for long-distance travel, such as different shares of the structural zero-trip regime, roles of modes, and size of airport catchment areas.

CHAPTER 5. LATENT CLASS MODELS WITH AN ERROR STRUCTURE

Paper title: *Latent class models with an error structure: Investigating potential dependence between latent segmentation and behavior generation*

5.1 Introduction

In the year 2000, James L. Heckman and Daniel L. McFadden won the Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel. Their main contributions were "for his [Heckman's] development of theory and methods for analyzing **selective samples**" (emphasis added) and "for his [McFadden's] development of theory and methods for analyzing **discrete choice**" (emphasis added)[42]. Indeed, their theories and methods are now ubiquitous and have provided major horsepower to research in various fields such as economics, other social sciences, and transportation. Their seminal works are distinct, but interrelated. Heckman recollected the background behind his development of methods in his Nobel Prize lecture (Heckman, 2001, p .686): "The two sets of tools available to me were classical Cowles Commission simultaneous equations theory and models of discrete choice originating in mathematical psychology that were introduced into economics by Quandt (1956, 1970), McFadden (1974, 1981), and Domencich and McFadden (1975). My goal was to unite these two literatures in order to produce an economically motivated, low dimensional, simultaneous equations model with both

[42] The Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel 2000. NobelPrize.org. Nobel Media AB 2020. Tue. 9 Jun 2020, https://www.nobelprize.org/prizes/economic-sciences/2000/summary/.

discrete and continuous endogenous variables that accounted for systematically missing wages for nonworkers and different dimensions of labor supply within a common framework, that could explain female labor supply, and that could be the basis for a rigorous analysis of policies never previously implemented."

Not only is discrete choice analysis embedded in sample selection models, but also the sample selection approach is closely related to endogenous sampling and relevant statistical treatments in discrete choice analysis, when the standard assumption of random sampling is violated (McFadden, 2001). Additionally, one of McFadden's important studies (McFadden and Train, 2000) demonstrated the power of mixed logit models by showing that any discrete choice model derived from random utility maximization can be approximated as closely as desired by a mixed logit model. An important comment in the study pertains to the *latent class model,* which is our interest in this chapter: it is a special case of mixed logit where the mixing distribution has finite (instead of continuous) supports. Our study exploits the fundamentals of these two economics giants' contributions as connecting some of the major ideas embedded in both the latent class model and sample selection. Although Heckman's sample selection model was originally devised for linear regression problems, it has been extended to many cases including choice problems.

As indicated, in this study we focus on latent class modeling. A basic idea of latent class models in behavioral modeling is to introduce a finite mixture into the discrete choice model (or any other outcome model) and to probabilistically segment the sample into (latent) subgroups that have different behavior functions. Latent class modeling (or finite mixture modeling) has been widely used in various applications such as residential location (Walker and Li, 2007), health care utilization (d'Uva and Jones, 2009), purchase channel

choice (Tang and Mokhtarian, 2009), rail ticket purchase timing (Hetrakul and Cirillo, 2013), financial satisfaction (Brown et al., 2014), obesity (Greene et al., 2014), college choice (Schmidt et al., 2019), choice of financial advisors (Amaral and Kolsarici, 2020), and so on. In particular, transportation has been an important application domain for latent class modeling, involving topics such as mode choice (Bhat, 1997; Vij et al., 2013), vehicle ownership (Anowar et al., 2014; Kim and Mokhtarian, 2018), preference for bus fare structure (Hess et al., 2013), and crash analysis (Eluru et al., 2012; Yasmin et al., 2014). A more comprehensive review is available in CHAPTER 2. Its popularity is mainly attributable not only to its performance, but also its conceptual attractiveness for treating various types of heterogeneity (Hess, 2014; CHAPTER 2).

One of the major differences of latent class modeling from other discrete segmentation models is, as we can infer from the name, that we do not know the true indicator for segmenting the population. Interestingly, there are a few implicit assumptions, whose tenability has been rarely discussed, associated with most latent class models. As noted in CHAPTER 3, a particular assumption of interest in this study is that the membership and outcome models are *independent* (i.e. those two behavioral processes are not associated and thus their error terms are not correlated). This is a crucial difference from sample selection modeling, which postulates that sample selectivity is endogenous to the outcome equations, aside from the other difference that typical sample selection models have a known group indicator (which is not the case for latent class models). What if the mechanisms governing how people belong to segments and how people generate behaviors share common unobservables? This could be likely in many contexts – for example, suppose it is revealed that there are health "nuts" and those who are indifferent to

maintaining healthy practices. If we model the number of walk/bike trips, such a model will likely share unobserved characteristics with membership in those segments. In this case, the propensity of belonging to the health-nuts class and the propensity to generate walk/bike trips could be positively associated.

In other words, we want to combine the main ideas of latent class modeling and sample selection modeling (or the endogenous switching model). To this end, we incorporate an error structure into the latent class modeling framework. There have been several types of latent class models incorporating an error structure. Errors could be correlated across decisions by an individual (Vij et al., 2013), across alternatives (Wen et al., 2012), between joint choices (Vij et al., 2013), and between an endogenous variable and the error term (Maness and Cirillo, 2016). However, another possible avenue, which we discuss in this chapter, is to have an error structure between the membership and outcome models. As noted by Greene et al. (2014), this approach resembles the switching regression model, but the key difference is that the individual's segment membership is *unobserved*. Greene et al. (2014) is the only attempt we are aware of that questions the implicit assumption of independence for latent class models and provides some supporting results. They proposed this approach, in particular with two latent classes, and applied it to obesity analysis in health economics. The present authors are unaware of discussions about the validity of the independence assumption elsewhere, including in the transportation domain, despite the numerous applications of latent class modeling. Hence, this study aims to introduce this approach to transportation analysts and latent-class model users, to propose models, and to demonstrate empirical applications with two different examples. In

addition, our study provides key equations (including for marginal effects) and discusses subtle but important methodological implications of the proposed approach.

The remainder of this chapter is organized as follows. Section 5.2 proposes the modeling framework. In Sections 5.3 and 5.4, we apply the proposed methodology to two empirical contexts. Section 5.5 summarizes the findings and discusses future directions.

5.2 Methodology

5.2.1 Model formulation

This study derives equations for two latent classes, and two types of outcome: binary or ordinal. Let us start with the membership model (or selection model). Suppose there are two segments in the population. If we formulate a binary probit model (i.e. assuming normality of the error term), the membership propensity and probabilities are given by Eqs. (5.1, 5.2, 5.3).

$$z_i^* = \boldsymbol{W}_i\boldsymbol{\alpha} + u_i \,, \tag{5.1}$$

$$P(z_i = 1) = \Phi(\boldsymbol{W}_i\boldsymbol{\alpha}) \,, \tag{5.2}$$

$$P(z_i = 0) = 1 - P(z_i = 1) = 1 - \Phi(\boldsymbol{W}_i\boldsymbol{\alpha}) = \Phi(-\boldsymbol{W}_i\boldsymbol{\alpha}) \,, \tag{5.3}$$

where i indexes the individual, z_i^* is a latent variable determining class membership, $\boldsymbol{W}$ is a vector of membership variables, $\boldsymbol{\alpha}$ is a vector of membership parameters, and u_i is an error term. Throughout the chapter, $\phi(\cdot)$ and $\Phi(\cdot)$ denote the standard normal density and cumulative probability functions, respectively. If $z_i^* > 0$ an individual belongs to class 1

($z_i = 1$), and otherwise belongs to class 0 ($z_i = 0$). The class-specific outcome utility functions are:

$$Y_{i1}^* = \boldsymbol{X}_i\boldsymbol{\beta}_1 + \varepsilon_{i1} \tag{5.4}$$

$$Y_{i0}^* = \boldsymbol{X}_i\boldsymbol{\beta}_0 + \varepsilon_{i0}\ , \tag{5.5}$$

where Y_z^* denotes the latent outcome propensity for class z (1 or 0), $\boldsymbol{X}$ is a vector of explanatory variables, $\boldsymbol{\beta}$ is a vector of parameters, and ε is an error term. The unconditional (marginal) outcome probabilities, Eq. (5.6), can be obtained by summing over z the joint probability of belonging to class z ($= \{1,0\}$) and obtaining an outcome y ($= \{1,0\}$ if binary and $= \{1, 2, \ldots, j, \ldots, J\}$ if ordinal). Each of the overall probabilities will be specified shortly.

$$P(y_i = j|\boldsymbol{X}_i, \boldsymbol{W}_i) = P(z_i = 1, y_i = j|\boldsymbol{X}_i, \boldsymbol{W}_i) + P(z_i = 0, y_i = j|\boldsymbol{X}_i, \boldsymbol{W}_i) \tag{5.6}$$

The heart of the proposed methodology is to introduce an error correlation structure over the equations for joint modeling. The error terms in the system are assumed to follow the trivariate normal distribution, as in Eq (5.7):

$$\begin{pmatrix} u_i \\ \varepsilon_{i1} \\ \varepsilon_{i0} \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho_1 & \rho_0 \\ \rho_1 & 1 & 0 \\ \rho_0 & 0 & 1 \end{pmatrix}\right]. \tag{5.7}$$

Here, the variances of u_i, ε_{i1}, and ε_{i0} are fixed at 1 for identification. ρ_1 and ρ_0 correlate the unobserved variables associated with the membership model with those of the

respective outcome models. Note that the covariance between ε_{i1} and ε_{i0} is fixed at zero since everyone can belong to only one of the two classes and thus the correlation is unidentifiable. As noted by Greene (2012), the choice of zero is merely for convenience and it does not play a role in the estimation of the model coefficients. This error structure, in fact, is equivalent to that of the *endogenous switching* model (cf. Cameron and Trivedi, 2005; Greene et al., 2014; CHAPTER 3), a variation on Heckman's original sample selection model (in which, originally, an outcome was observed for only one of the two segments; Heckman, 1979).

Allowing the error structure in the modeling system (specifically between the segmentation and outcome equations) calls for some important remarks. The standard latent class model is formulated with (finite) mixture modeling, which implies that Eqs. (5.4) and (5.5) are *conditional* (on class) distributions and ε_1 and ε_0 are *defined over the subpopulations* of class 1 and class 0 respectively. On the other hand, the proposed method takes the nature of a latent class model by treating the class membership indicator (z) as latent, but at the same time, it takes on endogenous switching model characteristics. In the spirit of the endogenous switching model, ε_1 and ε_0 are *defined over the population* because u, which is correlated with ε_1 and ε_0, is defined over the population (Maddala, 1986).[43] In this case, Eqs. (5.4) and (5.5) are specified at the population (rather than the subpopulation) level and we observe Y_1 only if $z^* > 0$ and Y_2 only if $z^* < 0$.

[43] Greene et al. (2014) provided the inspiration for this approach, but were not clear about this point. For example, they claimed (p.5), "This paper extends the finite mixture/latent class model literature by explicitly defining a latent variable for class membership as a function of both observables and unobservables, thereby allowing the equations defining the class membership and observed outcomes to be correlated". However, the phrase "finite mixture model" may introduce confusion. Indeed, this proposed method can be considered latent class modeling and perhaps, loosely speaking, finite mixture modeling, in that the model contains finite latent classes that have different behavioral processes (or functional forms).

The joint choice probabilities are as follows:

In the binary case,

$$
\begin{aligned}
P(z_i = 1, y_i = 1 | \boldsymbol{X}_i, \boldsymbol{W}_i) &= \Phi_2[\boldsymbol{W}_i\boldsymbol{\alpha}, \boldsymbol{X}_i\boldsymbol{\beta}_1, \rho_1] \\
P(z_i = 0, y_i = 1 | \boldsymbol{X}_i, \boldsymbol{W}_i) &= \Phi_2[-\boldsymbol{W}_i\boldsymbol{\alpha}, \boldsymbol{X}_i\boldsymbol{\beta}_0, -\rho_0] \\
P(z_i = 1, y_i = 0 | \boldsymbol{X}_i, \boldsymbol{W}_i) &= \Phi_2[\boldsymbol{W}_i\boldsymbol{\alpha}, -\boldsymbol{X}_i\boldsymbol{\beta}_1, -\rho_1] \\
P(z_i = 0, y_i = 0 | \boldsymbol{X}_i, \boldsymbol{W}_i) &= \Phi_2[-\boldsymbol{W}_i\boldsymbol{\alpha}, -\boldsymbol{X}_i\boldsymbol{\beta}_0, \rho_0]
\end{aligned}
\tag{5.8}
$$

In the ordinal case,

$$
\begin{aligned}
P(z_i = 1, y_i = 1 | \boldsymbol{X}_i, \boldsymbol{W}_i) &= \Phi_2\big[\boldsymbol{W}_i\boldsymbol{\alpha}, (\mu_{1,1} - \boldsymbol{X}_i\boldsymbol{\beta}_1), \rho_1\big] \\
P(z_i = 1, y_i = j | \boldsymbol{X}_i, \boldsymbol{W}_i) &= \Phi_2\big[\boldsymbol{W}_i\boldsymbol{\alpha}, \big(\mu_{1,j} - \boldsymbol{X}_i\boldsymbol{\beta}_1\big), \rho_1\big] - \\
&\qquad \Phi_2\big[\boldsymbol{W}_i\boldsymbol{\alpha}, (\mu_{1,(j-1)} - \boldsymbol{X}_i\boldsymbol{\beta}_1), \rho_1\big] \\
P(z_i = 1, y_i = J | \boldsymbol{X}_i, \boldsymbol{W}_i) &= \Phi_2\big[\boldsymbol{W}_i\boldsymbol{\alpha}, \big(-\mu_{1,(J-1)} + \boldsymbol{X}_i\boldsymbol{\beta}_1\big), -\rho_1\big]
\end{aligned}
\tag{5.9}
$$

$$
\begin{aligned}
P(z_i = 0, y_i = 1 | \boldsymbol{X}_i, \boldsymbol{W}_i) &= \Phi_2\big[-\boldsymbol{W}_i\boldsymbol{\alpha}, (\mu_{0,1} - \boldsymbol{X}_i\boldsymbol{\beta}_0), -\rho_0\big] \\
P(z_i = 0, y_i = j | \boldsymbol{X}_i, \boldsymbol{W}_i) &= \Phi_2\big[-\boldsymbol{W}_i\boldsymbol{\alpha}, \big(\mu_{0,j} - \boldsymbol{X}_i\boldsymbol{\beta}_0\big), -\rho_0\big] - \\
&\qquad \Phi_2\big[-\boldsymbol{W}_i\boldsymbol{\alpha}, (\mu_{0,(j-1)} - \boldsymbol{X}_i\boldsymbol{\beta}_0), -\rho_0\big] \\
P(z_i = 0, y_i = J | \boldsymbol{X}_i, \boldsymbol{W}_i) &= \Phi_2\big[-\boldsymbol{W}_i\boldsymbol{\alpha}, \big(-\mu_{0,(J-1)} + \boldsymbol{X}_i\boldsymbol{\beta}_0\big), \rho_0\big]
\end{aligned}
\tag{5.10}
$$

where $\Phi_2(\cdot)$ denotes the bivariate cumulative standard normal distribution, $\mu_{z,j}$ represents the jth threshold parameter of class z, and the other notation is as defined earlier. Given

However, statistically speaking, allowing the error correlation structure puts this model closer to the endogenous switching model family than to the mixture model family, due to the reasons described above.

the joint probabilities and marginal probabilities of membership, the class-specific conditional choice probabilities are given by Eqs. (5.11) and (5.12).

$$P(y_i = j|z_i = 1, \boldsymbol{X}_i, \boldsymbol{W}_i) = \frac{P(z_i=1, y_i=j)}{P(z_i=1)} \text{ , and} \tag{5.11}$$

$$P(y_i = j|z_i = 0, \boldsymbol{X}_i, \boldsymbol{W}_i) = \frac{P(z_i=0, y_i=j)}{P(z_i=0)} \text{ .} \tag{5.12}$$

The joint probabilities are equivalent to bivariate probit models. In conventional bivariate probit models, however, we know the choice indicators for both dimensions, whereas here one of the dimensions is *latent*. If ρ_1 and ρ_0 are zeros, the models are reduced to standard latent class models with a probit link function (instead of the usual logit link, cf. CHAPTER 2). Hence, in this case, the joint probabilities are just the products of the two associated marginal probabilities.

Returning to the more general (error correlation) case, the log-likelihood (LL) can be obtained by summing the logged marginal probabilities of the chosen outcomes over individuals (binary and ordinal respectively):

$$LL = \sum_{i=1}^{n} [y_i \ln P(y_i = 1|\boldsymbol{X}_i, \boldsymbol{W}_i) + (1 - y_i) \ln P(y_i = 0|\boldsymbol{X}_i, \boldsymbol{W}_i)] \text{ ,}^{44} \tag{5.13}$$

$$LL = \sum_{i=1}^{n} \sum_{j=1}^{J} I(y_i = j) \ln P(y_i = j|\boldsymbol{X}_i, \boldsymbol{W}_i) \text{ ,} \tag{5.14}$$

[44] In the empirical application (Study 2), the model will deviate from this standard form. We will derive the log-likelihood function under the assumption that only one of the regimes has an outcome equation. This is a type of confirmatory latent class model (CHAPTER 2 and CHAPTER 4), and is analogous to the zero-inflated model. The revised likelihood equation will be provided in Section 5.4.

where $I(y_i = j) = \begin{cases} 1, & if\ y_i = j \\ 0, & otherwise \end{cases}$.

For statistical inference, we calculate standard errors via bootstrapping that provides asymptotically consistent estimates. Bootstrapping approximates the distribution of a statistic by a Monte Carlo simulation. In other words, the basic idea is to re-estimate many times on different samples taken from the original sample (with replacement). This bootstrapping method also relies on asymptotic theory like other conventional methods (Cameron and Trivedi, 2005). The asymptotic covariance matrix based on bootstrapping is as follows:

$$V(\widehat{\boldsymbol{\theta}}) = \frac{1}{B-1}\sum_{b=1}^{B}\left[\widehat{\boldsymbol{\theta}}_b^* - \overline{\widehat{\boldsymbol{\theta}}}_B\right]\left[\widehat{\boldsymbol{\theta}}_b^* - \overline{\widehat{\boldsymbol{\theta}}}_B\right]' \tag{5.15}$$

where b indexes bootstraps ($b = 1, 2, \dots, B$), $\widehat{\boldsymbol{\theta}}_b^*$ indicates the vector of estimated parameters $\left\{\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}, \hat{\rho}_1, \hat{\rho}_0\right\}_b$ for the b^{th} bootstrapped sample, and $\overline{\widehat{\boldsymbol{\theta}}}_B$ is the average of the B bootstrapped estimates.

A major decision is the number of bootstraps, which has been discussed in several studies (e.g. Efron and Tibshirani, 1986; Davidson and McKinnon, 2000). In this study, we employ 500 bootstraps, which is considered sufficient. For details about the bootstrapping method, please refer to Efron and Tibshirani (1986), Greene (2012), or Train (2009). One of the drawbacks of using bootstrapping is estimation time, since we need to repeat the estimation process multiple times with resampled cases. As a partial remedy, we employ parallel computation (i.e. giving computation tasks to multiple cores instead of one; this

study uses 40 logical processors). Since each computation task (i.e. estimation of each bootstrapped sample) does not need to communicate with any others (i.e. no dependence), parallelization is particularly straightforward and beneficial.

5.2.2 Marginal effects

It is often of interest to see how choice probabilities change with respect to changes in an explanatory variable. Here we present several derivatives of probabilities (i.e. partial/marginal effects) with respect to variables of interest. For the binary case, derivatives of the joint probabilities with respect to W_m and X_m are given by:

$$\frac{\partial P(z_i=1,y_i=1)}{\partial W_{im}} = \frac{\partial \Phi_2[\boldsymbol{W_i\alpha}, \boldsymbol{X_i\beta_1}, \rho_1]}{\partial W_{im}} = \phi(\boldsymbol{W}_i\boldsymbol{\alpha})\Phi\left(\frac{\boldsymbol{X_i\beta_1}-\rho_1\boldsymbol{W_i\alpha}}{\sqrt{1-\rho_1^2}}\right)\alpha_m\,, \tag{5.16}$$

$$\frac{\partial P(z_i=1,y_i=1)}{\partial X_{im}} = \frac{\partial \Phi_2[\boldsymbol{W_i\alpha}, \boldsymbol{X_i\beta_1}, \rho_1]}{\partial X_{im}} = \phi(\boldsymbol{X}_i\boldsymbol{\beta}_1)\Phi\left(\frac{\boldsymbol{W_i\alpha}-\rho_1\boldsymbol{X_i\beta_1}}{\sqrt{1-\rho_1^2}}\right)\beta_{1m}\,, \tag{5.17}$$

and analogously for the remaining combinations of values for z and y.

Derivatives of the choice probabilities conditional on membership (for $z_i = 1$, $y_i = 1$), with respect to X_m and W_m, are given by:

$$\frac{\partial P(y_i=1|z_i=1)}{\partial X_{im}} = \frac{\partial \frac{P(z_i=1,y_i=1)}{P(z_i=1)}}{\partial X_{im}} = \frac{1}{\Phi(\boldsymbol{W_i\alpha})}\phi(\boldsymbol{X}_i\boldsymbol{\beta}_1)\Phi\left(\frac{\boldsymbol{W_i\alpha}-\rho_1\boldsymbol{X_i\beta_1}}{\sqrt{1-\rho_1^2}}\right)\beta_{1m}\,, \tag{5.18}$$

$$\frac{\partial P(y_i=1|z_i=1)}{\partial W_{im}} = \frac{\partial \frac{P(z_i=1,y_i=1)}{P(z_i=1)}}{\partial W_{im}}$$

$$= P'(z_i = 1, y_i = 1) * \frac{1}{P(z_i=1)} + P(z_i = 1, y_i = 1) * \left(\frac{1}{P(z_i=1)}\right)'$$

$$= \frac{1}{\Phi(\boldsymbol{W}_i\boldsymbol{\alpha})}\phi(\boldsymbol{W}_i\boldsymbol{\alpha})\Phi\left(\frac{\boldsymbol{X}_i\boldsymbol{\beta}_1 - \rho_1\boldsymbol{W}_i\boldsymbol{\alpha}}{\sqrt{1-\rho_1^2}}\right)\alpha_m +$$

$$\Phi_2[\boldsymbol{W}_i\boldsymbol{\alpha}, \boldsymbol{X}_i\boldsymbol{\beta}_1, \rho_1]\left(-\frac{1}{\Phi(\boldsymbol{W}_i\boldsymbol{\alpha})^2}\right)\phi(\boldsymbol{W}_i\boldsymbol{\alpha})\alpha_m \tag{5.19}$$

The derivative of the marginal membership probability (for $z_i = 1$) with respect to W_m is:

$$\frac{\partial P(z_i=1)}{\partial W_{im}} = \phi(\boldsymbol{W}_i\boldsymbol{\alpha})\alpha_m \ . \tag{5.20}$$

As a special case, when the membership and outcome models are independent (i.e. $\rho_1 = 0$),

$$\frac{\partial P(z_i=1,y_i=1)}{\partial W_{im}} = \frac{\partial \Phi_2[\boldsymbol{W}_i\boldsymbol{\alpha}, \boldsymbol{X}_i\boldsymbol{\beta}_1, \rho_1=0]}{\partial W_{im}} = \phi(\boldsymbol{W}_i\boldsymbol{\alpha})\Phi(\boldsymbol{X}_i\boldsymbol{\beta}_1)\alpha_m \ , \tag{5.21}$$

$$\frac{\partial P(z_i=1,y_i=1)}{\partial X_{im}} = \frac{\partial \Phi_2[\boldsymbol{W}_i\boldsymbol{\alpha}, \boldsymbol{X}_i\boldsymbol{\beta}_1, \rho_1=0]}{\partial X_{im}} = \phi(\boldsymbol{X}_i\boldsymbol{\beta}_1)\Phi(\boldsymbol{W}_i\boldsymbol{\alpha})\beta_{1m} \ , \tag{5.22}$$

$$\frac{\partial P(y_i=1|z_i=1)}{\partial X_{im}} = \frac{\partial \frac{P(z_i=1,y_i=1)}{P(z_i=1)}}{\partial X_{im}}$$

$$= \frac{1}{\Phi(\boldsymbol{W}_i\boldsymbol{\alpha})}\phi(\boldsymbol{X}_i\boldsymbol{\beta}_1)\Phi(\boldsymbol{W}_i\boldsymbol{\alpha})\beta_{1m} = \phi(\boldsymbol{X}_i\boldsymbol{\beta}_1)\beta_{1m} \ , \tag{5.23}$$

and similarly for $z_i = 0$ when $\rho_0 = 0$.

Hence, when membership function and outcome function errors are uncorrelated, the derivative of the conditional outcome probability with respect to one of *its* explanatory variables is reduced to the marginal effect of a simple binary probit model (i.e. the effect on the outcome probability is independent of membership, as expected). The derivative of the conditional outcome probability with respect to one of the *membership function's* explanatory variables is 0:

$$\begin{aligned}\frac{\partial P(y_i=1|z_i=1)}{\partial W_{im}} &= \frac{1}{\Phi(\boldsymbol{W}_i\boldsymbol{\alpha})}\phi(\boldsymbol{W}_i\boldsymbol{\alpha})\Phi(\boldsymbol{X}_i\boldsymbol{\beta}_1)\alpha_m + \\ &\quad \Phi[\boldsymbol{W}_i\boldsymbol{\alpha}]\Phi[\boldsymbol{X}_i\boldsymbol{\beta}_1]\left(-\frac{1}{\Phi(\boldsymbol{W}_i\boldsymbol{\alpha})^2}\right)\phi(\boldsymbol{W}_i\boldsymbol{\alpha})\alpha_m \\ &= \frac{1}{\Phi(\boldsymbol{W}_i\boldsymbol{\alpha})}\phi(\boldsymbol{W}_i\boldsymbol{\alpha})\Phi(\boldsymbol{X}_i\boldsymbol{\beta}_1)\alpha_m - \\ &\quad \Phi[\boldsymbol{X}_i\boldsymbol{\beta}_1]\frac{1}{\Phi(\boldsymbol{W}_i\boldsymbol{\alpha})}\phi(\boldsymbol{W}_i\boldsymbol{\alpha})\alpha_m = 0\end{aligned} \tag{5.24}$$

It is worth noting two things. First, the derivations above are for cases where variables in the membership model ($\boldsymbol{W}$) and outcome model ($\boldsymbol{X}$) are mutually exclusive. In other words, variables do not have double roles in both models (otherwise, the derivatives would need to account for their role in both models). Second, when a variable is transformed in the model, an additional term is required. For example, in Study 2, W_{im} (population density) is actually modeled with $\ln W_{im}$ and thus to calculate the derivative with respect to the original variable, $\frac{\partial P(z_i=1)}{\partial W_{im}} = \phi(\boldsymbol{W}_i\boldsymbol{\alpha})\alpha_m * \left(\frac{1}{W_{im}}\right)$ by the chain rule ($\frac{dP}{dW} = \frac{dP}{dr}\frac{dr}{dW}$ where $r = \ln W$ and $\frac{dr}{dW} = \frac{1}{W}$). This applies to other derivatives as well. For the ordinal case, the derivations are basically the same, after making the logical replacements in the bivariate probit derivatives.

5.2.3 Overview of empirical applications

In this study, we present two empirical applications of the proposed methodology. We want to note several things before getting into the empirical results. First, our case studies do not necessarily aim to demonstrate the superiority of the proposed method over conventional ones. For example, although the proposed method has, in theory, a more realistic assumption (relaxing a previous restriction), the goodness-of-fit improvement could be marginal. In our applications, the goodness-of-fit and parameter estimates turned out to be fairly similar between the standard and proposed models (we will discuss the reason for this in Section 5.5.1). However, we will show how the proposed method could exhibit different behavioral implications. In addition, in both cases, although we attended to conceptualization based on the literature and experimented with various model specifications, the final models may not be necessarily the best ones for explaining the behaviors in question. Rather, we aim to present the potential of the proposed method for travel behavior applications. Particularly, variables related to emerging transportation services are selected (ridehailing services and autonomous vehicles, AVs). This is because although they have gained substantial attention recently, factors related to use of ridehailing services and willingness to use AV-based services are relatively less studied. Hence, we may not be able to correctly specify the models and thus there could be greater potential for having substantial influence of unobserved variables (or, variables that are in fact observed, but inadvertently omitted from the models). If so, specifying error correlations might be helpful. Figure 5-1 illustrates the two modeling frameworks; details will be delineated in the following subsections.

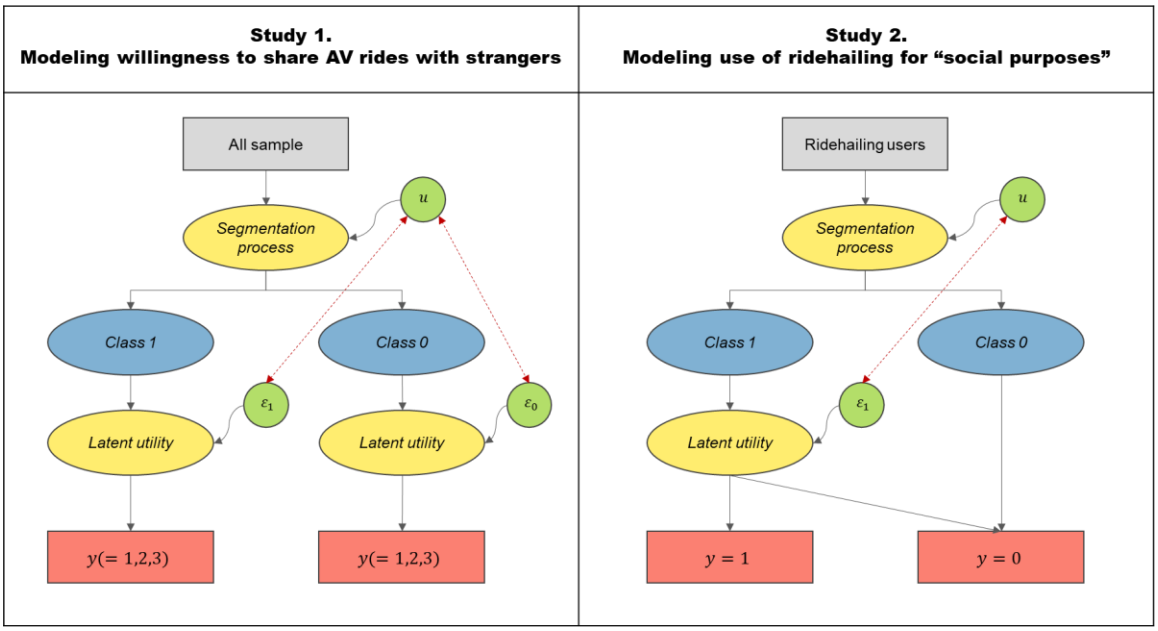


Figure 5-1. Conceptual diagrams of the two empirical models

5.3 Empirical application (1)

Study 1. Modeling willingness to share AVs with strangers (ordinal outcome variable)

5.3.1 Background

Studies on autonomous vehicles (AVs) have proliferated over recent years, demonstrating the lofty expectations regarding their potential impacts. Since the primary goal of this study is to show an application of the proposed method, we do not present a fine-grained snapshot of AV studies here. In this study, we are interested in the willingness to use AVs, particularly for ridehailing and even more particularly for sharing a ride in one. In other words, we want to learn not just about the willingness to use AVs for ridehailing services (i.e. sequential sharing), but also about the willingness to (simultaneously) share the ride with strangers (as in currently-available UberPOOL services). It is important to

envision such willingness because, although everything is uncertain yet, AV-based dynamic ridesharing is considered to be a competitive business model that might affect overall vehicle ownership and vehicle miles traveled (e.g. Fagnant and Kockelman, 2018; Gurumurthy and Kockelman, 2018).

Due to its importance, several studies have examined the willingness to use AVs. However, there are some variations because the measurement scales and AV use configurations vary across studies. Barbour et al. (2019) modeled the binary willingness (yes/no) to use shared AVs (SAVs; e.g. the willingness to share one's own AV with strangers, or to share an AV ride with strangers) with a random parameters logit model. Lavieri et al. (2017) jointly modeled interest in AV adoption (with options of no interest, AV sharing only, AV ownership only, and both AV sharing and AV ownership; the meaning of sharing was not specified), current behavioral choices, and vehicle ownership. Nazari et al. (2018) developed a multivariate ordered probit model of the level of interest in private AVs and four types of SAVs (renting an AV, using an AV with and without a backup driver, and using an AV for a short trip to get to a vehicle; the presence of strangers in the vehicle was not specified). Lavieri and Bhat (2019) conducted a stated preference survey to see tradeoffs between solo and shared options (with strangers) when choosing AV rides.

Although relevant studies are available, the current study mainly focuses on the general willingness to share AV rides with strangers and takes a different methodological approach which aims to identify distinct segments having heterogeneous sensitivities to factors. Note that there could be volatility in general willingness, given the uncertainty that is embedded in any AV study. The current study collected data in 2017-2018 and thus

general willingness was measured based on 2017-2018 knowledge about AVs. In addition, the COVID-19 pandemic brings another dimension of uncertainty. The general willingness to share rides could be reshaped based on the post-pandemic "new normal". For example, aside from supply-side issues (e.g. how ridehailing companies change their operation strategy), a sizable number of consumers may not fully lose their fear of exposure to the virus. It is not clear to what extent the general willingness will change, but we can certainly expect it to be lower, to some extent, than if we had not had the pandemic.

5.3.2 Data and modeling approach

This study employs the GDOT data (Section 1.2.3). In the AV section of the survey, respondents were asked to respond to the statement "I would use a driverless taxi with other passengers who are strangers to me (like UberPOOL)". Originally, a five-level option was provided ("very unlikely" to "very likely"), but we collapsed them into three options ("unlikely", "somewhat likely", and "likely") for modeling purposes. There are companion papers using the same dataset, but envisioning AV futures from different angles such as short-term mode use propensities (Kim et al., 2019a), possible mid-term behavioral changes (Kim et al., 2020a), and longer-term residential and vehicle ownership changes (Kim et al., 2020b).

For descriptive purposes, the sample was initially weighted to represent the population on nine selected variables (refer to Kim et al., 2019b for more details about weighting; however, the data were not reweighted, nor were factor scores restandardized, after some cases were excluded due to missing data or ineligibility). For simplicity (and reflecting the fact that models portray conditional relationships between variables rather

than absolute distributions, and therefore that population representativeness is less important in that context), the models were estimated on the unweighted data. Table 5-1 shows the descriptive statistics for the unweighted and weighted sample used in Study 1.

The basic model structure is as illustrated in Figure 5-1. Among the major decisions related to latent class modeling, we need to determine membership factors and outcome factors. As noted in CHAPTER 2, such model specification issues are is less-often discussed in the literature. In the early generation of marketing literature, no prior information was used for segmentation – i.e. segmentation was modeled with only intercepts and thus every individual had the same class membership probabilities – (Kamakura and Russell, 1989). Gupta and Chintagunta (1994) introduced demographic information for segmentation in latent class models. Swait (1994) provided a conceptual framework for the latent segmentation model, where class membership is formulated with general attitudes/perceptions as well as socio-demographics. Kim and Mokhtarian (2018) employed attitudes as the segmentation basis of the latent class model, based on the premise that attitudes moderate the effects of socio-demographics and the built environment on the vehicle ownership decision. In this application, we follow the approach of Kim and Mokhtarian (2018) and thus specify attitudes as the segmentation basis on the expectation that they moderate demographic effects on the outcome. Again, the primary goal of this application is to illustrate the proposed method rather than to find the best model specification *per se*; hence the final solution may not be the best one.

Table 5-1. Descriptive statistics of the sample (Study 1, N=3,215)

Variable	Category	Unweighted share	Weighted share
Willingness to share AV rides with strangers	Unlikely	0.706	0.697
	Somewhat likely	0.183	0.182
	Likely	0.111	0.121
Gender	Female	0.487	0.518
Education	High school or less	0.126	0.316
	Some college	0.295	0.322
	4-year degree or higher	0.579	0.362
Age	18-44	0.193	0.402
Use of ridehailing services	Have used ridehailing	0.344	0.398
Variable	**Category**	**Unweighted mean**	**Weighted mean**
Attitudes [a]	Pro-no-car-modes	0.004	-0.007
	Tech-savvy	0.014	0.180
	Urbanite	-0.001	0.132
	Sociable	-0.001	-0.039

a. Attitudes are estimated factor scores (standardized) obtained by applying factor analysis to attitudinal statements. Statements highly loading on each factor, with their loadings, are reported in Table 1-1.

5.3.3 Results

Table 5-2 exhibits the estimation results of two models: a standard latent class model, and a latent class model with correlated errors. At a glance, the two models are similar to each other with respect to their estimated coefficients and final log-likelihood values. They also produce consistent parameter interpretations. Two latent classes are identified as a function of attitudes. Class 1 consists of those who are relatively less favorable to non-car modes, less tech-savvy, less urbanite, and less sociable. In both models, class 1 has notable features. As shown in Table 5-3, class 1 mostly produces the conditional choice of "unlikely" – even more drastically in the proposed model than in the standard latent class model. Considering that most estimated parameters (except intercepts) in class 1 are not statistically different from zeros, the latent class models have

distinguished a group of people who are "structurally unwilling" to share AV rides with strangers regardless of any factors included. Unlike class 1, class 0 presents average choice probabilities more distributed over the alternatives and has significant sensitivities to factors. Males, more educated, and those who have used ridehailing services are more willing to share AV rides with strangers. Females may be, on average, more afraid of taking rides with strangers and education might be associated with familiarity with new technologies and AVs. Experiences with using ridehailing services would give individuals more confidence in taking a ride with strangers.

There are several differences between the two models. First, the proposed model presents a negative error correlation for class 0 that is statistically significant at the 0.01 level. This means that unobserved variables that increase the propensity to belong to the "structurally unwilling" class (i.e. that increase u_i of Eq. (5.1)) tend to reduce the propensity to share (i.e., decrease ε_{i0} of Eq. (5.5)) for people who actually belong to the "potentially willing" class. Thus, the negative correlation for class 0 seems reasonable.[45] This is an important finding, that dependency between latent segmentation and behavior generation is corroborated in this empirical context. Due to this error structure, we observe that the two models bring about different results. For example, we can confirm that the two models even give different pictures in a scenario analysis. Figure 5-2 presents the expected effect of the pro-no-car-mode attitude (a membership variable) on the conditional and joint

[45] On the other hand, if the same logic is applied to class 1, then its estimated positive correlation could be counterintuitive. Although the estimate is not small, it is not significant at the 0.05 level. If taken at face value, however, the opposite signs of the two error correlations mean that unobserved factors that increase the propensity to be in class 1 (increasing u_i) tend to decrease the willingness to share (decreasing ε_{i0}) for class 0 (as expected), while increasing it (increasing ε_{i1}) for class 1. It is challenging to identify such factors, but in principle the phenomenon is quite possible, just as the same *observed* variable can have coefficients with opposite signs in the two outcome models (thereby demonstrating how valuable it is to *have* segmented models).

choice probabilities for class 0. The slope of the tangent line at each point of the graphs is the marginal effect of the selected variable that is derived in Section 5.2. As the pro-no-car-modes attitude becomes stronger, the expected conditional choice probability of being "unlikely to share" decreases and probabilities of the other options increase, whereas they are constant in the standard latent class model (this is already derived in Section 5.2). In other words, when there is no endogeneity, the standard latent class model is not sensitive to factors related to segmentation, whereas the proposed model is. This demonstrates that even if the standard and proposed models have seemingly similar parameter values, they may produce different implications for prediction. For instance, in this case, if the overall attitudinal propensities are changing over the years, then the landscape of classes would change and the distribution of their behavioral decisions as well.

Table 5-2. Estimation results of Study 1 (N=3,215)

Membership component (for class 1)	***Standard latent class model***		***Latent class with error correlations***	
	Estimate	t-value	Estimate	t-value
Intercept	0.454	1.71	-0.068	-0.22
Pro-no-car-modes	**-0.711**	**-3.35**	**-0.523**	**-3.17**
Tech-savvy	-0.191	-1.42	-0.140	-1.45
Urbanite	**-0.312**	**-2.38**	**-0.241**	**-2.35**
Sociable	**-0.232**	**-2.57**	**-0.175**	**-2.59**
Outcome component (class 1)				
	Estimate	t-value	Estimate	t-value
Intercept 1 | 2	**1.125**	**7.90**	**0.917**	**4.23**
Intercept 2 | 3	**1.759**	**10.36**	**1.288**	**4.80**
Gender (female=1)	-0.312	-1.31	-0.399	-1.22
Education (college=1)	-0.232	-0.66	-0.344	-0.83
Education (bachelor's or graduate=1)	0.039	0.20	-0.053	-0.23
Age (18-44=1)	0.025	0.12	-0.005	-0.02
Have used ridehailing services (=1)	0.342	1.73	0.190	0.61
Outcome component (class 0)				
	Estimate	t-value	Estimate	t-value
Intercept 1 | 2	0.275	1.31	0.143	0.69
Intercept 2 | 3	**1.273**	**7.34**	**1.046**	**5.99**
Gender (female=1)	**-0.302**	**-2.60**	**-0.235**	**-2.28**
Education (college=1)	**0.489**	**2.68**	**0.402**	**2.17**
Education (bachelor's or graduate=1)	**0.463**	**2.81**	**0.403**	**2.95**
Age (18-44=1)	0.117	0.90	0.084	0.78
Have used ridehailing services (=1)	**0.442**	**3.61**	**0.418**	**3.95**
Error correlations				
Rho parameter (class 1)	-	-	0.375	1.80
Rho parameter (class 0)	-	-	**-0.474**	**-2.64**
Summary				
Class share (class 1, class 0)	0.632	0.368	0.451	0.549
Number of parameters	19		21	
Log-likelihood at zero	-3532.0385		-3532.0385	
Log-likelihood at convergence	-2397.3835		-2394.8090	
McFadden's R^2	0.3212		0.3225	

Table 5-3. Average marginal, conditional, joint choice probabilities

	Standard latent class model			*Latent class with error correlations*		
	Unlikely	Somewhat likely	Likely	Unlikely	Somewhat likely	Likely
Marginal choice (class 1)	-	-	-	0.873	0.060	0.067
Marginal choice (class 0)	-	-	-	0.403	0.329	0.267
Conditional choice (class 1)	0.877	0.085	0.038	0.938	0.034	0.028
Conditional choice (class 0)	0.438	0.348	0.215	0.555	0.295	0.150
Joint choice (class 1)	0.558	0.051	0.023	0.444	0.018	0.015
Joint choice (class 0)	0.152	0.130	0.086	0.266	0.164	0.093

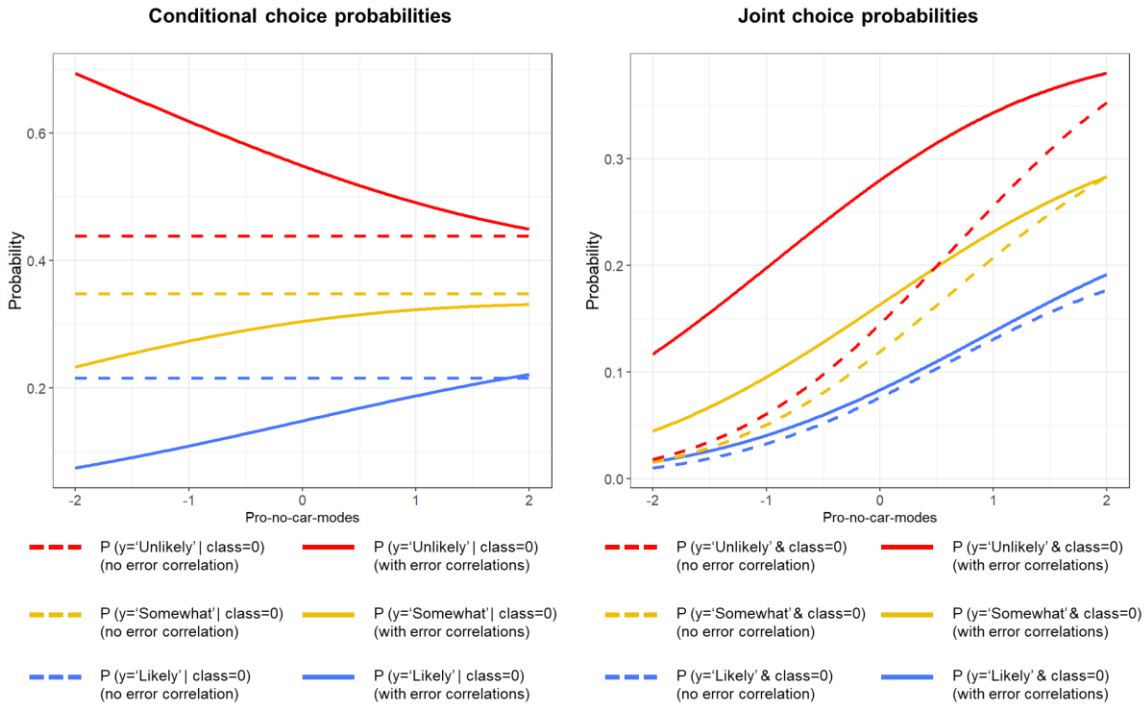


Figure 5-2. Scenario analysis of the impact of the pro-no-car-mode attitude on class 0's choice probabilities

5.4 Empirical application (2)

Study 2. Modeling ridehailing use for social purposes (binary outcome variable)

5.4.1 Background

There is a growing body of literature uncovering behaviors related to the use of ridehailing services. A few studies have started to analyze actual trip data to understand

general trip characteristics (e.g. Dias et al., 2019; Yan et al., 2020; Soria et al., 2020). However, important variables such as user demographics and detailed trip information are usually missing in such actual trip data. Such data can play useful roles, but they lack the ability to improve our understanding in key ways. An important piece of information, we think, is trip purpose. Why do people use ridehailing services and what factors motivate them to ridehail for such purposes?

Ridehailing usually exhibits an evening/night-peak of trip generation that is distinct from other modes. For example, demand peaked around 7-8 pm and exhibited a high plateau in the late evening in San Francisco (Castiglione et al., 2016). Dias et al. (2019) reported that 55% of RideAustin trips were at night (10 pm-7 am). From this fact we can infer that many trips might be for personal/social purposes. Several studies presented descriptive statistics for trip purposes. Young and Farber (2019) used a household travel survey in Southern Ontario, Canada, and found that "other" purpose trips (including entertainment, personal business, social and recreational trips) comprised about 28% of the ridehailing total (this is substantial given that 14% of trips by any mode are "other" purpose trips). Tirachini and del Rio (2019) surveyed residents in the Santiago (Chile) metropolitan region in 2017, and found that the most common purpose of ridehailing use was leisure (55.4%) followed by work (17.4%) and others. This finding – that the major trip purpose of ridehailing users is leisure/social – is consistent with other studies (e.g. Clewlow and Mishra, 2017; de Souza Silva et al., 2018). A limited number of studies aimed to model trip purposes. Dias et al. (2019) employed RideAustin trip data for modeling the number of ridehailing trips by destination trip purpose; they inferred trip purpose based on land use characteristics of trip origins/ destinations since trip data do not have such information.

Lavieri and Bhat (2019) is the only study we are aware of that aimed to identify factors affecting trip purpose. Using a sample from the Dallas-Fort Worth-Arlington Metropolitan Area of Texas, they developed a multivariate multinomial probit model of several characteristics (including purpose as well as time-of-day, companionship, and self-reported mode substitution) of the individual's last ridehailing trip.

5.4.2 Data and modeling approach

This study employs the GDOT data (Section 1.2.3). Aside from the fact that the Study 1 and Study 2 used partially different variables, this study uses only a subset of the sample (people who have used ridehailing services), given its focus on the purpose of ridehailing trips (Table 5-4; the data were not reweighted, nor were factor scores restandardized, after some cases were excluded due to missing data or ineligibility). In the survey, respondents were asked to check the purposes of the trips that they had made by ridehailing or shared ridehailing services. The dependent variable is a binary indicator of whether the respondent (who has used ridehailing services for any purpose) has used either ridehailing or shared ridehailing services for any of shopping/eating/drinking/social/ recreational purposes.

In this study, we hypothesize that some people use ridehailing services only for mandatory (e.g. work-related) or occasional events (e.g. airport, sports event), while others use them for a wider array of purposes including social purposes. Accordingly, in the context of identifying the relevant factors motivating people to use ridehailing services for more general purposes, particularly social purposes, we speculate that there is a type of person who has a greater *potential* of using ridehailing services for social purposes (even

if not all such people have actually have *used* them for those purposes) and another type of person who never considers ridehailing services for social purposes (and thus structurally generates no choice for social purposes), even if using such services for other reasons. This model takes a *confirmatory* latent class approach as opposed to the more common exploratory approach (cf. CHAPTER 2). In other words, we hypothesize certain data generation processes for each of the classes and check if this presumption is valid based on the data, whereas typical latent class models do not impose particular forms of data generation processes and thus each of the classes has the same form (but parameters are freely estimated to allow parameter heterogeneity). The proposed model resembles a zero-inflated model, which detaches the generation of "structural/systematic zeros" from the usual behavior generation process (e.g. CHAPTER 4).

We derived the probability and log-likelihood functions in Section 5.2 under the assumption that the two regimes have outcome equations with parameter heterogeneity (i.e. assuming the usual exploratory latent class approach). In this case study, because of the hypothesis described above, we slightly modify the probability functions (Eq. (5.25) to demonstrate that the proposed method can be also applied to the confirmatory latent class approach. In particular, we assume two regimes, where one regime ($z_i = 1$) has the potential to generate various outcomes (here, $y_i = 1$ or $y_i = 0$) and the other ($z_i = 0$) structurally generates a particular outcome (here, the outcome "not used for such purposes", i.e. only $y_i = 0$). Hence, in this setting, as illustrated in Figure 5-1, there is a single error correlation to be estimated, whereas the usual model may have the number of error correlations equivalent to the number of latent classes. For estimation of this model, we replace the marginal choice probabilities of Eq. (5.6) and error structure of Eq. (5.7)

with Eqs. (5.25) and (5.26), respectively, and insert Eq. (5.25) into the log-likelihood function, Eq. (5.13). Since $z = 0$ (i.e. a structural zero) is assumed not to have a behavioral process in this confirmatory approach, $P(z_i = 0, y_i = 1|\boldsymbol{X}_i, \boldsymbol{W}_i)$ and $P(z_i = 0, y_i = 0|\boldsymbol{X}_i, \boldsymbol{W}_i)$ in Eq. (5.8) are not specified in this application. Notation follows the definitions in Section 5.2.

$$P(y_i = 1|\boldsymbol{X}_i, \boldsymbol{W}_i) = P(z_i = 1, y_i = 1|\boldsymbol{X}_i, \boldsymbol{W}_i) = \Phi_2[\boldsymbol{W}_i\boldsymbol{\alpha}, \boldsymbol{X}_i\boldsymbol{\beta}_1, \rho_1] ,$$

$$\begin{aligned} P(y_i = 0|\boldsymbol{X}_i, \boldsymbol{W}_i) &= P(z_i = 1, y_i = 0|\boldsymbol{X}_i, \boldsymbol{W}_i) + P(z_i = 0|\boldsymbol{X}_i, \boldsymbol{W}_i) \\ &= \Phi_2[\boldsymbol{W}_i\boldsymbol{\alpha}, -\boldsymbol{X}_i\boldsymbol{\beta}_1, -\rho_1] + \Phi[-\boldsymbol{W}_i\boldsymbol{\alpha}] \text{ , and} \end{aligned} \quad (5.25)$$

$$\begin{pmatrix} u_i \\ \varepsilon_{i1} \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho_1 \\ \rho_1 & 1 \end{pmatrix}\right] . \quad (5.26)$$

Table 5-4. Descriptive statistics of the sample (Study 2, N=1,105 ridehailing users)

Variable	Category	Unweighted share	Weighted share
Use of ridehailing services	Have used ridehailing for social trips	0.700	0.710
Gender	Female	0.484	0.512
Age	18-34	0.179	0.404
Household income	Below $50,000	0.176	0.247
	$50,000-$99,999	0.340	0.350
	$100,000+	0.483	0.403
Education	4-year degree or higher	0.740	0.591
Race	White	0.784	0.612
Variable	**Category**	**Unweighted mean**	**Weighted mean**
Residential characteristics	Population density (per acre)	3.698	1.052
Attitudes	Tech-savvy	0.381	0.612
	Pro-no-car-mode	0.245	0.211

5.4.3 Results

Table 5-5 exhibits the estimation results for two confirmatory latent class models, having uncorrelated and correlated errors, respectively. At a glance, as in Study 1, the two models are similar to each other with respect to the face values of coefficient estimates and final log-likelihoods. Both models present logical estimation results. As imposed, we identify two latent classes: one for those who are "potential social-purpose ridehailers" (class 1) and the other for those who are "structurally zero social-purpose ridehailers" (class 0). As expected, class 0 tends to include those living in less-populated areas, and their share in the sample is marginal (about 10%).[46] In all, 30% of the sample did not ridehail for social purposes, and our models decomposed that group into two: those who made zero social ridehailing trips because they are presumably structurally uninterested in ridehailing for social trips (the 10% of cases belonging to class 0), and those who incidentally did not make social trips but may do so in the future (the remaining 20%, who belong to class 1). In both models, for those belonging to class 1, several demographics and attitudes are related to trip-making behaviors. Unsurprisingly, younger people, higher-income people, and whites are more likely to have used ridehailing services for social trips. The education dummy is not statistically significant in either model. Females are more likely to use ridehailing services for social trips, although not significantly so at the 0.01 level. We found one attitude – tech-savviness – to be positively significant in the models, indicating that those who are tech-savvy are more likely to use ridehailing services for social trips. In our model experiments (not presented in this chapter), other attitudes had significant impacts

[46] This, and the other shares mentioned in this paragraph, should not be considered representative of the population shares, since the model was estimated on the unweighted sample.

on whether to use ridehailing services *at all* or not, but, *given* that a person has used such services, tech-savviness was the only factor significant to using them for social purposes. So far, the standard latent class model and the proposed model seem fairly similar. A difference comes from the correlation parameter ρ_1, which turns out to be statistically significant and positive. It indicates that unobserved characteristics influencing whether people belong to class 1 (the potential social-purpose ridehailers) and those influencing their choice to ridehail for social purposes are positively associated. The effects of correlation can be confirmed when we examine a scenario analysis of how the probabilities of interest change by various factors.

Figure 5-3 exhibits the average probabilities of class 1 membership and (conditional) choice to ridehail for social purposes, as functions of (raw) population density (which is a membership variable) while holding other variables at their sample values. This result shows that the two models (uncorrelated and correlated errors) could provide different pictures for prediction, although the parameter estimates and final log-likelihoods do not seem to be markedly different. As the membership factor (population density) changes, both models show similar increase patterns in class 1 membership probabilities. However, the conditional (on being in class 1) *behavioral* probabilities are rather different between the two models. They remain constant with respect to changes in population density for the model with uncorrelated errors, because the model assumed independence of the segmentation and behavior generation processes, and population density only affects the former. On the other hand, as population density increases, the conditional choice probabilities for the proposed method decrease, due to the association between segmentation and choice processes. Viewed the other way, when population density is very

low, then conditional on belonging to the *potential* social-purpose ridehailer class in the first place, the individual is far more likely to actually *choose* to ridehail for social purposes – presumably because *without* a strong motivation to actually ridehail for social purposes, someone in a very low-density area would otherwise belong to the structurally zero social-purpose ridehailer class. The general principle is that when unobservables of latent segmentation are associated with those of the behavior generation process, the assumption of independence could be incorrect, and a model which imposes that assumption could provide biased estimates of effects.

Table 5-5. Estimation results of the error-independent and error-correlated models (Study 2, N=1,105)

Membership component (for class 1)	***Latent class model with uncorrelated errors***		***Latent class model with correlated errors***	
	Estimate	t-value	Estimate	t-value
Intercept	**1.256**	**5.59**	**1.183**	**5.36**
Log-transformed population density	**0.208**	**2.05**	**0.247**	**2.16**
Outcome component (class 1 specific)				
	Estimate	t-value	Estimate	t-value
Intercept	**-0.478**	**-2.96**	**-0.551**	**-3.37**
Gender (female=1)	0.209	1.82	0.202	1.78
Age (18-34=1)	**0.842**	**3.55**	**0.854**	**3.69**
Income (medium income=1)	**0.401**	**2.38**	**0.395**	**2.40**
Income (higher income=1)	**0.480**	**2.78**	**0.478**	**2.90**
Education (bachelor's or graduate =1)	0.170	1.30	0.168	1.30
Race (white=1)	**0.660**	**5.25**	**0.637**	**5.24**
Attitude: Tech-savvy	**0.168**	**2.67**	**0.161**	**2.57**
Attitude: Pro-non-car-modes	0.051	0.85	0.052	0.86
Rho parameter	-	-	**0.546**	**6.65**
Summary				
Class 1 share	0.914		0.905	
Number of parameters	11		12	
Log-likelihood at zero	-765.9276		-765.9276	
Log-likelihood at convergence	-624.0779		-623.9649	
McFadden's R^2	0.1852		0.1853	

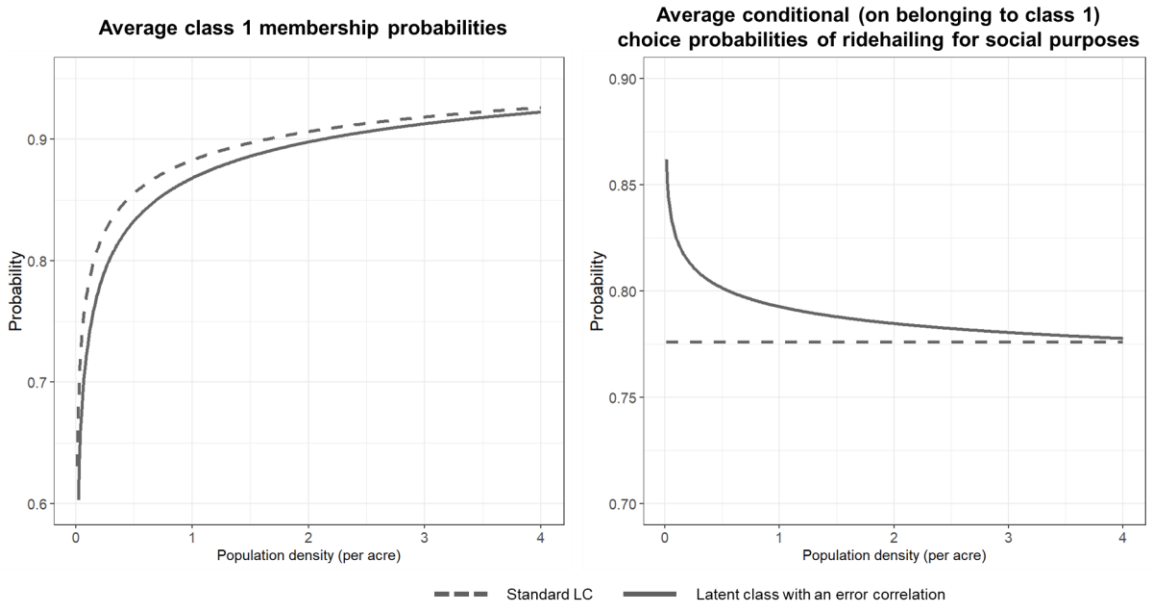


Figure 5-3. Scenario analysis of population density change

5.5 Conclusions

5.5.1 A note about the performance of the proposed models

Study 1 and Study 2 presented two empirical applications that compared standard latent class and the proposed models. As illustrated, the proposed method could be meaningful in its relaxation of the restrictive assumption of independence. We did not expect dramatic improvements in goodness of fit given that we are only adding one or two parameters (which are not even based on "new" information), but the improvements we obtained were modest, to say the least. The degree of improvement might be dependent on the empirical context, but we think improvement, in general, could be marginal due to the nature of latent class models.

To see this, first consider how we evaluate *conventional* latent class models. In general, compared to a single-class model, latent class models (with two or more classes) bring notably better performance with respect to log-likelihood-based information criteria. Although model assessment based on information criteria is important, in fact, such improvements do not represent all the benefits of latent class models. A real benefit (and perhaps the main goal of using latent class models in many cases) is that (to echo previous discussion) the model decomposes the sample into subgroups that might have different behavior generation processes (this "heterogeneity" can be viewed from many angles – please refer to CHAPTER 2). The implication is that this step allows us to understand/interpret (1) class-specific behavior generation processes and (2) how the sample (population) consists of subsegments – this includes the overall shares of classes or probabilities of belonging to classes of individuals, as well as profiles of the average member of each class. However, the final model log-likelihood only partly captures this benefit. Since we do not know true class membership, we evaluate the log-likelihood (Eq. (5.13)) using *unconditional* (marginal) probabilities (Eq. (5.6) or the first equalities of Eq. (5.25)) rather than *conditional* probabilities. If the latent class models captured true heterogeneity adequately, there should be a notable improvement in the overall likelihood, but the unconditional likelihood is an imperfect measure of the effect of segmentation.

Turning now to the introduction of an error correlation structure, a key merit of doing so stems from modeling how two (or more) decisions are associated and thus how *joint* probabilities behave. Since the unconditional likelihood does not explicitly represent the joint choice likelihood, likelihood-based information criteria, again, cannot fully capture the benefit of latent segmentation with an error correlation structure. Because, as

already mentioned, the proposed method is a variant of the standard joint model, it is easier to think of this issue in comparison with a standard bivariate probit model. When estimating a bivariate probit model (say, two binary choices), we evaluate the likelihood with the information on two choice indicators whose values we know. Hence, in this example, the likelihood is evaluated using 2×2 joint choice cells. From this viewpoint, if error correlations adequately captured joint decisions, their roles are reflected in the *joint-choice-based likelihood* evaluation of the model; hence the improvement of goodness of fit with the aid of an error structure could be significant. On the other hand, in latent class models, (again) we do not know the true membership indicator and thus the likelihood is only evaluated with respect to the outcome choice indicator (i.e. based on only one of the choice dimensions). Thus, since latent class models – with or without an error structure – only evaluate *marginal-choice-based likelihoods*, the benefit of introducing an error structure is undervalued. This is perhaps why the standard latent class model and the latent class model with an error structure may have similar log-likelihoods, even if they present different pictures of jointness.

5.5.2 Summary and contributions

In spite of the recent great popularity of latent class models, discussions about whether their basic assumptions are valid in empirical contexts are scarce. This study questioned whether one of those basic assumptions, namely that of independence between latent segmentation and the behavior generation process, is tenable. The study formulated the latent class model where unobserved influences on latent segmentation and behavior generation are correlated, by combining the key concepts of latent class and endogenous switching modeling. The proposed method is implemented in two empirical applications.

In the first application, the dependent variable was measured on an ordinal scale and the model took an "exploratory" latent class approach (i.e. there was no imposed difference across class-specific outcome models). In the context of modeling willingness to share AV rides with strangers, standard latent class and the proposed method provided reasonable results. In the second application, the dependent variable was binary and the model took a "confirmatory" latent class approach (i.e. there was an imposed difference between the two class-specific behaviors). Here, in modeling whether a person has used ridehailing services for social purposes, class 0 is posited to produce systematic zeros, whereas class 1 followed a typical behavior process. In both applications, error correlations were statistically significant, indicating that segmentation and behavior generation processes were jointly determined. Our scenario analyses showed how the proposed model could be useful due to its consideration of jointness with correlations. Improvements in goodness-of-fit were relatively small in our applications, and thus we discussed major reasons for such phenomena.

The main contribution of the study is to question the validity of a basic assumption of the standard latent class model and to provide a potential avenue of methodological development. The usefulness of the method may depend on the purpose and research question of the study. For example, if the study aims for better predictive ability per se, the method may not be necessarily appealing (the issue of explanation versus prediction has been discussed in studies such as Shmueli, 2010; Mannering et al., 2020). Rather, the proposed method could provide a framework with a conceptually more realistic assumption. If the association between segmentation and behavior processes is true given the empirical context, then estimates should be more reliable than those under the

independence assumption, which is more restrictive. Models using the proposed method can therefore provide more appropriate predictions of joint choices. As well, this approach may open the door to an avenue of evaluating "treatment effects" in the latent class modeling context. In the standard latent class model, latent classes act as intrinsically-fixed characteristics of individuals. However, as an example, if individuals change their attitudinal propensities (and thus "switch" their classes) by an education program, then the study may want to measure its treatment effect as statistically controlling for potential self-selection effects.

5.5.3 Future directions

A number of further studies could profitably be pursued. The study proposed methodologies under the two-class context. As noted by CHAPTER 2, more than half of the empirical studies employing latent class models (in the transportation domain) have used two classes for the final solution. Hence, two-class solutions are expected to cover many research contexts. However, the methodology can be expanded to three or more classes. In such cases, the error structure should be enlarged to accommodate more error correlations. The authors think, however, that such an effort should be preceded by solid hypotheses, because adding complexity could be less beneficial compared to the effort when classes are *latent* and thus the conceptual validity of error correlations could be less straightforward. In other words, in any case, allowing a less-restrictive error structure should be more realistic in theory, but testing the conceptual validity of any particular such structure could be challenging. Second, the study employed probit link functions for membership models instead of the conventional logit link function. This assumption was for convenience in that joint or marginal probability densities of the bivariate normal

distribution are well-known. However, other link functions are worth trying. One rationale is that since probit has thin tails, sometimes it could be computationally less stable. For example, Dubey et al. (2020) proposed a t-distributed error kernel (so-called robit link) for multinomial response models. Adopting this idea, a possibility is to use the bivariate t-distribution instead of the bivariate normal distribution. Another plausible avenue is to introduce the copula-based approach (cf. Bhat and Eluru, 2009), if the normality assumption is considered too strict. Lastly, although this study proposed the methodology and applied it to two empirical contexts (here, willingness to share AVs with strangers and the purpose of ridehailing use), more theoretical and empirical studies should be pursued to confirm the usefulness and potential of this approach. For example, as we discussed, the proposed method can be a way of evaluating treatment effects when endogeneity exists. Thus, future studies may explore this research avenue.

CHAPTER 6. MIXTURE OF EXPERTS AND NONLINEAR/INTERACTION EFFECTS

Paper title: *Mixture of experts as a data-driven exploratory tool for improving conventional model specifications*

6.1 Introduction

6.1.1 Use of machine learning in the transportation domain

Machine learning, or so-called "data-driven", approaches have proved their success with respect to performance over the past several decades. Machine learning has been built within a community of its own, but it has now penetrated into a variety of domains, and is seemingly becoming ubiquitous. Many scholars in various domains have been pondering the implications of introducing machine learning (or deep learning) into their applications, such as in economics (Mullainathan and Spiess, 2017; Athey, 2018; Athey and Imbens, 2019), psychology (Yarkoni and Westfall, 2017; Urban and Gates, 2019; Orru et al., 2020), sociology (Molina and Garip, 2019), choice modeling (Timothy et al., 2017; Hillel et al., 2021), and so on. Common conclusions mostly include consensus on the usefulness of machine learning models with respect to their prediction ability and, at the same time, some reservations on their usefulness due to their lack of interpretability. In response to this lament, interpretable/explainable machine learning (or artificial intelligence) has become a popular topic of research.

In the travel behavior and choice modeling fields, applications of machine learning are notably increasing. In the context of mode choice modeling, artificial neural

networks (ANN) have been the most popular machine learning technique following the logit family (Hillel et al. 2021). Beyond the mechanical applications, a recent mainstream of work is to exploit machine learning models to extract behavioral implications. Some initial efforts were devoted to understanding how the learned model maps input features to the outcome, for which a useful tool is the partial dependence plot (cf. Friedman, 2001; Hastie et al., 2009; Molnar, 2020; Zhao and Hastie, 2021). For example, Zhao et al. (2020) compared the random forest and logit models (multinomial and mixed logit) with the aid of partial dependence plots and concluded, in their empirical context (p. 22), that "However, we find that the random forest model produces behaviorally unreasonable arc elasticities and marginal effects when these behavioral outputs are computed from a standard approach." Alwosheel et al. (2019) were inspired by an approach in computer vision and proposed to synthesize prototypical examples with a trained ANN to diagnose whether the learned model behaves reasonably. One avenue has been to understand the neural networks structure as a generalized logit model that contains implicit random utility maximization in the black box (Bentz and Merunka, 2000; Wang et al. 2020a; Wang et al. 2020b; Zhang et al. 2020).

Another approach, also related to the initial one, is to combine classical model structure and neural network structure under the premise of achieving a balance between interpretability and performance. Sifringer et al. (2020) proposed the learning multinomial logit (L-MNL) and learning nested logit (L-NL) models, which embrace a knowledge-driven part and a data-driven one in the systematic utility specification (i.e. embedding the neural network in the standard logit model). Han et al. (2020) proposed the TasteNet-MNL model. In this model, taste parameters in MNL are functionalized using neural networks to

capture systematic taste heterogeneity. Wang et al. (2021) developed the TB-ResNets model, which connects a classical discrete choice model and neural networks with a weighting parameter to represent deterministic utilities.

6.1.2 Challenge of model specification and the usage of machine learning

Despite the breakthrough of machine learning, theory-driven classical models have long played a critical role in informing public policy. In addition to the simplicity and interpretability of the classical models, there is arguably an inertia on the part of analysts and/or policymakers that favors the use of familiar models, plus the fact that the benefits of using more complex models are seemingly limited (e.g. datasets may not be large enough, or the "first-order" approximations of simpler models are sufficient). In these simpler models (such as classical choice models), the model specification (or the utility specification in choice modeling) is the key step. Misspecification could result in biased estimates and thus lead to an erroneous understanding of the behavioral process and lower prediction power (Abe, 1998; Torres et al., 2011; Van Der Pol et al., 2014). This misspecification could take the form of omitted relevant variables, the inclusion of irrelevant variables, and failure to capture nonlinear effects or interaction effects, among others (Greene, 2012). The key issue is that we almost always have limited prior knowledge of the correct specification (Ben-Akiva and Lerman, 1985) or even if prior knowledge is valid it may not hold in a particular empirical context, and thus we mostly rely on time-consuming trial-and-error model experiments. However, the trial-and-error approach is not only time-/effort-intensive, but also the solution space is too vast to be systematically searched; hence some specification possibilities, especially complex effects such as nonlinear or interaction effects, are often excluded from the outset.

A promising approach to tackling this challenge might be to take advantage of machine learning techniques *to aid in the process of specifying a conventional model*. Machine learning techniques, in general, have a larger search space for the solution and they can *automatically* detect nonlinear and interaction effects without prior information. Some studies have tried to use such benefits. For example, Hillel et al. (2019) built gradient boosting decision trees (GBDT) and extracted the learned structure to inform the utility specification for a choice model. Ortelli et al. (2021) translated the utility specification problem into a combinatorial (optimization) problem. In their study, the proposed method produced a better out-of-sample performance than a benchmark (simple) model and at the same time it ensured behavioral realism. Another method is to use some "approximator" such as neural networks. It is known that neural networks are universal approximators (Hornik et al., 1989; Cybenko, 1989); in other words, neural networks can approximate any function without any a priori assumptions. Inspired by this property, several studies utilized neural networks to obtain a better choice model specification (Bentz and Merunka, 2000; Hruschka et al., 2002; Hruschka et al., 2004; Hruschka, 2007). For example, Bentz and Merunka (2000) identified interaction and threshold effects in brand choice with the aid of neural networks and revised MNL models based on the learned knowledge. Although there was only a marginal improvement of model performance in the empirical application, indicating that there was only a weak nonlinear component in the utility function, the study offered an insightful pathway for using neural networks as a diagnostic tool.

In the present study, we take this pathway to improve our behavioral models. In other words, we use machine learning models as ***data-driven exploratory tools*** *to automatically identify nonlinear and interaction effects* and thus to improve model

specifications for “simpler” models that are more intuitive. To our knowledge, this is the first study to introduce the idea of mixture of experts (MoE), which will be described in the following section, to travel behavior research. More importantly, we illuminate the linkage between MoE and the latent class model, and suggest using MoE to capture complex effects. The expected benefits include the improvement of performance and the reduction of potential bias by taking into account *systematic heterogeneity* while keeping the model as simple (and, ideally, interpretable) as possible. The remainder of this chapter is as follows. Section 6.2 describes the pertinent methodologies. Section 6.3 presents model experiments with synthetic data and Section 6.4 exhibits the application of the approach to the well-known Swissmetro data. Section 6.5 concludes the chapter with some remarks.

6.2 Methodology

6.2.1 Mixture of experts

In this study, we use the *mixture of experts* (MoE) architecture proposed by Jacobs et al. (1991a). The basic idea is to design a model architecture that decomposes tasks and assigns them to “experts” (cf. Jacobs et al. 1991b). In other words, the model splits the input space into homogeneous regions, and different *local experts* (i.e. *models,* or also called *learners*) are “responsible for” (i.e. operating in) the different regions (Masoudnia and Ebrahimpour, 2014; Baldacchino et al., 2016). This approach is called “divide-and-conquer” from a problem-solving perspective (which is comparable to the concept of market segmentation). Then, results from the different experts are combined by a *gating network* (mostly employing the so-called softmax function, which is also known as multinomial logistic regression). This is a type of *ensemble method* that combines multiple

local models instead of having a "global" model (cf. Kotsiantis et al., 2006; Bishop, 2006; Rokach, 2010).

The main reasons for using MoE in this study are twofold. First, in the field of information theory, it has been proved that the MoE is a universal approximator (Zeevi et al., 1998; Nguyen et al., 2016). That is, it has the capability to approximate any unknown true mapping function between input features and outcome, to any specified degree of accuracy. Second, it is connected to the well-known *latent class* (or *finite mixture*) *model*. We can translate "homogeneous input spaces" into "latent classes" or "market segments", "gating network" into "membership/segmentation model", and "local expert/learner" into "class-specific outcome model".

The structures of the standard latent class model and the MoE are basically identical:

$$f(y|\boldsymbol{X}) = \sum_{z=1}^{Z} P(z|\boldsymbol{X};\boldsymbol{\alpha}) f_z(y|z,\boldsymbol{X};\boldsymbol{\beta}_z) \tag{6.1}$$

where y is an outcome variable, $\boldsymbol{X}$ is a vector of input features, z is a *discrete latent* segment or subgroup indicator ($z = 1,2,\dots Z$), $P(\cdot)$ denotes a (finite) mixture density function (or segment membership probability), and $f_z(\cdot)$ denotes an outcome function (or a local expert) for segment z. $P(\cdot)$ is typically represented by the softmax function and $f_z(\cdot)$ depends on the type of problem (cf. CHAPTER 2). In this study, as we focus on the binary classification problem, the $f_z(\cdot)$'s are logit functions. We maximize a log-likelihood function with an additional regularization term (if needed).

$$\max LL_{\lambda}(\boldsymbol{\theta}) = \max(\sum_{i=1}^{N} \ln[\sum_{z=1}^{Z} P(z_i|\boldsymbol{X}_i) f_z(y_i|z_i, \boldsymbol{X}_i)] + \lambda\|\boldsymbol{\theta}\|_2^2) \quad (6.2)$$

where $\boldsymbol{\theta}$ denotes all parameters in the model ($= \{\boldsymbol{\alpha}, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2 \dots, \boldsymbol{\beta}_Z\}$), $\boldsymbol{\alpha}$ represents a vector of parameters in $P(\cdot)$, $\boldsymbol{\beta}_z$ represents a vector of parameters in $f_z(\cdot)$, λ is the regularization coefficient, and $\lambda\ \|\boldsymbol{\theta}\|_2^2$ is the squared L2 regularization term[47].

Despite this similarity, there are a few practical differences. In most standard latent class models, as noted in CHAPTER 2 and CHAPTER 3, input features are split into two parts: variables that characterize latent classes (denoted as $\boldsymbol{W}$) and variables that feed to class-specific outcome functions (mostly alternative-specific characteristics, such as travel time/cost, in many choice models). On the other hand, in applications of MoE, $\boldsymbol{W}$ are usually not distinguished from $\boldsymbol{X}$; rather all $\boldsymbol{X}$ are used in $P(\cdot)$ and $f_z(\cdot)$. The former strategy could have value in that it might offer interpretation benefits (e.g. testing of hypotheses regarding whether certain characteristics are related to distinctive groups), whereas the latter could have value in that it can avoid a key dilemma of this model form (which variables should go into the membership versus outcome models?) and allow the data to "confess" the best specification (from a purely mechanical standpoint).

Second, as alluded to earlier, individual local experts (or latent classes) are mostly not the main focus in the MoE method; rather, it focuses on how they collectively produce the overall outcome. On the other hand, the interpretation of individual latent classes is the key in most standard latent class models. As noted in psychometrics, there are two

[47] $\|\boldsymbol{\theta}\|_2^2 = \sum_{m=1}^{M} \theta_m^2$, where m indexes parameters ($m = 1,2, \dots M$). Other regularizations (e.g. L1, $\lambda\|\boldsymbol{\theta}\|_1 = \lambda \sum_{m=1}^{M} |\theta_m|$) are also possible. If $\lambda = 0$, then there is no regularization on parameters. In this study, we confine ourselves to the squared L2 regularization for experiments.

approaches taken with finite mixture modeling: "direct" and "indirect" applications (cf. Bauer, 2005; Masyn, 2013; Cole and Bauer, 2016). In *direct* applications, the assumption of heterogeneous subgroups is the key idea and thus the main focus is on latent classes – studies in the travel behavior domain almost exclusively take this approach. Although less-often used, in *indirect* applications, the finite mixture structure is just used to build up a more tractable semi-parametric model (Masyn, 2013).[48] For example, by using an indirect approach, Bauer (2005) aimed to capture nonlinear relationships between latent variables with the aid of the finite mixture structure. In this regard, MoE builds a set of local experts to capture nonlinear or higher-order interaction effects (like the indirect application) and thus MoE approximates a true mapping function.

It is also pertinent to remark on the distinction between the seemingly similar concepts of MoE and *Bayesian model averaging*[49]. As noted by Bishop (2006), MoE is a way of combining models to improve performance. As aforementioned, the key idea of the MoE approach is to have multiple experts that are each responsible for part of the input space. On the other hand, in Bayesian model averaging, models ($h = 1, \dots, H$) are weighted by prior probabilities, $P(h)$, to reflect *uncertainty* on which model is the global model. Simply put, MoE postulates that the dataset is generated by *multiple* data generation processes (represented by local experts), whereas in Bayesian model averaging, the dataset is generated by a *single* model and prior probabilities are applied to reflect model

[48] Thus, the finite mixture model (or the latent class model) is often called a *nonparametric* model if focusing on the ad hoc distribution of model coefficients across segments (e.g. Vij and Krueger, 2017). This implicitly takes the viewpoint of the direct application. On the other hand, the model is often called a *semiparametric* model if taking the viewpoint of indirect application (cf. McLachlan and Peel, 2001; Bauer, 2005).

[49] Bayesian model averaging can be expressed as $f(y) = \sum_{h=1}^{H} P(h) f_h(y|h)$ (adapted from Bishop, 2006). h indexes models and $P(h)$ characterizes the *uncertainty* about which model $f_h(y|h)$ is the *global* model.

uncertainty. For example, Hancock et al. (2020) discussed the use of the "sequential latent class approach" for model averaging. This sequential approach implies estimating individual (pre-specified) models on the same (single) dataset independently in the first stage and averaging them by estimating the weights of each model (through fixing the individual first-stage log-likelihood values) in the second stage. Thus, the equation for this approach is similar to that of MoE, but the philosophy for solving the problem is distinct.

An important question is, which local experts are to be used? In the original study of MoE (Jacobs et al., 1991a), individual local experts were described as neural networks. From the broader perspective on the general model architecture of MoE, any models can be used as local experts (class-specific outcome models in the latent class modeling language). As reviewed by CHAPTER 2, these class-specific outcome models are often formulated with classical econometric models such as logit; in the choice modeling community, the MNL model is the most popular one. In addition, in theory, we can customize and design different types of models for each of the local experts following the *confirmatory* latent class approach (cf. Hess, 2014; CHAPTER 2; CHAPTER 4). In this study, we confine our attention to the logit model for local experts rather than making more complex models, for three reasons. First, the main goal of the study is to discuss the general idea of using MoE as an approximator and extracting the information the model learned; hence the simplicity is helpful for delivering the main idea. Second, as a practical reason, having neural networks as the local experts is parameter-intensive, and thus our dataset with its limited number of cases might not be large enough to learn such complicated models. Lastly, in fact, we expect that having "simple" local experts is not a bad idea,

because it is common to have “weak learners” (even only slightly more accurate than random) when using ensemble models (cf. Hastie et al., 2009; Rokach, 2010).

6.2.2 Neural networks

Our focus is on the MoE in this study, but we also train neural networks as another benchmark, given their popularity and ability of approximation. There are two types of activation functions in neural networks: one in hidden layer(s) and another in the output operation. An activation function serving as an output operation is subject to the type of outcome variable (e.g. logistic for binary, softmax for multinomial, linear for continuous). An activation function in hidden layer(s) serves as a sort of hyperparameter for neural networks; hence analysts generally use the one of the three most popular (nonlinear) activation functions (logistic, hyperbolic tangent, and ReLU) that offers the best performance given the data. In this study, we use the logistic function because of its popularity as well as to achieve greater comparability with the mixture of experts which is the main interest of this study. More detailed equations and descriptions can be found in standard machine learning textbooks (e.g. Bishop, 2006; Hastie et al., 2009).

6.2.3 General approach of the study

We feed input features, i.e. explanatory variables (without specifying nonlinear or interaction effects), to three models:

- **Benchmark (logit) model**: this serves as the baseline, representing the simplest model;
- **Neural network (NN)**: this is an alternative machine learning model, which also (like MoE) has the property of a universal approximator. If NN and MoE present

"identical" results, then we may claim that both models have reached the "ceiling" of possible performance (given the data);

- **Mixture of experts (MoE)**: this model is the main interest of the study.

In the experiments, after training the models, we plot the true choice probabilities and (systematic) utilities versus estimated probabilities/utilities of the three models. Choice probabilities are the ordinary final output of the models, but utilities are not clear for NN and MoE. We convert choice probabilities into systematic utilities by fixing the reference alternative's systematic utility to zero, focusing on the relative utility of the other alternative, and assuming that the true data generation process follows the logit model formulation.[50] Thus, in the binary choice context, the estimated systematic utility of the non-reference alternative is:

$$\widehat{\boldsymbol{V}} = -\log\left(\frac{1}{\hat{P}} - 1\right) = \log\hat{P} - \log\left(1 - \hat{P}\right), \tag{6.3}$$

where $\hat{P}$ is the estimated choice probability of that alternative ($\hat{P} = \frac{1}{1+\exp(-\widehat{\boldsymbol{V}})}$).[51]

After training the models and obtaining choice probabilities and utilities, we explore whether they approximate the true values. To examine the effects of travel time and cost, we examine how the output (choice probabilities or utilities) behaves with respect

[50] For experiments with synthetic data (Section 6.3), we generate the data with logit models (i.e. the true data generation process follows the logit model). For the empirical application (Section 6.4), we do not know the true data generation process. We assume, for convenience, that it follows the logit model. Note that we are not (here) trying to reverse-engineer the *specific equation for V*, only the total value of V. The equation for V could be quite complicated (e.g. with random coefficients, interaction terms, inclusive value terms, etc.), but as long as we can express total utility (difference) as the sum of V and a random component having a logistic distribution – assumptions that are quite standard even for complex discrete choice models – then this is a reasonable approach. Of course, as with any model, the reality could differ from the assumption we make.

[51] Since "only differences in utility matter" (Train, 2009), in effect we are finding the difference in systematic utilities between the non-reference alternative and the reference alternative.

to changes in the variable of interest (while fixing other variables as they are). This is how we obtain *marginal effects* in choice modeling (cf. Hensher et al., 2015), and what is known in machine learning as the *partial dependence plot* (cf. Friedman, 2001; Hastie et al., 2009; Molnar, 2020). For interaction effects, we additionally plot how the slopes of utilities vary by the interacted variables. The following sections have separate purposes:

- Section 6.3 experiments with synthetic data, for which we know the true data generation processes. The main goal of the section is to demonstrate the ability of MoE to detect and approximate nonlinear and interaction effects. If its abilities are verified, then it implies we can use MoE as an exploratory tool to find effects that would otherwise have not been identified.
- Section 6.4 applies the methodology to empirical data. We train and find the best model given the data. Then we discuss how to extract the information learned from those exploratory tools, and devise a general process of informing simpler models of better specifications based on knowledge that can be gained from machine learning.

6.3 Experiments with synthetic data

6.3.1 Experimental setting

This section aims to evaluate the ability of the two approximators, MoE and NN, to identify atypical functional forms by experimenting with synthetic data. In particular, we focus on how well they can recover the true parameters, compared to a simple logit benchmark model that does not consider complex effects. To this end, we generate a training set (N=8,000) and a testing set (N=4,000) of observations simulating a binary mode choice application. First, we draw input features with distributions defined in Table 6-1. By applying the true systematic utility equations (Table 6-2), we obtain choice probabilities for each individual and, based on those probabilities, draw a chosen

alternative in each case. This procedure is analogous to those in Vij and Krueger (2017) and Han et al. (2020). We have four experiments, specifically focusing on two nonlinear effects and two interaction effects involving travel time. These four effects can be considered instances of (systematic as opposed to unobserved) *parameter heterogeneity*, in which parameters vary by region of the input space. For nonlinear effects, the sensitivity of the outcome to a given input is a function only of the size of that single input; for interaction effects, the parameter associated with one input variable is a function of other variables (e.g. the travel time coefficient depends on the individual's income status). Note that, in this study, we assume that the true data generation processes follow random utility maximization theory and the binary logit function. However, it is likely that in the real world, other types of behavioral generation processes can be mixed (e.g. regret minimization, lexicography). Thus, future research may need to explore this avenue.

Table 6-1. Description of synthetic data

Variable	Notation	Data generation process
Cost of car ($)	CO_{car}	$U(1,10)$
Cost of bus ($)	CO_{bus}	$U(1,4)$
Travel time for car (min)	TT_{car}	$U(5,40)$
Travel time for bus (min)	TT_{bus}	$U(TT_{car}, 1.2 * TT_{car})$
Lower-income dummy	INC	$Bern(0.3)$
Attitudinal propensity (continuous) for monochronicity	AT	$N(0,1)$

Table 6-2. True utility equations generated

	Effect	Utility equation
1.1	Nonlinear – polynomial	$V_{bus} - V_{car} = 0.4 - 0.2\ CO - (0.6 + 0.1\ TT)\ TT$ *Example:* The travel time differential has a quadratic effect, where the longer travel time for bus initially decreases the attractiveness of bus (relative to car), but past a certain point increases its relative utility (perhaps due to having a meaningful amount of travel time during which other activities can be conducted).
1.2	Nonlinear – threshold (saturation)	$V_{bus} - V_{car} = 0.4 - 0.2\ CO + (0 - 0.4\ t)TT + 0.8\ t$, where t is a dummy variable ($t = 1$ if $TT > 2$, $t = 0$ otherwise) *Example:* If the travel time differential is minimal, then it does not affect preference among alternatives. However, when the time differential is greater than a certain value, then it would affect her choice behavior.
2.1	Interaction – binary	$V_{bus} - V_{car} = 0.4 - 0.2\ CO + (-0.1 - 0.3\ INC)\ TT + 0.1\ INC$ *Example:* It is known that income groups (or certain demographic groups) have different sensitivity to travel time.
2.2	Interaction – continuous	$V_{bus} - V_{car} = 0.4 - 0.2\ CO + (-0.4 - 0.2\ AT)\ TT + 0.3\ AT$ *Example:* How people perceive travel time or cost is a function of their personalities or attitudes. For example, people who cannot or do not wish to use in-vehicle time productively have greater sensitivity to travel time.

* $CO = CO_{bus} - CO_{car}, TT = TT_{bus} - TT_{car}$

6.3.2 Results

Figure 6-1 and Table 6-3 show how well the three models approximate the true values of the choice probabilities and utilities. Figure 6-1 visualizes how close the estimates are to the true values (black lines). Overall, NN and MoE present good approximations, whereas the benchmark model (logit without specifying nonlinear/interaction effects) deviates substantially from the true values. Table 6-3 quantifies their closeness with R-squared measures for the regression of estimated against true values.

The next question is how well the three models are recovering the true parameters defined in Table 6-2. To this end, parameters are compared across models (Table 6-4).

First, the logit model with the true specification has, overall, the lowest MAPE (i.e. comes closest to the true values). The model is replicating the true parameters at satisfactory levels, but not exactly replicating the true values because of randomness. We also find that NN and MoE well approximate the true parameters (note that their MAPE values are similar to those of the logit model with the true specification), with MoE generally performing a bit better than NN. On the other hand, the benchmark model (logit with the simple specification) gives substantially biased estimates for all experiments. In particular, due to dismissing nonlinear or interaction effects, the relevant parameter (i.e. the travel time difference coefficient) is significantly affected. Along with this analysis, Figure 6-2, Figure 6-3, and Figure 6-4 visually show how well the three models are capturing nonlinear and interaction effects. First, Figure 6-2 illustrates whether the variable of interest has a linear or nonlinear effect on systematic utility. The slopes of the black lines portray the true marginal effects for the experiments (1.1 and 1.2) involving polynomial and threshold nonlinear effects, respectively. When we specify the logit model without such nonlinear effects, the model fails to capture those effects and produces a constant marginal effect of the time differential on systematic utility (see the yellow lines). On the other hand, even if we do not inform them of such nonlinearity, NN and MoE replicate the true nonlinear effects (see the blue and red lines).

For the interaction effects, we refer to Figure 6-3 and Figure 6-4. The left panel of Figure 6-3 plots utilities as a function of the travel time differential, and shows how the slopes vary by the interacting binary dummy (experiment 2.1). The two slopes are the same for the benchmark model, whereas NN and MoE exhibit different slopes based on values of the dummy variable and thus we can confirm that the two models capture the interaction

effect between travel time and the income dummy. The bottom panel in Figure 6-3 shows the same plot for the cost variable; none of the three models present an interaction effect. Thus, this demonstrates that NN and MoE automatically identify which variables should be interacted "under the hood". Figure 6-4 visualizes interaction effects between two continuous variables (experiment 2.2). The benchmark model gives a plane surface of utility as a function of the two variables (due to the linearity), whereas NN and MoE identify the nonlinear surface (saddle shape). Thus, throughout a series of experiments, we demonstrate that MoE (and NN) can capture nonlinear and interaction effects *without* prior information.

Table 6-3. The squared correlation of true with estimated probabilities and utilities

		Probability			Utility		
	Effect	Logit	NN	MoE	Logit	NN	MoE
1.1	Nonlinear – polynomial	0.746	0.993	0.991	0.716	0.993	0.992
1.2	Nonlinear – threshold (saturation)	0.957	0.990	0.991	0.952	0.988	0.989
2.1	Interaction – binary	0.900	0.990	0.997	0.893	0.989	0.996
2.2	Interaction – continuous	0.879	0.989	0.987	0.859	0.988	0.984

Table 6-4. Parameter estimates by model

Experiment	Variable	True value	Logit with the true specification	Benchmark (logit)	NN [a]	MoE [a]
ex 1.1	Intercept	0.400	0.461	-0.106	0.460	0.428
	CO	-0.200	-0.213	-0.207	-0.212	-0.213
	TT	-0.600	-0.659	-0.017	-0.658	-0.628
	TT^2	0.100	0.108	-	0.108	0.103
	MAPE (%) [b]		**9.8**	**81.8**	**9.7**	**5.3**
ex 1.2	Intercept	0.400	0.460	0.752	0.490	0.465
	CO	-0.200	-0.209	-0.207	-0.211	-0.213
	TT	0.000	-0.040	-0.299	-0.076	-0.051
	$TT * t$	-0.400	-0.368	-	-0.327	-0.353
	t	0.800	0.752	-	0.691	0.713
	MAPE (%)		**8.3**	**72.9**	**15.0**	**11.3**
ex 2.1	Intercept	0.400	0.435	0.675	0.452	0.449
	CO	-0.200	-0.203	-0.198	-0.203	-0.203
	TT	-0.100	-0.102	-0.198	-0.108	-0.108
	$TT * INC$	-0.300	-0.317	-	-0.292	-0.306
	INC	0.100	0.101	-0.619	0.055	0.074
	MAPE (%)		**3.8**	**197.4**	**14.0**	**10.0**
ex 2.2	Intercept	0.400	0.428	0.401	0.443	0.412
	CO	-0.200	-0.211	-0.207	-0.205	-0.210
	TT	-0.400	-0.408	-0.380	-0.402	-0.398
	$TT * AT$	-0.200	-0.205	-	-0.175	-0.176
	AT	0.300	0.239	-0.198	0.189	0.180
	MAPE (%)		**7.5**	**55.0**	**12.7**	**12.1**

a. Parameters are estimated by regressing the estimated utilities of Eq. (6.3) on the variables in the table.
b. MAPE (mean absolute percentage error) is calculated based on true values.

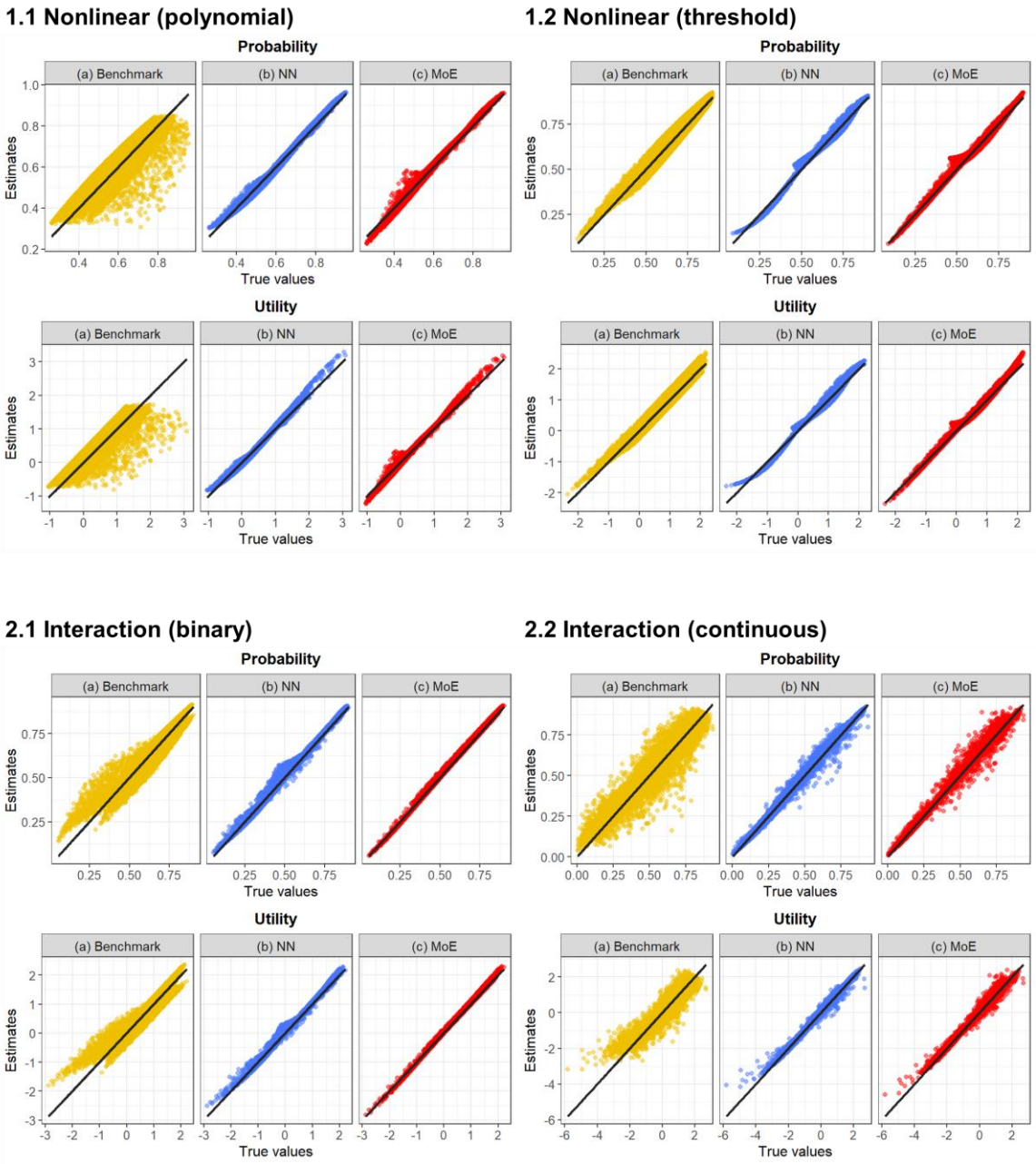


Figure 6-1. Estimated versus true values of probabilities and utilities

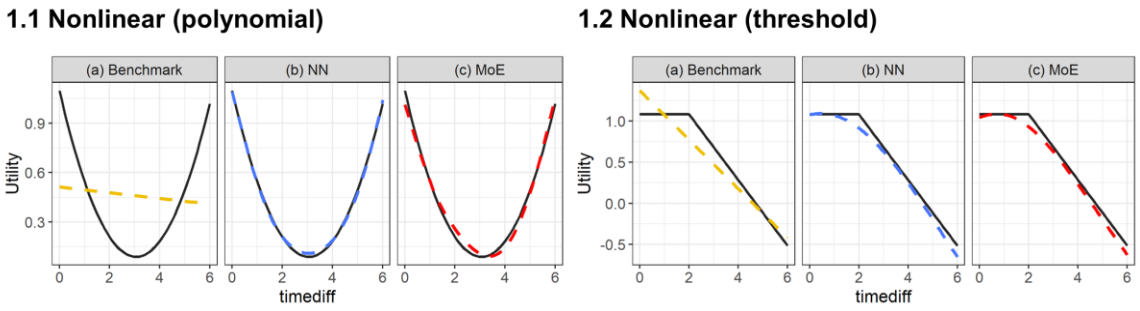


Figure 6-2. Systematic utilities for polynomial/threshold models

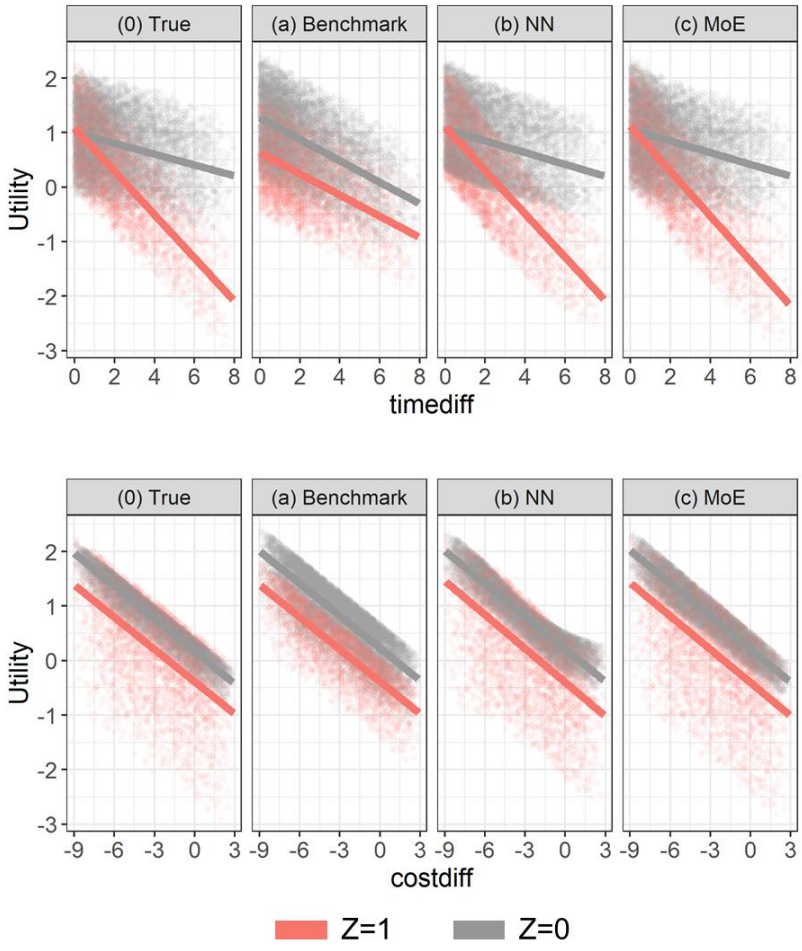


Figure 6-3. Identified interaction effect (with binary dummy, experiment 2.1)

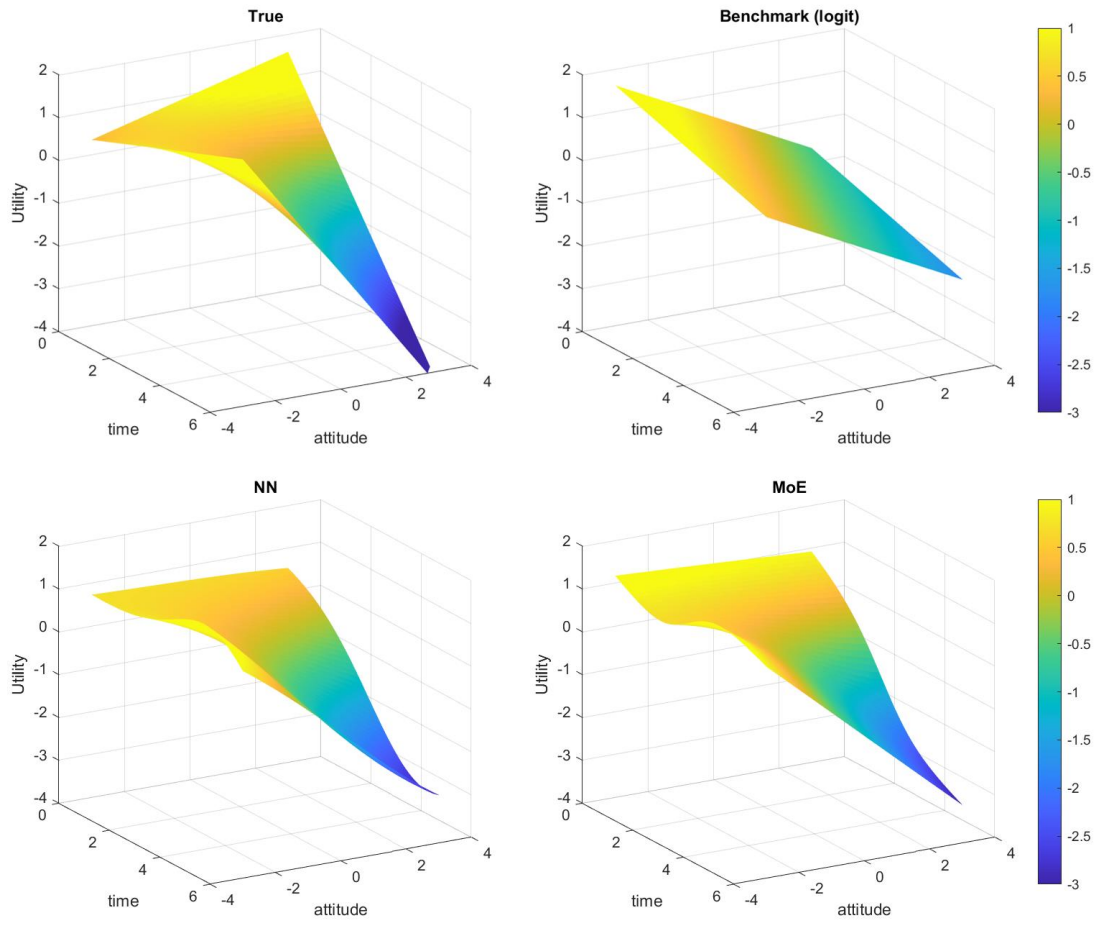


Figure 6-4. Identified interaction effect (with continuous variable, experiment 2.2)

6.4 Empirical application

6.4.1 Data

In this study, we employ the Swissmetro data,[52] which have been widely used in the choice modeling community. The dataset consists of survey data collected on trains between St. Gallen and Geneva, Switzerland. The original dataset contains stated preference (SP) measures over three modes (train, Swissmetro, which is a hypothetical

[52] The data are available at https://biogeme.epfl.ch/data.html.

high-speed rail alternative, and car) for intercity travel. More details about the data are available at the source link (footnote 52). We confine our attention to the binary choice context (choice between conventional train and car, where car is the reference alternative) and analyze 3859 observations. We randomly split the data into a training set (3299, 75%) and a test set (560, 25%).

Table 6-5. Variables used in modeling (N=3,859)

Variable	Coding	Mean
Choice	Choice (1 if train, 0 if car)	0.202
Travel time difference (min)	reference: car	0.413
Travel cost difference (CHF scaled by 100)	reference: car	0.121
Sex dummy	male = 1	0.792
Age dummy	39 or less = 1	0.295
Age dummy	65 or greater = 1	0.100
Lower-income dummy	Under 50k CHF* annually = 1	0.147
Luggage dummy	1+ piece of luggage = 1	0.596
First-class traveler dummy	First class = 1	0.507
Purpose dummy	commute/business = 1	0.578
Train headway dummy	30 min = 1 (60 or 120 min = 0)	0.354

* CHF stands for Swiss Franc (approximately similar to USD)

6.4.2 Training and performance

We estimate three models with the variables described in Table 6-5. To pick the best models, we grid-search some combinations of hyperparameters. The number of hidden nodes (for NN) and classes (for MoE) are explored over [2, 12] and the lambda value (for the L2 regularization term) is explored over [0, 0.01, 0.1], with 5 different seeds for random starting points[53]. We find the best set of hyperparameters using 5-fold cross-validation.

[53] This search space may be narrower than for typical machine learning applications. For example, for NN, we fix a single layer and use logistic activation functions. As well, the numbers of hidden nodes (for NN) and classes (for MoE) are fairly small. One reason for this is that the purpose of the study is simply to illustrate the idea of using MoE as an approximator that finds nonlinear and interaction effects. More importantly, however, due to the (quite small, for machine learning) sample size, it is less beneficial to increase the complexity of the models (i.e. the number of parameters is mainly affected by the number of hidden nodes and classes).

Specifically, for each combination of hyperparameters, the training set is split into five parts and the model is trained after holding out one part; this is repeated while holding out each of the five parts in turn. Then, we average the performance on the five hold-out validation sets for each combination of hyperparameters, change the combination of hyperparameters, and repeat the process. Finally, we select the model with the best average performance across all combinations tested. Based on this approach, the 10-node NN model and 6-class MoE model are selected.

Table 6-6 presents the performance of the three models. MoE and NN show improvement in predictive accuracy over the benchmark logit model. The ρ^2 measure improves substantially, from 0.46 (logit) to 0.57 (MoE and NN). The final log-likelihoods are similar on the test set for MoE and NN, and both better than for the benchmark model. The comparisons for the Akaike and Bayesian information criteria (AIC and BIC), where smaller values are better, are more ambiguous, but overall, by employing MoE and NN, we gained better performance over the benchmark logit model.

Table 6-6. Model performance

Model	Dataset	Accuracy (probability-weighted) *	Log-likelihood	ρ^2	AIC	BIC
Benchmark (logit)	Training set	0.743	-1364.82	0.403	2751.65	2818.76
	Test set	0.762	-208.49	0.463	438.98	486.59
	Pooled	0.746	-1573.31	0.412	3168.63	3237.47
NN	Training set	0.831	-873.40	0.618	1988.81	2727.07
	Test set	0.822	-164.74	0.576	571.48	1095.16
	Pooled	0.829	-1038.14	0.612	2318.28	3075.52
MoE	Training set	0.815	-954.31	0.583	2150.63	2888.89
	Test set	0.816	-165.72	0.573	573.44	1097.12
	Pooled	0.815	-1120.03	0.581	2482.07	3239.30

* The share correctly classified, using probability-weighted predictions.

6.4.3 Identifying nonlinear effects

Our interest is in what the models learned from the data, particularly regarding time and cost effects. To find the nonlinear effects learned from the data, we plot choice probabilities and systematic utilities as a function of time and cost (Figure 6-5), and consider the marginal effects of these variables by examining the slopes of these functions. For both time and cost, as expected, we find that the benchmark model (logit) produces constant marginal effects on utilities (i.e. constant parameters). On the other hand, both MoE and NN capture the nonlinear effects of both variables. If the time difference (between train and car) is less than zero, the slope is gentle or close to zero, whereas the slope is steeper when the time difference lies in [0, 1]. Cost also shows nonlinear effects. The estimated marginal effects are not identical, but both MoE and NN show consistent nonlinear effects of time and cost.

We can approximate these nonlinearities by interacting the main variable with dummy variables (similar to a piecewise regression approach) in an ordinary binary logit model. To see the benefit of capturing nonlinear effects, we compare three models in this section (Table 6-7): the model without nonlinear effects (N1), the model having a two-region piecewise linear effect (N2), and the model having a three-region piecewise linear effect (N3). As we add more nonlinearity dummies, the new parameters play significant roles and the models are substantially improved on all performance metrics. Thus, by exploring the identified marginal effects learned from MoE (or NN), they can inform us regarding how to specify nonlinear effects in a *conventional* logit model.

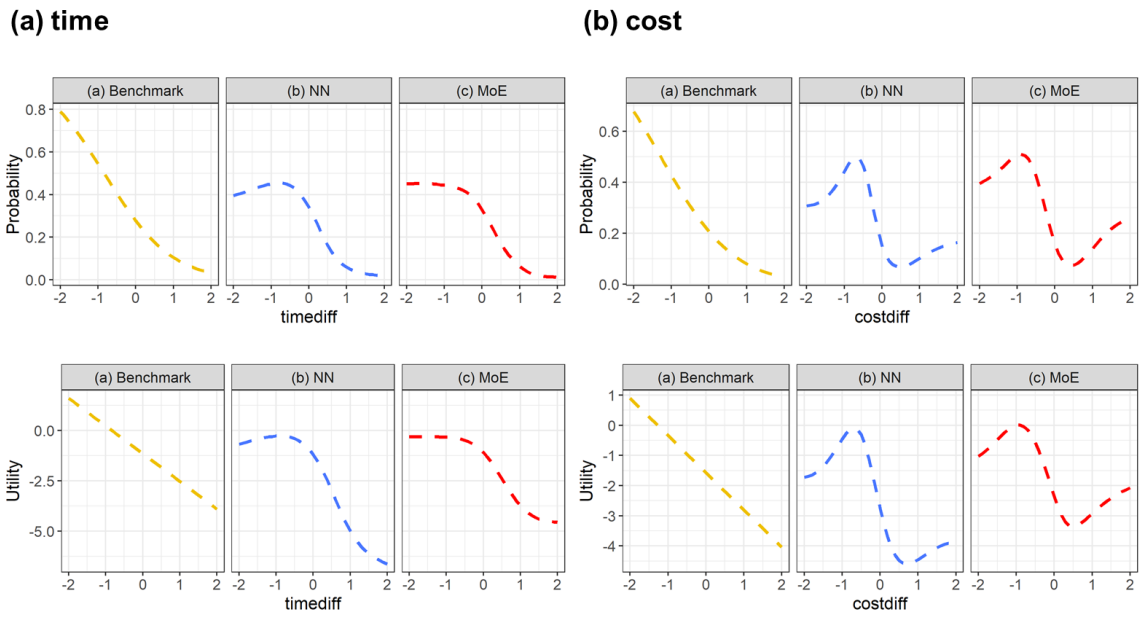


Figure 6-5. Choice probabilities and systematic utilities as functions of (a) time and (b) cost

Table 6-7. Binary logit model results incorporating nonlinear effects learned from MoE

Estimates

	N1		N2		N3	
	Estimate	**t value**	**Estimate**	**t value**	**Estimate**	**t value**
Intercept	-0.896	-17.29	0.383	0.65	0.413	0.69
TT	-1.427	-13.72	-0.073	-1.08	-0.124	-1.78
CO	-1.263	-14.43	-0.231	-0.62	-0.211	-0.57
Dummy_TT (TT>0)	-	-	-0.033	-0.26	0.082	0.61
Dummy_CO_1 (CO<-1)	-	-	-0.795	-1.35	-1.186	-2.00
Dummy_CO_2 (CO>0.5)	-	-	-	-	-2.248	-7.24
TT * Dummy_TT	-	-	-2.572	-12.41	-2.878	-13.04
CO * Dummy_CO_1	-	-	-1.266	-3.27	-3.107	-7.55
CO * Dummy_CO_2	-	-	-	-	4.826	16.58

Performance

	Accuracy	**Log-likelihood**	**ρ^2**	**AIC**	**BIC**
N1	0.728	-1682.595	0.371	3371.190	3389.965
N2	0.740	-1604.068	0.400	3222.136	3265.943
N3	0.764	-1443.377	0.460	2904.755	2961.078

6.4.4 Finding the best specification of the conventional model

By training MoE models, they can learn not only nonlinear effects, but also interaction effects among variables. In theory, we can detect interaction effects by following steps similar to those in Sections 6.3.2 and 6.4.3. However, in real-world contexts numerous combinations of interaction terms can be possible, and testing such combinations individually can be tedious and haphazard. It is of interest to explore whether a somewhat more (though not completely) automated process can produce meaningful and useful results. Accordingly, here we conduct regression modeling to approximate what we learned with MoE. Specifically, we regress the *estimated utilities* (the $\widehat{\boldsymbol{V}}$s of Eq. (6.3)) on the available explanatory variables, including nonlinear and interaction terms, and select the specification that best predicts those utilities and thus best approximates the results of MoE (in a spirit similar to that of stepwise regression, or actually in this instance, all-possible-subsets regression). In this study, we purposefully allow interactions only up to three variables rather than allowing any possible higher-order interactions. This "degeneration" might be seen as unnecessary, but we consider that allowing higher-order interaction severely impedes interpretability and thus we may lose the benefit of using conventional models over just using machine learning models. However, if needed, analysts can allow higher-than-three-level interactions following the same process described in this study.

After finding the best specifications through regression, we estimate choice models (i.e. based on the *observed choices*) with those best specifications. Figure 6-6 plots the ratios of the ρ^2 measure of the MoE result to that of each choice model in turn, thereby illustrating the level of approximation to MoE achieved by the various specifications. The

same specifications and their performance in modeling the observed choices are listed in Table 6-8. We consider that MoE represents the practical "ceiling" on how close we can come to the true data generation process in the sample. Thus, the question is how closely more conventionally-specified models can approximate the data generation process captured by the MoE. Here are several observations from Figure 6-6.

- Overall, specifying nonlinear effects of time/cost substantially improves performance. In particular, their contributions are even greater than those obtained by adding other explanatory variables (M0b vs. M1a).
- Specifying interaction effects improves performance compared to the model without interaction effects. The model with three-variable interactions is better than that with two-variable interactions, but the improvement is not significant (M2 vs. M3).
- Combinations of nonlinear and interaction effects bring the best approximation to the ceiling (M4a vs M5a, M4a vs M5b).
- Having two splits in the piecewise linear effect of cost approximates the ceiling better than having one split (M1a vs. M1b; M4a vs M4b; M5a vs. M5b).

From this application, two key observations emerge. First, as in the experiments with synthetic data (Section 6.3), we can confirm that MoE captures nonlinear and interaction effects without prior knowledge. We can identify nonlinear effects by plotting the estimated choice probabilities and systematic utilities, and important interaction effects by conducting regression on the MoE results. In this empirical context, time interacts with MALE and PURP; cost interacts with MALE and FIRST. The final model is in Table 6-9.

Second, by using these discoveries, we can inform the specifications of conventional (logit) models. Here, however, we face a dilemma regarding the decision on the final model specification. The best model specification with respect to the degree of

approximation to the ceiling (i.e. data-driven) is M5a. In particular, the nonlinear effects of time and cost bring significant improvements. Such a model may, however, contradict conventional theory, because it renders positive coefficients of time and cost for certain segments of people. Hence, analysts should decide which one to pick depending on their purpose. If the analyst pursues better performance, either accepting an inconsistency between the solution and theory or having some theories to explain the positive coefficient of cost, then the final specification might be M5a. However, if the analyst places more weight on theory or if the positive coefficient can pose a problem for decision-making, then the analyst should adopt the best model giving results that are consistent with theory (while sacrificing a bit of performance).

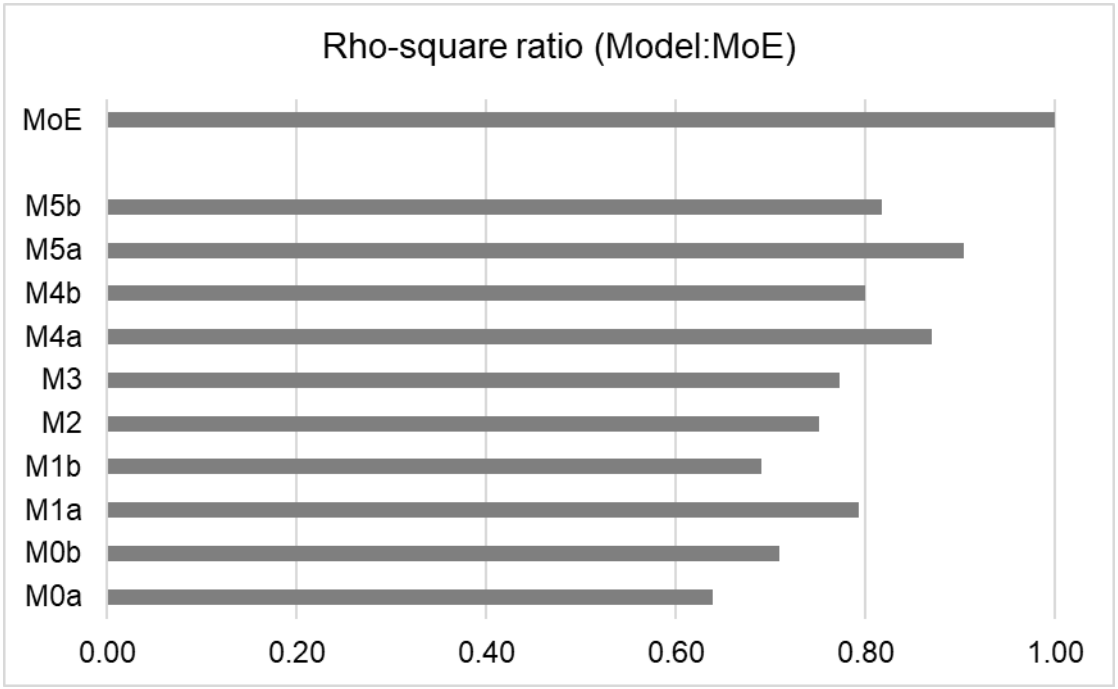


Figure 6-6. Approximation to MoE result by various specifications

Table 6-8. Model performance based on the final specifications

Training set

Model	Time	Cost	Other variables	Accu-racy	LL	ρ^2
M0a	Linear	Linear	NA	0.726	-1454	0.364
M0b	Linear	Linear	Linear	0.743	-1365	0.403
M1a	Nonlinear (1 split)	Nonlinear (1 split)	NA	0.763	-1239	0.458
M1b	Nonlinear (1 split)	Nonlinear (2 splits)	NA	0.738	-1376	0.398
M2	Linear	Linear	Interact with time & cost (1 level)	0.750	-1301	0.431
M3	Linear	Linear	Interact with time & cost (2 levels)	0.757	-1269	0.445
M4a	Nonlinear (1 split)	Nonlinear (2 splits)	Interact with time & cost (1 level)	0.783	-1135	0.504
M4b	Nonlinear (1 split)	Nonlinear (1 split)	Interact with time & cost (1 level)	0.764	-1225	0.464
M5a	Nonlinear (1 split)	Nonlinear (2 splits)	Interact with time & cost (2 levels)	0.793	-1089	0.524
M5b	Nonlinear (1 split)	Nonlinear (1 split)	Interact with time & cost (2 levels)	0.771	-1195	0.478

Test set

Model	Time	Cost	Other variables	Accu-racy	LL	ρ^2
M0a	Linear	Linear	NA	0.742	-228	0.411
M0b	Linear	Linear	Linear	0.762	-208	0.463
M1a	Nonlinear (1 split)	Nonlinear (1 split)	NA	0.767	-205	0.473
M1b	Nonlinear (1 split)	Nonlinear (2 splits)	NA	0.749	-228	0.411
M2	Linear	Linear	Interact with time & cost (1 level)	0.764	-209	0.463
M3	Linear	Linear	Interact with time & cost (2 levels)	0.767	-207	0.466
M4a	Nonlinear (1 split)	Nonlinear (2 splits)	Interact with time & cost (1 level)	0.788	-189	0.514
M4b	Nonlinear (1 split)	Nonlinear (1 split)	Interact with time & cost (1 level)	0.773	-208	0.465
M5a	Nonlinear (1 split)	Nonlinear (2 splits)	Interact with time & cost (2 levels)	0.793	-184	0.526
M5b	Nonlinear (1 split)	Nonlinear (1 split)	Interact with time & cost (2 levels)	0.772	-213	0.452

Table 6-9. Best model results

M5a

	Estimate	t value
(Intercept)	0.418	0.56
TT	-0.843	-1.89
Dummy_TT_1	0.067	0.38
CO	1.527	2.83
Dummy_CO_1	-2.807	-3.91
Dummy_CO_2	-2.251	-5.95
MALE	0.501	2.25
PURP	1.493	6.31
FIRST	1.142	4.15
AGE12	0.265	2.07
AGE5	0.589	3.32
INC1	0.933	6.20
LUG	0.922	7.50
T_HE1	0.611	5.42
TT* Dummy_TT_1	-3.203	-8.83
CO* Dummy_CO_1	-4.610	-8.84
CO* Dummy_CO_2	5.811	15.49
MALE*PURP	-1.771	-6.25
TT*MALE	0.421	0.88
TT*PURP	0.962	2.18
MALE*FIRST	-1.220	-3.97
CO*MALE	-0.522	-1.70
CO*FIRST	-2.374	-5.98
TT*MALE*PURP	-0.254	-0.45
CO*MALE*FIRST	1.870	4.17

6.5 Conclusions

This study examined the possibility of using the mixture of experts (MoE) as an exploratory tool to capture nonlinear and interaction effects (particular types of parameter heterogeneity) when having no prior knowledge of those effects. Firstly, we explained that MoE is connected to the popular latent class model; more specifically, the usage of MoE is connected to the "indirect application" (in psychometric parlance) of finite mixture

modeling. To demonstrate its usefulness, we experimented with synthetic data for which we know the true nonlinear/interaction effects. In the experiments, MoE was able to identify those true effects. In a separate application to empirical data, MoE identified significant nonlinear effects of time and cost on mode choice and captured interaction effects as well. By using the information from the MoE results, we were able to revise the specifications of conventional logit models and thus improved model performance.

There are several avenues for future research. MoE could be extremely parameter-greedy, as a function of the number of experts and the complexity of each expert. As one possibility to reduce complexity, we may feed explanatory variables only to the segmentation model or only to the experts (i.e. the class-specific outcome functions). This can lessen the number of parameters, but whether its performance can be equivalent to that of standard MoE (in which variables are fed to both segmentation and outcome functions) has not been studied. Second, the current study required the analyst's judgment to translate the identified effects into particular model specification elements (e.g. definition of the number and location of piecewise linear splits; specific interaction terms). It might be helpful if we could have a more systematic framework for performing this translation (e.g. algorithmic detections). Third, the analyst's engagement is also required to determine whether "statistically significant" effects are conceptually meaningful/interpretable or not. Some automation of this step could also be helpful, but this avenue could be challenging because analysts would have to codify domain knowledge in a form suitable for machine-based evaluation. This would not be easy, because domain knowledge itself may contain effects that are complex, and there may not be consensus in the domain on the directionality or composition of many effects. At a minimum, some constraints in estimation might be

used to ensure directionality of effects where such a consensus does exist. Lastly, we applied the models to simple empirical data (i.e. a binary decision with small sample sizes), but more empirical studies are needed to verify the usefulness of MoE.

CHAPTER 7. DISCUSSION AND CONCLUSION

7.1 Summary

This thesis aimed to pave the way to improving our behavioral/demand models by taking an in-depth look at the *heterogeneity* in human behavioral processes, and ways of incorporating that heterogeneity into our models. It specifically focused on finite-valued forms of heterogeneity, embracing concepts of *market/data segmentation* and *finite mixture modeling*. The thesis set up the objectives:

1. To build a framework for modeling finite mixture heterogeneity that connects seemingly less related models and various methodological ideas across domains;
2. To tackle various heterogeneity-related research questions in travel behavior and thus show the empirical usefulness of the models under the framework;
3. To examine the potential, challenges, and implications of the framework with conceptual considerations and practical applications.

For these purposes, five inter-related studies were conducted on this journey.

CHAPTER 1 and CHAPTER 2 started with discussions about the necessity of studying heterogeneity, related key concepts, and an overview of modeling finite mixture heterogeneity. Through a comprehensive and systematic review, the study (1) provided a broader understanding of the usage landscape of finite mixture modeling, (2) shed light on various typologies related to methodological approaches to treat heterogeneity, and (3) discussed alternative model configurations. Transportation researchers may benefit from this study by understanding the general idea of finite mixture heterogeneity and where we are now in this modeling. As well, analysts can use this study as a compass while designing their models.

CHAPTER 3 discussed parameter heterogeneity, which is the most popular type of heterogeneity. Specifically, the chapter connected three alternative approaches to treating finite-valued parameter heterogeneity: deterministic segmentation, endogenous switching, and latent class models. The study (1) expanded the typology of mixture modeling by embracing "observed classes" and (2) connected the finite mixture model with the switching model family by way of detailed discussions about their similarities and differences from conceptual and empirical standpoints. Specifically, with equation-rich discussions the study pointed out the distinctive usefulness of each approach: the often-better performance of the latent class model over competing models, and the proper framework for estimating treatment effects offered by the endogenous switching model (including an in-depth interpretation of treatment effects). Analysts may benefit from this study by understanding the connections between two modeling families (thus supporting model selection appropriate to satisfying their ends) and obtaining the correct equations for calculating treatment effects, especially when the dependent variable is log-transformed.

CHAPTER 4 dealt with the confirmatory latent class approach, which has been less discussed in the literature. The study illustrated the usefulness of the confirmatory latent class approach with an empirical application (modeling leisure trip frequencies by car and air). Specifically, the zero-inflated model is embraced under the finite mixture heterogeneity framework, given the expanded typology of heterogeneity. Analysts may gain inspiration from this study on how to operationalize behavioral models when dealing with data showing a particular pattern and when having some behavioral hypotheses on such a pattern.

CHAPTER 5 expanded the latent class model by combining it with the endogenous switching model. It relaxed the latent class model's implicit assumption of independence between the unobserved influences on class membership and outcome. With two empirical applications (modeling the willingness to share autonomous vehicle rides with strangers and the adoption of ridehailing for social-purpose trips), the study showed how the proposed models may give different insights compared to standard latent class models, even when parameter estimates and goodness-of-fit measures appear to be similar. Specifically, when conducting scenario analysis, the proposed method provides distinct marginal and conditional (on class) expectations, whereas the standard model only focuses on conditional expectations. The study opens the door to an avenue for evaluating "treatment effects" in the latent class modeling context, which analysts may wish to pursue in the future.

CHAPTER 6 conceptually connected latent class modeling to the mixture of experts (MoE) approach arising from the machine learning domain. This study used MoE as a data-driven exploratory tool to identify nonlinear and interaction effects (which are special types of parameter heterogeneity) and used what we learn from MoE to improve the performance of conventional models. Through experiments with synthetic data and an empirical application (to mode choice), the study showed that MoE can automatically detect nonlinear/interaction effects and can be used to inform our model specifications. To our knowledge, this study is the first in the transportation domain to use the "indirect application" (as it is known in the psychometrics field) of latent class modeling. Hence, the study expands the usage of finite mixture structures and thus helps to diversify applications for analysts.

7.2 Challenges

Although the preceding five studies showed promising aspects of the finite mixture framework, it is not free from challenges as well. Thus, this section describes several challenges, which have been little discussed in the literature, that analysts need to contemplate.

7.2.1 Sample representativeness

One of the potential issues of the market segmentation or finite mixture approach might stem from the sample: is the sample representative of the target population? This question would be embedded in any study using empirical data and it is an important issue. However, the issue could be even more crucial for studies involving segmentation. Numerous studies conduct modeling to identify structural relationships among variables, or the effects of key variables on a target outcome. In such cases, sample bias may not necessarily be critical, in that the studies focus on conditional relationships (Babbie, 2012): *given* these characteristics $\boldsymbol{X}$, what is the expected outcome Y? However, even aside from estimator bias, if we apply finite mixture modeling to sample data that are not representative of the target population, then projecting the results onto the target population could yield inappropriate answers. In particular, the shares of classes and compositions of classes may not represent those in the population of interest. For example, if the sample is highly skewed toward wealthy people, then the shares of classes related to such demographics could be inflated. Even worse, analysts may fail to identify substantive latent classes that exist in the real population. Another possibility is misleading interpretations. For example, suppose tech-savviness measures are used to identify latent classes from each of two samples: one of the general population and one of "tablet PC users". Although each

sample may produce a “tech-savvy class” (compared to the other classes in the sample), would “tech-savviness” have the same meaning in each sample? Probably not, because mixture modeling would identify *relative* classes within the sample. This issue also makes it challenging to compare latent classes across different studies.

In our review, relatively few studies reported or discussed this issue (or how they handled it). For instance, several studies reported that the sample was fairly representative of the target population with respect to some key demographics (e.g. Srinivasan et al., 2009; Nayum et al., 2013; Fu and Juan, 2017; Mouter et al., 2017; Rahmani and Loureiro, 2019; Saxena et al., 2019b; Gong et al., 2020). Or some studies acknowledged that the sample may not be fully representative of the target population, such as the whole cycling population of Santiago (Rossetti et al., 2019), or the German population (Hackbarth and Madlender, 2016). Several studies applied sample weights to make the results more representative (e.g. Bailey and Axsen, 2015; Prato et al., 2017; Vij et al., 2017). However, this issue may not be pertinent in some studies, if the data represent “all” populations of interest or the data are “big” enough. Alternatively, remedies such as sample weighting may be adequate to address the issue, although that could be debated (since we cannot weight with respect to unobserved characteristics). Either way, however, representativeness should certainly be considered when applying mixture modeling to a sample and then making inferences for the population.

7.2.2 Overfitting and generalizability

Overfitting issues are a concern in any model estimation process, but they can be obscured when it comes to latent class modeling. First, this is because latent class

modeling, by nature, simultaneously involves both "unsupervised learning" and "supervised learning". In latent class choice models or latent class linear regression problems, the problems *per se* are supervised learning problems (outcomes are observed, and the model is oriented toward predicting those outcomes as well as possible), but clustering is embedded in the model (and cluster membership is not predetermined). In unsupervised learning problems such as cluster analysis, it is not easy to say whether the solution is overfitting or not, because the purpose of the analysis is to find latent structure in the data and there is no ground truth against which to test whether the structure is "right" or not.

In addition, an important rationale behind using mixture models or segmentation is to find heterogeneity. If the model identifies some substantial types of heterogeneity, then it is interesting news for analysts who look for heterogeneity. However, the million-dollar question is, *are such differences an indication of true heterogeneity, or a consequence of overfitting*? The more complicated the models are (e.g. having more classes or more parameters), the more likely it is that both true heterogeneity and overfitting are confounded.

One might argue that we can determine whether the solution is overfitted or not by checking whether it is generalizable to other datasets (e.g. using a holdout sample, or cross-validation). However, this question is also tricky to answer. That is because *it is unclear how to properly evaluate model performance correctly* for latent class modeling. If we knew the true class membership in the test sample and if we knew how to properly "guess" class membership using the model, then we could compare those two things and assess the performance of the model adequately. However, by nature, classes are *latent*,

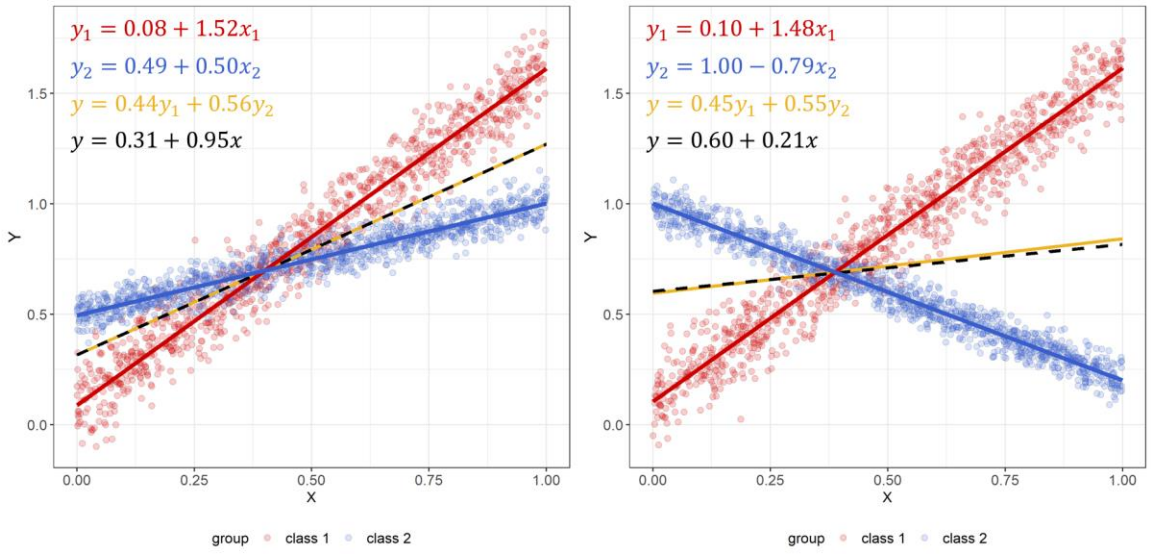
and thus neither of those conditions are satisfied. Hence, although we *interpret* latent class models using their *class-specific* (conditional) outcome functions (which is a key virtue of such models), for *evaluating model fit* we end up relying on the *expected* (marginal) outcomes obtained by probability-weighting the class-specific predictions (since we do not know which conditional, or class-specific, prediction applies to any particular case). This implies that there is an inconsistency between how we interpret the models and how we evaluate the models, and this inconsistency leads us to a systematic undervaluation of latent class models, because the true value of latent class models lies in how well they explain the behavioral processes of each class, not some average outcome.

The two (extreme) examples shown in Figure 7-1 illustrate how an accuracy test can undervalue the latent class model. We generate two Y-X relationships[54] (red and blue; the left panel shows a case where the two relationships are positive, but with impacts of different magnitudes; the right panel exhibits a case where the two relationships reflect impacts of different signs as well as magnitudes). By applying a simple linear regression on the pooled data (black dashed lines), the results deviate from the true data generation processes; this becomes even more severe when the two data generation processes are notably different (e.g. different signs of parameters; see case 2). On the other hand, the latent class model (with two classes) recovered the true parameters for both classes (i.e. the two data generation processes are recovered). The problem is that when obtaining new data points, we are not sure whether a given point is generated from Class 1 or Class 2. Hence, we obtain the marginal prediction (the yellow lines; the weights in the linear combination

[54] Here, for simplicity of discussion, the examples are very extreme cases and are presented as simple regression problems.

of the two class-specific models are the estimated class membership probabilities), but in so doing we lose the original idea of latent class modeling and prediction becomes poor as well.

An obvious (and commonly exercised) option is to assign individuals to classes based on their estimated class probabilities, specifically assigning a case to the class with the highest predicted membership probability for that case. There are three issues here. First, misclassification of at least some cases is almost inevitable. Second, it is common to use posterior probabilities (i.e. considering the information provided by the outcome and thus updating the class membership probability using Bayes' Rule) when assigning in-sample individuals into classes (cf. NLOGIT 2016; Vermunt and Magidson, 2016). However, this is arguably "cheating" in that we are using information about the outcome, which is supposed to be unknown while predicting (Kim and Mokhtarian, 2018). In addition, in a real prediction exercise, the outcome values are likely to be unknown (which is why we are modeling/predicting in the first place) and thus we may not be able to obtain posterior probabilities in a holdout or entirely new sample. Third, the meaning of class probabilities is somewhat misused in this case. As Train (2009) and Hensher et al. (2015) point out, such a "hard classification" based on the highest-probability class (the so-called unit-weighted method) violates the basic idea of a probabilistic choice model. For example, if there were 100 people, each of whom had a 0.51 probability of belonging to Class 1 and 0.49 for Class 2, it would be unrealistic to assign all of them to Class 1 (as the unit-weighted method would do) – instead, we would expect only about 51 of them to belong to Class 1.



	Case 1	Case 2
Pooled model	$y = 0.31 + 0.95x$	$y = 0.60 + 0.21x$
Class 1 model in the mixture	$y_1 = 0.08 + 1.52x_1$	$y_1 = 0.10 + 1.48x_1$
Class 2 model in the mixture	$y_2 = 0.49 + 0.50x_2$	$y_2 = 1.00 - 0.79x_2$
Probability-weighted marginal	$y = 0.44y_1 + 0.56y_2$	$y = 0.45y_1 + 0.55y_2$

Figure 7-1. Illustration of how evaluation of finite mixture models can be misleading

7.2.3 The "Rashomon effect"

Given the flexibility offered by the concept of finite mixture heterogeneity, as highlighted in this thesis, we can use that approach to model numerous types of heterogeneity problems. However, this flexibility may possibly turn into a pitfall since it can raise the question, "when the finite mixture model performs better than competing ones, how can we be sure that the model captured the effects intended by the analyst?" In other words, couldn't there be multiple ways of interpreting the model or formulating the model to fit the data? In machine learning, Breiman (2001) used the term "*Rashomon*

effect"[55] to describe situations in which a multitude of different models can explain the data with about the same performance. A similar phenomenon has been discussed in psychometrics, the so-called "*equivalent models*". When building structural equations models, there could be equivalent models that provide the same general statistical fit indexes, but they may imply very different interpretations of the data (cf. MacCallum et al., 1993; Raykov and Marcoulides, 2001). As well, in the transportation and choice modeling community, as noted in Section 2.3.1, Hess et al. (2013b) and Hensher et al. (2013) pointed out that attribute non-attendance (a type of heterogeneity) and regular taste heterogeneity could be confounded while employing the mixture modeling framework.

The situation becomes even more complex under the expanded typology and framework discussed throughout this thesis. There are at least four aspects that can possibly introduce a Rashomon effect.

1. **Typology of heterogeneity**: There are various types of heterogeneity, but possibly they can be confounded. The findings in Hess et al. (2013b) and Hensher et al. (2013) belong to this category.
2. **Confirmatory latent class approach**: Since analysts can design finite mixture structures depending on their needs, there could be an almost infinite number of possibilities.
3. **Direct vs. indirect applications**: Two usages of finite mixture modeling have been identified in the psychometric literature. In so-called direct applications, we interpret the class-specific behavioral models. On the other hand, we can also use the method to capture overall nonlinear/interaction effects of explanatory variables, neglecting the class-specific functions – a so-called indirect application.

[55] *Rashomon* is a Japanese film (1950), which is known for a plot device that involves various characters providing subjective, alternative, self-serving, and contradictory versions of the same incident. The word Rashomon, afterward, has been used to describe a situation in which multiple different descriptions exist about the same incident.

Then, given a particular empirical context, should a certain model *necessarily* be interpreted as if it is a direct application (or alternatively an indirect application)?

4. **True heterogeneity vs. overfitting**: As described in Section 7.2.2, the more complex models become, the more likely it is that true heterogeneity and overfitting can be intertwined.

Even worse, these four situations can themselves be confounded in a given empirical application. Thus, future research needs to explore the Rashomon effect, and seek possible remedies to alleviate it.

7.2.4 Revisiting the usefulness of finite mixture modeling

Finite mixture modeling is well suited to explore various types of heterogeneity (Section 2.3.1), and it has been verified in numerous empirical studies that the method gives behavioral insights and better performance over other competing models (Section 2.3.7). However, it is often questioned how *useful* the finite mixture model really is, and what would happen if we dismiss it. One may argue that a population-level perspective, which melds differences into an aggregate "average", could be sufficient for planning purposes.

We still contend that uncovering population segments will render both more correct and more useful assessments of infrastructure plans or policy instruments. First, identifying the population average does not mean that everyone exhibits such behavior. Consider a simple situation in which we know that there are two groups of people, respectively willing to pay fees of $6 and $20 for toll lane access. A homogeneous model might give us a willingness-to-pay estimate of around $13, but in fact no one has a willingness to pay near $13. Hence, planning based on population means can be misleading. Second, a key is that

such segments are not randomly distributed, particularly geographically. For example, "urbanite" people are more likely to live in urban areas than in suburban/rural areas and they would be more transit-friendly than "non-urbanite" people. Hence, an understanding of latent population segments (via finite mixture modeling in this study), and particularly how they are distributed, could be useful for the assessment of the effectiveness of planning/policy instruments.

Obviously, we are subject to falling prey to *confirmation bias*, since we typically hypothesize the existence of heterogeneity and mixture models seem to "look better" than the simpler models. Almost no one would reject the model if it is what the analyst wanted/ expected and it exhibits satisfactory goodness of fit. Then, what should we do to truly corroborate the usefulness of the finite mixture model? The first step might be comparisons with simpler models. As covered in Section 2.3.7, comparisons with some baseline models have been reported in many studies, but still, a non-trivial fraction of studies skipped this process. Having such baseline models could be considered redundant in that more complicated models will almost always outperform the simpler models. However, it can still be useful to see *how much* worse the simpler models are (or how well the mixture models actually work).

Furthermore, even when studies compared models, such comparisons were often limited to simple statistical goodness of fits (e.g. particularly in-sample information criteria; cf. Parady et al., 2021). Ideally, future studies should be able to *validate*[56] the models and show how our *estimates*/*predictions* (e.g. willingness to pay, modal shift,

[56] However, as alluded to in Section 7.2.2, we need more discussion on "how to validate".

vehicle-miles reduction) could change by comparing the predictions that do and do not account for heterogeneity in the population. Scenario or sensitivity analysis is one of the possible avenues. For example, what outcomes will each segment produce in response to key factors or policy variables (e.g. Vij et al., 2013; Bailey and Axsen, 2015; Seelhorst and Liu, 2015; Lin et al., 2017; El Zarwi et al., 2017; Kim and Mokhtarian, 2018; Kormos et al., 2019)? The real value of mixture modeling may come from some additional analyses that would not have been possible for simpler models.

Lastly, transportation modelers may need to think about *how to use "latent class solutions"*. In many academic papers, analysts have enjoyed using latent class models and those studies have reported numerous interesting findings. However, we seem to be missing deeper discussions about how to use latent classes in subsequent studies or how to exploit them for actual (demand) forecasting practice. This could be because (1) we might be less confident to use latent classes because they are "latent" and thus too intangible compared to other deterministic segmentation indicators (e.g. gender, age, income), and (2) it is unclear whether latent classes are temporally and spatially stable and transferable. These issues remain unanswered.

7.3 What's next?

Pursuing better behavior/demand modeling (specifically, in this thesis, accounting for finite mixture heterogeneity), we've come a long way, and have a long way to go[57]. Then, what's next? Obviously, the challenges listed in Section 7.2 should be addressed or at least formally discussed in future research. Aside from responses to those challenges,

[57] This was the title for a panel discussion, lectern session 1087, which discussed advances in travel behavior research (moderated by the AEP30 committee; 100th TRB Annual Meeting, Jan. 25, 2021).

due to the versatility and flexibility of the finite mixture approach, there could be a variety of avenues to enrich our behavior/demand modeling. Considering recent developments, here we briefly list some possibilities.

7.3.1 Combining continuous and discrete mixtures

There has been a growing interest in combining different types of heterogeneity in the model. In particular, combining continuous and discrete natures of heterogeneity becomes of interest. Two distinct approaches are possible: *(1) latent class random parameters – random parameters considered for the class-specific outcome models* and *(2) random parameter or latent variable models where the random parameters (or latent variables) follow finite mixtures of (generally continuous) distributions.* The former can be considered as adopting continuous heterogeneity within each segment. A major rationale is to relax the homogeneity assumption *within* a class of a latent class model, by introducing random parameters. In other words, unlike the usual latent class model that focuses solely on *inter-class* heterogeneity, this approach assumes there is also *intra-class* heterogeneity. Boeri et al. (2014) found, in the context of choice of traffic calming scheme, that the outcome models for the two latent classes each had significant standard deviation parameters, confirming taste heterogeneity within each class. This is also supported by the goodness of fit measures compared to those for the standard latent class model. Haghani and Sarvi (2016) modeled pedestrian exit choice behaviors with standard latent class and random parameter latent class models. In this application, both models are significantly better than the ordinary logit model, but both models also exhibited almost identical results, indicating that within-class heterogeneity was not significant given the data. Orvin and Fatmi (2020) examined the destination choice behavior of users of a dockless bike sharing

service with a random parameter latent-segmentation logit model; they found heterogeneity in the means of the random parameter distributions across two classes. This modeling approach has also been recently gaining popularity in safety analysis (cf. Mannering et al., 2016; Li et al., 2018; Yu et al. 2019a).

The second approach constitutes a type of random parameter model where the parameters are assumed to follow a certain mixture distribution (e.g. Gaussian mixture). Random parameters are generally assumed to follow a particular parametric distribution (most often the normal distribution, but others are possible such as log-normal, triangular, and uniform). Adopting mixture distributions, which can approximate any arbitrary continuous distribution in theory, the distribution of a parameter can be more flexible (e.g. allowing asymmetry, multimodality). As one example, Buddhavarapu et al. (2016) modeled a random parameter negative binomial model with a finite mixture multivariate normal structure on the random parameters for crash count data. Alternatively, the finite mixture can be used to specify distributions of latent variables in an ICLV model. In other words, although a latent variable is often specified as having a parametric distribution (e.g. normal, lognormal), allowing a finite mixture of such distributions offers more flexibility in the shape of the overall distribution, while retaining the advantages of parametric distributions. For example, Brey and Walker (2011) modeled flight choice with a latent temporal preference that follows a mixture of normals.

An approach that is related in the sense that it is another way to approximate complex parameter distributions, albeit not specifically combining continuous and discrete heterogeneity, is to introduce more structured support of a distribution, such as a grid (i.e. “casting a net in the coefficient space”, Vij and Krueger, 2017, p. 81), as opposed to the

parameters having unstructured distributions in the parameter space as is the case in conventional latent class models. With this method, it is possible to have a large number of mass points (i.e. classes) while retaining computational tractability. This could be a useful approach not only for computational ease but also for better explaining heterogeneity. With sufficient mass points, the discrete distribution can approximate any distribution to a high degree of accuracy (Heckman and Singer, 1984). However, as we discussed in Section 2.3.6, the number of classes is confined to a limited number in practice due to interpretability as well as estimability, and "finite mixture models with a smaller number of mass points may inadequately capture the full extent of heterogeneity in the data" (Allenby and Rossi, 1999). Dong and Koppelman (2014) first proposed a mixed logit with a discrete distribution with finite support, referring to it as the discrete mixed logit model (DMXL), and compared it with the latent class model, which is a special case of a DMXL. Vij and Krueger (2017) proposed a more flexible modeling framework using the EM algorithm and, with applications to mode choice modeling, showed its usefulness in interpretation as well as predictive ability.

7.3.2 Latent variable sub-models

Many studies have acknowledged the role of some latent constructs (e.g. attitudes) to explain target outcomes (e.g. behaviors). Hence, incorporating such constructs has been a key interest. This has been done in three different ways. First, a few studies have used *raw attitudinal statements* as variables (e.g. Beck et al., 2014; Bailey and Axsen, 2015; Hackbarth and Madlender, 2016; Ferguson et al., 2018). Second, a sizable number of studies have used a two-step approach to incorporate attitudes or other type of latent construct. Specifically, they have *estimated latent factor scores first* (often via exploratory

factor analysis) and then used the estimated scores in latent class modeling (e.g. Olaru et al., 2011; Araghi et al., 2016; Molesworth and Koo, 2016; Molin et al., 2016). This two-step approach avoids model complexity at the expense of losing the benefits of simultaneous estimation (e.g. incorporating measurement errors, estimation efficiency).

A third, more conceptually rigorous, approach is to embed a latent variable model into the finite mixture modeling system. For example, constructing a membership model as a function of a (latent) attitude, Hess et al. (2013) incorporated a measurement model into the membership model. Motoaki and Daziano (2015) constructed a membership function containing a latent variable of bicyclist status by adopting a multiple indicator and multiple causes (MIMIC) structure. Pan et al. (2019) estimated latent class models with a risk aversion latent variable to understand expected electric vehicle charging behavior. Alizadeh et al. (2019) modeled route choice behavior with a latent class integrated choice and latent variable (LC-ICLV) model. By adding two latent variables (consciousness and cautiousness), the model fitted better than a benchmark standard latent class model. Krueger et al. (2018) suggested a framework incorporating latent variable (mode-specific and ecological normative beliefs) and latent class (modality styles) sub-models. A more intuitive approach to understanding this framework is to consider it as a latent class cluster analysis where the class membership is a function of latent variables (having multiple indicators). This is analogous to hybrid choice models where utilities are a function of latent variables.

Another avenue of incorporating latent variable models was proposed in Hurtubia et al. (2014). Specifically, they added class-specific measurement models (with the indicators being ordinal psychometric indicators). By doing so, in two empirical

applications, they found different estimates for the class-membership model and gained behavioral insights with the additional measurement models.

7.3.3 Marriage with other "machine learning" architectures

So far, the latent class models discussed in this review (CHAPTER 2) have been operationalized with "conventional" statistical models. As reviewed in Section 2.3.4, the membership model is almost exclusively specified as a softmax function (multinomial logit) and the outcome models have been represented with well-known classical models (e.g. linear regression, logit, Poisson, etc.; see Section 2.3.5). However, we may view finite mixture modeling as a general model architecture and expand it by marrying it with other machine learning architectures. Specifically, we can embed other architectures (e.g. neural networks) in the mixture modeling framework by replacing membership and/or outcome models with new methods. In fact, this idea is analogous to the *mixture of experts* proposed by Jacobs et al. (1991a) in neural computation (CHAPTER 6 presented a preliminary analysis of this avenue). In the original proposal, the class-specific outcome functions were neural networks. In this case, given that neural networks can automatically capture nonlinear and higher-order interactions, individual class-specific functions may be able to capture complex relationships among variables, or *intra-class* heterogeneity, that would have not been identified by simpler functions (without explicit specifications). In theory, class-specific functions can take any form if one is able to specify the objective function and properly optimize it. Considering the confirmatory approach described in Section 2.3.2, class-specific outcome functions do not need to be the same as well (e.g. class 1 follows nested logit and class 2 follows neural networks, etc.). In this case, of course, the

rationale of such specifications and how to interpret them are the responsibility of the analyst.

The same logic can be applied to the membership model as well, if the membership probability can satisfy conditions of $0 \leq \pi_z \leq 1$ and $\sum_{z=1}^{Z} \pi_z = 1$. For example, Ishaq et al. (2014) used a fuzzy c-means method to construct membership probabilities. Han (2019) represented the class membership model with a feed-forward neural network. The rationale behind this effort is that "In contrast to choice models [outcome models], where we more clearly know about the trade-offs between attributes, the class membership model specification is less clear to define" (Han, 2019, p. 86). By representing the class membership model with a neural network, nonlinear effects can be automatically captured in the membership model. In a nutshell, the mixture modeling framework can be generalized to embrace other methods in the architecture, with the premise of achieving better performance and gaining more information on heterogeneity (at the expense of model complexity, corresponding computational cost, and the challenge of interpretability).

7.4 Outlook and concluding remarks

Throughout the thesis, we explored the possibility of *teaching "older" models "newer" tricks*[58], finding clues for the tricks from the ideas of heterogeneity, finite segmentation, and mixture modeling architecture.

[58] This borrowed the name of the University Transportation Center TOMNET, Center for Teaching Old Models New Tricks, which funded this thesis.

We believe that a real value of modeling finite mixture heterogeneity is its capability and flexibility of allowing a variety of research questions to be tested that are of particular interest to analysts.[59] The methodology has been widely used, but we expect it will gain more attention. For example, as new mobility technologies and services (e.g. shared mobility, autonomous vehicles, e-commerce) keep emerging, human responses to such technologies will be diverging. Hence, the necessity for models to consider unobservable influences on behavior will keep growing. In addition, using "big data" will increasingly require analysts to unearth and model the seemingly undetectable heterogeneity in the data. From this perspective, finite mixture modeling could serve as an appealing approach. Lastly, as we have observed, advances in computational power will allow more complicated models (and thus more sophisticated and comprehensive research questions). Thus, we conjecture that modeling finite mixture heterogeneity will be shouldering a crucial role in transportation studies, and expect that various methodical developments will enrich the analyses and thus bring more insights that would not have been uncovered with simpler models. As a final note, we believe that the true value in employing the finite mixture heterogeneity framework lies not just in permitting the estimation of slightly more complicated statistical models; rather, it is to broaden the horizons of research questions that can be built upon a solid statistical framework. Hence, the thoughtful conceptualization of research questions constructed upon a theoretical

[59] An analogy to the flexibility offered by the finite mixture heterogeneity approach is offered by a documentary on Lego blocks ("A Lego Brickumentary", 2014). For only six bricks (that have the same color and size, each with eight studs), there are 950,103,765 options for putting them together. Mathematics professor Søren Eilers comments, "By mathematical definition, this is a finite system. We have a finite number of bricks. They have a finite number of studs and holes. But for all practical human purposes, these bricks are infinitely flexible and not only that, they define a mathematical problem of infinite complexity. So I would say that, 'Yes, it is finite but in a way it's also infinite.' " This is the source of creativity of Lego designs.

foundation is essential for maximizing the usefulness of models involving finite mixture heterogeneity.

APPENDIX A. TECHNICAL DETAILS ABOUT TREATMENT EFFECTS (CHAPTER 3)

This appendix reports technical details about treatment effects, which were dealt with in CHAPTER 3. Although the treatment effect is not the main focus of the chapter, this appendix explains why we have such a large average treatment effect (ATE) in Section 3.5.1. We also include some notes related to using the sample selection model with a log-transformed dependent variable, which we hope will be useful to the transportation research community.

The heart of the sample selection idea is *selection on unobservables*, as opposed to selection on observables (Cameron and Trivedi, 2005; Greene, 2012). In other words, the continuous latent variable z^* related to selection in Eq. (3.10) is not observed; rather we observe the discrete state defined by its sign, which we denote with the binary variable z (here, the indicator of urban residence). If z^* is associated with outcome generation (here, of ln(VMD)) through correlation of their unobserved influences, Eq. (3.15), then conditional expectations are the sum of the unconditional expectation and the selection effect.

The unconditional (on selection) expectations are:

$$E(Y_{i1}|\boldsymbol{X}_i) = \boldsymbol{X}_i\boldsymbol{\beta}_1 \text{ and} \tag{A1}$$

$$E(Y_{i0}|\boldsymbol{X}_i) = \boldsymbol{X}_i\boldsymbol{\beta}_0 \, . \tag{A2}$$

Eq. (A1) represents what the expected value of Y (i.e. ln(VMD)) would be for a *randomly-selected* person with characteristics $\boldsymbol{X}_i$ if she were to live in an urban area, and Eq. (A2) represents what the expected value would be for a *randomly-selected* person with those observed characteristics if she were to live in a non-urban area.

The conditional (on selection) expectations are (as in Eqs. (3.16) and (3.17)):

$$\begin{aligned} E(Y_{i1}|z_i^* > 0, \boldsymbol{X}_i, \boldsymbol{W}_i) &= \boldsymbol{X}_i\boldsymbol{\beta}_1 + E[\varepsilon_{i1}|z_i^* > 0, \boldsymbol{X}_i, \boldsymbol{W}_i] \\ &= \boldsymbol{X}_i\boldsymbol{\beta}_1 + \rho_1\sigma_1 \frac{\phi(\boldsymbol{W}_i\boldsymbol{\alpha})}{\Phi(\boldsymbol{W}_i\boldsymbol{\alpha})} \text{ and} \end{aligned} \tag{A3}$$

$$\begin{aligned} E(Y_{i0}|z_i^* < 0, \boldsymbol{X}_i, \boldsymbol{W}_i) &= \boldsymbol{X}_i\boldsymbol{\beta}_0 + E[\varepsilon_{i0}|z_i^* < 0, \boldsymbol{X}_i, \boldsymbol{W}_i] \\ &= \boldsymbol{X}_i\boldsymbol{\beta}_0 + \rho_0\sigma_0 \left[\frac{-\phi(\boldsymbol{W}_i\boldsymbol{\alpha})}{1-\Phi(\boldsymbol{W}_i\boldsymbol{\alpha})}\right]. \end{aligned} \tag{A4}$$

Eq. (A3) represents the expected value of Y for a person with characteristics $\boldsymbol{X}_i, \boldsymbol{W}_i$ who is *living in an urban area*, and Eq. (A4) represents the expected value for a person with those observed attributes who is *living in a non-urban area*. These quantities differ from their counterparts in Eqs. (A1) and (A2) because *unobserved* characteristics relevant to the outcome differ, on average, between residents of urban and non-urban areas.

We can also define the *counterfactual* conditional expectations as:

$$\begin{aligned} E(Y_{i1}|z_i^* < 0, \boldsymbol{X}_i, \boldsymbol{W}_i) &= \boldsymbol{X}_i\boldsymbol{\beta}_1 + E[\varepsilon_{i1}|z_i^* < 0, \boldsymbol{X}_i, \boldsymbol{W}_i] \\ &= \boldsymbol{X}_i\boldsymbol{\beta}_1 + \rho_1\sigma_1 \left[\frac{-\phi(\boldsymbol{W}_i\boldsymbol{\alpha})}{1-\Phi(\boldsymbol{W}_i\boldsymbol{\alpha})}\right] \text{ and} \end{aligned} \tag{A5}$$

$$\begin{aligned} E(Y_{i0}|z_i^* > 0, \boldsymbol{X}_i, \boldsymbol{W}_i) &= \boldsymbol{X}_i\boldsymbol{\beta}_0 + E[\varepsilon_{i0}|z_i^* > 0, \boldsymbol{X}_i, \boldsymbol{W}_i] \\ &= \boldsymbol{X}_i\boldsymbol{\beta}_0 + \rho_0\sigma_0 \frac{\phi(\boldsymbol{W}_i\boldsymbol{\alpha})}{\Phi(\boldsymbol{W}_i\boldsymbol{\alpha})}. \end{aligned} \tag{A6}$$

Eq. (A5) represents what the expected value of *Y would be* for a non-urban resident with characteristics $\boldsymbol{X}_i, \boldsymbol{W}_i$ *if she were to live in an urban area*, and conversely for Eq. (A6).

For both types of conditional expectations, the selection effect term is a product of the appropriate inverse Mills ratio quantity, IMR ($\frac{\phi(\boldsymbol{W}_i\boldsymbol{\alpha})}{\Phi(\boldsymbol{W}_i\boldsymbol{\alpha})}$ or $\frac{-\phi(\boldsymbol{W}_i\boldsymbol{\alpha})}{1-\Phi(\boldsymbol{W}_i\boldsymbol{\alpha})}$), and its coefficient ($\rho\sigma$, which is often known as λ in Heckman's selection model), and it is the consequence of the *truncation* of the normal distribution of z^* by conditioning on its sign. In the context of the bivariate normal distribution, if the two normal distributions are independent, a truncation of one normal random variable (by selection) does not affect the other normal random variable. However, since the two distributions are correlated here, the truncation of one also truncates the other normal distribution (and the expected value of the truncated normal distribution is the IMR term).

In addition to the ATE, defined as the difference between Eq. (A1) and Eq. (A2), integrated over the distribution of $\boldsymbol{X}$, we can define:

- *the (average) treatment effect on the treated (TT)* (in log-transformed terms) as the difference between Eq. (A3) (the urban resident's expected ln(VMD)) and Eq. (A6) (the urban resident's counterfactual expected ln(VMD) if she were to live in a non-urban area), integrated over the distribution of $\boldsymbol{X}$; and
- *the (average) treatment effect on the untreated (TUT)* (in log-transformed terms) as the difference between Eq. (A5) (the non-urban resident's counterfactual expected ln(VMD) if she were to live in an urban area) and Eq. (A4) (the non-urban resident's expected ln(VMD) where she currently lives), integrated over the distribution of $\boldsymbol{X}$.

Returning to our issue, Table A1 presents various estimated expected values of ln(VMD), distinguishing urban and non-urban components (for simplicity, we do not apply the sample weights here; also, we explain the relevant quantities in log-scale and then discuss the back-transformation later). It can be seen that the estimated unconditional means, after properly accounting for selection effects (obtained by estimating Eqs. (A1) and (A2) and averaging the respective results over the entire sample), have a greater difference (which is the average treatment effect, ATE, in log scale) than the observed means. This may initially be surprising, as discussed in Section 6.1, but two issues arise when we simply compare our estimated unconditional means with the observed group means in the sample. First, if self-selection is in effect (i.e. significant correlations ρ exist), then those observed group means are the consequence of self-selection. That is, those values have already absorbed the selection effects. Second, the observed group means are based on the segmented subsamples (since they are already self-selected), not on the pooled sample (as is the case for the unconditional means, per Eq. (3.31)). As shown in Table A1, the conditional means on the segmented sample replicate the observed group means. Hence, a more proper interpretation is that when we eliminate self-selection effects (i.e. if we could randomly assign individuals to neighborhood type), then the expected outcomes for each class are 3.62 and 4.93, respectively.

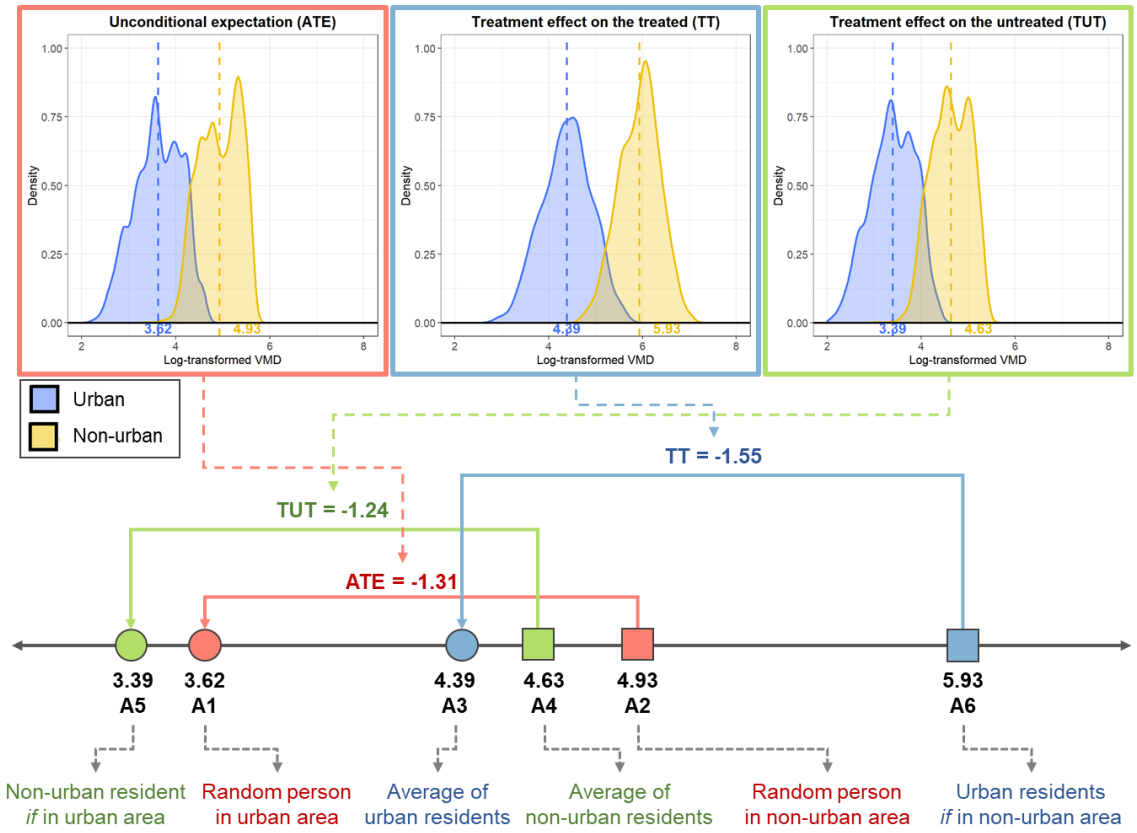
Stereotypically, unobserved "urbanite" predispositions would reduce VMD relative to that of an average person, thus leading us to expect negative ρ_1 and ρ_0. This is important because, by Eqs. (A3) and (A4), the directions of the self-selection correction are determined by the signs of ρ_1 and ρ_0, since σ_1, σ_0, and $\frac{\phi(\boldsymbol{W}_i\boldsymbol{\alpha})}{\Phi(\boldsymbol{W}_i\boldsymbol{\alpha})}$ are strictly positive and

$\left[\frac{-\phi(\boldsymbol{W}_i\boldsymbol{\alpha})}{1-\Phi(\boldsymbol{W}_i\boldsymbol{\alpha})}\right]$ is strictly negative. However, in our empirical application we obtained $\hat{\rho}_1 = 0.66$ and $\hat{\rho}_0 = 0.84$. This means we obtained opposite-from-expected signs for $\rho_1$and ρ_0, and the two error correlations have substantial magnitudes. The consequences are:

- For the urban resident group, given that $E(Y_{i1}|z_i^* > 0, \boldsymbol{X}_i, \boldsymbol{W}_i)$ approximates the observed outcome (since the latter is conditional on the same variables), Eq. (A3) shows that $E(Y_{i1}|z_i^* > 0, \boldsymbol{X}_i, \boldsymbol{W}_i) > \boldsymbol{X}_i\boldsymbol{\beta}_1$, and thus the unconditional expectation of $\boldsymbol{X}_i\boldsymbol{\beta}_1$ is (unexpectedly) smaller than the observed mean.
- For the non-urban resident group, given that $E(Y_{i0}|z_i^* < 0, \boldsymbol{X}_i, \boldsymbol{W}_i)$ approximates the observed outcome, Eq. (A4) shows that $E(Y_{i0}|z_i^* < 0, \boldsymbol{X}_i, \boldsymbol{W}_i) < \boldsymbol{X}_i\boldsymbol{\beta}_0$, and thus the unconditional expectation is greater than the observed mean. This direction of correction is also unexpected, and such a high correlation makes the corrected expected value rather large.

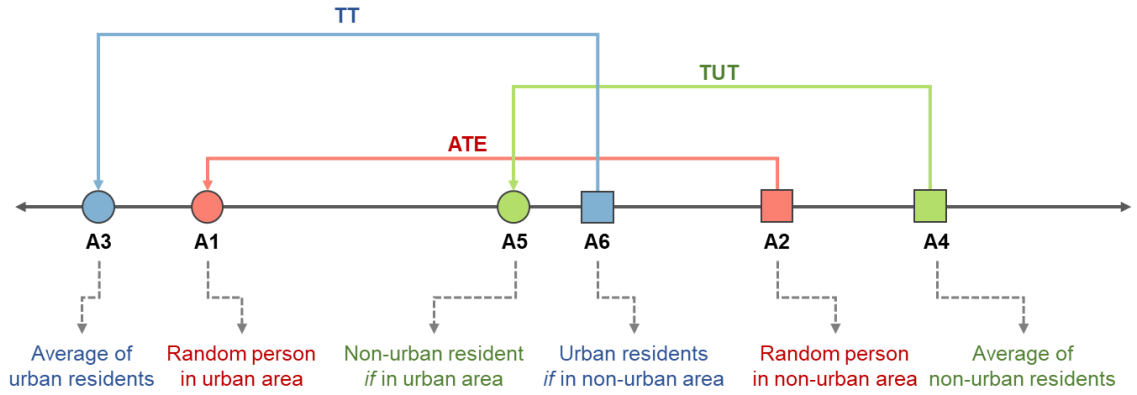
In a nutshell, in this empirical application, we end up having an ATE that is greater than the observed difference.

Table A1 also presents the components of the treatment effects on the treated (TT) and on the untreated (TUT), and all three treatment effects are illustrated in Figure A1. For reference, Figure A2 schematically portrays the situation when ρ_1 and ρ_0 are both negative.



Note1: The top graphs portray the empirical distributions of the individual-specific values of the various quantities under discussion.
Note2: In the lower portion of the figure, the notations "A5", etc. refer to the equation numbers used to compute each quantity.

Figure A1. Relationships among the various components of interest in our sample (when ρ_1 and ρ_0 are positive)



Note: The relationship between the Eqs. A5 and A6 components could be reversed, depending on the parameter estimates and the data.

Figure A2. Schematic of relationships among the various components of interest when ρ_1 and ρ_0 are negative

Table A1. Estimated expected values of Y (log-transformed VMD; unweighted)

	Formula for the Estimated Difference	Urban	Non-urban	Difference
Observed mean	$\frac{1}{n_1}\sum_{z=1}\boldsymbol{Y}_{i1} - \frac{1}{n_0}\sum_{z=0}\boldsymbol{Y}_{i0}$	4.39	4.63	-0.24
Conditional (on selection) mean on segmented sample (Eq. A3 – Eq. A4)	$\frac{1}{n_1}\sum_{z=1}\left\{\boldsymbol{X}_{i1}\,\widehat{\boldsymbol{\beta}}_1 + \hat{\rho}_1\hat{\sigma}_1\,\frac{\phi(\boldsymbol{W}_i\widehat{\boldsymbol{\alpha}})}{\Phi(\boldsymbol{W}_i\widehat{\boldsymbol{\alpha}})}\right\} -$ $\frac{1}{n_0}\sum_{z=0}\left\{\boldsymbol{X}_{i1}\,\widehat{\boldsymbol{\beta}}_0 + \hat{\rho}_0\hat{\sigma}_0\,\frac{-\phi(\boldsymbol{W}_i\widehat{\boldsymbol{\alpha}})}{\Phi(-\boldsymbol{W}_i\widehat{\boldsymbol{\alpha}})}\right\}$	4.39	4.63	-0.24
Unconditional (on selection) mean (ATE; Eq. A1 – Eq. A2)	$\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_{i1}\,\widehat{\boldsymbol{\beta}}_1 - \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_{i0}\,\widehat{\boldsymbol{\beta}}_0$	3.62	4.93	-1.31
Treatment effect on the treated (TT; Eq. A3 – Eq. A6)	$\left\{\frac{1}{n_1}\sum_{z=1}\boldsymbol{X}_{i1}\,\widehat{\boldsymbol{\beta}}_1 + \frac{1}{n_1}\sum_{z=1}\hat{\rho}_1\hat{\sigma}_1\,\frac{\phi(\boldsymbol{W}_i\widehat{\boldsymbol{\alpha}})}{\Phi(\boldsymbol{W}_i\widehat{\boldsymbol{\alpha}})}\right\} -$ $\left\{\frac{1}{n_1}\sum_{z=1}\boldsymbol{X}_{i1}\,\widehat{\boldsymbol{\beta}}_0 + \frac{1}{n_1}\sum_{z=1}\hat{\rho}_0\hat{\sigma}_0\,\frac{\phi(\boldsymbol{W}_i\widehat{\boldsymbol{\alpha}})}{\Phi(\boldsymbol{W}_i\widehat{\boldsymbol{\alpha}})}\right\}$	4.39	5.93	-1.55
Treatment effect on the untreated (TUT; Eq. A5 – Eq. A4)	$\left\{\frac{1}{n_0}\sum_{z=0}\boldsymbol{X}_{i1}\,\widehat{\boldsymbol{\beta}}_1 + \frac{1}{n_0}\sum_{z=0}\hat{\rho}_1\hat{\sigma}_1\,\frac{-\phi(\boldsymbol{W}_i\widehat{\boldsymbol{\alpha}})}{\Phi(-\boldsymbol{W}_i\widehat{\boldsymbol{\alpha}})}\right\} -$ $\left\{\frac{1}{n_0}\sum_{z=0}\boldsymbol{X}_{i1}\,\widehat{\boldsymbol{\beta}}_0 + \frac{1}{n_0}\sum_{z=0}\hat{\rho}_0\hat{\sigma}_0\,\frac{-\phi(\boldsymbol{W}_i\widehat{\boldsymbol{\alpha}})}{\Phi(-\boldsymbol{W}_i\widehat{\boldsymbol{\alpha}})}\right\}$	3.39	4.63	-1.24

Another complication arises when we back-transform the dependent variable from the log scale to its original units (miles driven). It is common to model VMD with the log transformation to achieve greater normality, but reporting the results in the original scale is not straightforward. As mentioned in the body of the CHAPTER 3, since $\ln(VMD+1) = Y$ and $\sim N[\boldsymbol{X\beta}, \sigma^2]$, then $(VMD+1 \mid \boldsymbol{X}) \sim LN[\boldsymbol{X\beta}, \sigma^2]$ (throughout this passage, the individual subscript *i* is suppressed for simplicity). From known properties of the lognormal distribution, $[VMD+1 \mid \boldsymbol{X}] = E[\exp(\boldsymbol{X\beta}_z + \varepsilon) \mid \boldsymbol{X}] = \exp(\boldsymbol{X\beta}_z + \sigma_z^2/2)$. This equation gives the back-transformed counterparts to Eqs. (A1) (for *z* = 1) and (A2) (for *z* = 0).

To help increase awareness of little-known details associated with back-transforming the log transformation in the context of treatment evaluations, we provide the proper equations for the back-transformed conditional means. Adapting the equation found

in Yen and Rosinski (2008; the formal derivation of the equation is in their Appendix) to our notation, the conditional mean of V, where $\ln V = Y = \boldsymbol{X\beta} + \varepsilon$, is:

$$\begin{aligned} &E(V|c_1 - \boldsymbol{W\alpha} < u < c_2 - \boldsymbol{W\alpha}) \\ &\quad = \exp(\boldsymbol{X\beta} + \sigma^2/2)\,\frac{\Phi(c_2 - \boldsymbol{W\alpha} - \rho\sigma) - \Phi(c_1 - \boldsymbol{W\alpha} - \rho\sigma)}{\Phi(c_2 - \boldsymbol{W\alpha}) - \Phi(c_1 - \boldsymbol{W\alpha})}, \end{aligned} \tag{A7}$$

where c_1 and c_2 are constants to be properly specified depending on the context, as shown below.

The conditional expectation for the treatment group (i.e. urban residents; the back-transformed counterpart to Eq. (A3)) is ($c_1 = 0, c_2 = \infty$):

$$\begin{aligned} &E(V_1| - \boldsymbol{W\alpha} < u, \boldsymbol{X}, \boldsymbol{W}) \\ &\quad = \exp(\boldsymbol{X\beta}_1 + \sigma_1^2/2)\,\frac{\Phi(\infty - \boldsymbol{W\alpha} - \rho_1\sigma_1) - \Phi(0 - \boldsymbol{W\alpha} - \rho_1\sigma_1)}{\Phi(\infty - \boldsymbol{W\alpha}) - \Phi(0 - \boldsymbol{W\alpha})} \\ &\quad = \exp(\boldsymbol{X\beta}_1 + \sigma_1^2/2)\,\frac{1 - \Phi(-\boldsymbol{W\alpha} - \rho_1\sigma_1)}{1 - \Phi(-\boldsymbol{W\alpha})} \\ &\quad = \exp(\boldsymbol{X\beta}_1 + \sigma_1^2/2)\,\frac{\Phi(\boldsymbol{W\alpha} + \rho_1\sigma_1)}{\Phi(\boldsymbol{W\alpha})} \end{aligned} \tag{A8}$$

The conditional expectation for the untreated group (i.e. non-urban residents; the back-transformed counterpart to Eq. (A4)) is ($c_1 = -\infty, c_2 = 0$):

$$\begin{aligned} &E(V_0|u < -\boldsymbol{W\alpha}, \boldsymbol{X}, \boldsymbol{W}) \\ &\quad = \exp(\boldsymbol{X\beta}_0 + \sigma_0^2/2)\,\frac{\Phi(0 - \boldsymbol{W\alpha} - \rho_0\sigma_0) - \Phi(-\infty - \boldsymbol{W\alpha} - \rho_0\sigma_0)}{\Phi(0 - \boldsymbol{W\alpha}) - \Phi(-\infty - \boldsymbol{W\alpha})} \\ &\quad = \exp(\boldsymbol{X\beta}_0 + \sigma_0^2/2)\,\frac{\Phi(-\boldsymbol{W\alpha} - \rho_0\sigma_0) - 0}{\Phi(-\boldsymbol{W\alpha}) - 0} \\ &\quad = \exp(\boldsymbol{X\beta}_0 + \sigma_0^2/2)\,\frac{\Phi(-\boldsymbol{W\alpha} - \rho_0\sigma_0)}{\Phi(-\boldsymbol{W\alpha})} \end{aligned} \tag{A9}$$

Similarly, the counterfactual expectations are:

$$E(V_1 \,|u < -\boldsymbol{W\alpha}, \boldsymbol{X}, \boldsymbol{W}) = \exp(\boldsymbol{X\beta}_1 + \sigma_1^2/2)\,\frac{\Phi(-\boldsymbol{W\alpha}-\rho_1\sigma_1)}{\Phi(-\boldsymbol{W\alpha})} \tag{A10}$$

$$E(V_0|\,-\boldsymbol{W\alpha} < u, \boldsymbol{X}, \boldsymbol{W}) = \exp(\boldsymbol{X\beta}_0 + \sigma_0^2/2)\,\frac{\Phi(\boldsymbol{W\alpha}+\rho_0\sigma_0)}{\Phi(\boldsymbol{W\alpha})}\,, \tag{A11}$$

the respective counterparts to Eqs. (A5) (where here, $c_1 = -\infty, c_2 = 0$) and (A6) ($c_1 = 0, c_2 = \infty$).

Using Eqs. (3.32) and (A8)-(A11), Table A2 shows various expected values and treatment effects. Note that the observed and ATE quantities differ somewhat from their counterparts reported in Sections 5.2 (Table 4) and 6.1 because here, for simplicity, we have not weighted the results (Eq. (3.34) gives an example of how to do so). The non-urban counterfactual component of the TT is quite large (about 88 miles a day), but given the arguments in Section 6.1, this represents "what it would take" for an urban resident, predisposed toward an active lifestyle involving numerous and diverse destinations, to achieve such a lifestyle in a very low-density area (hence the self-selection into an urban area, whose built environment will be more supportive of such a lifestyle at a lower travel cost). On the other hand, it is likely that the sensitivity of the log transformation is also at play here, in that a fairly small reduction in the transformed counterpart to the 613.47 miles of Table A2 (i.e. in the 5.93 of Table A1) would lead to a sizable reduction in miles.

Table A2. Estimated expected values of VMD and treatment effects (unweighted)

	Urban	Non-urban	Difference
Observed	124.10	152.89	-28.79
ATE (Eq. (3.33))	71.71	255.30	-183.59
TT (Eq. (3.8) – Eq. (3.11))	149.66	613.47	-463.80
TUT (Eq. (3.10) – Eq. (3.9))	49.00	151.60	-102.60

APPENDIX B. AIRPORTS IN GEORGIA AND 2017 STATISTICS (CHAPTER 4)

Table B1. Airports in Georgia and 2017 statistics [a]

Airport		Number of flights			Number of passenger enplanements		
Origin airport	Airport code	Domestic	Inter-national	Total	Domestic	Inter-national	Total
Atlanta Hartsfield Intl. Airport	ATL	384,134	38,914	423,048	44,351,038	5,891,684	50,242,722
Savannah / Hiltonhead International Airport	SAV	15,452	179	15,631	1,182,353	7,420	1,189,773
Augusta Regional Airport- Bushfield Airport	AGS	4,905	0	4,905	268,228	0	268,228
Columbus Metropolitan Airport	CSG	1,208	0	1,208	44,767	0	44,767
Valdosta Regional Airport	VLD	1,020	0	1,020	43,734	0	43,734
Southwest Georgia Regional Airport	ABY	991	0	991	37,900	0	37,900
Brunswick Golden Isles Airport	BQK	995	0	995	36,219	0	36,219
Middle Georgia Regional Airport	MCN	69	0	69	1,292	0	1,292
Athens-Ben Epps Airport [b]	AHN	2	0	2	1	0	1

a. Created based on data provided by the Bureau of Transportation Statistics of the U.S. Department of Transportation.

b. AHN had a small and declining number of annual flights (from 1,112 in 2010 to 361 in 2014), mainly connecting to Nashville, Tennessee, because Athens is one of the communities subsidized by the US government "Essential Air Service" program, which aims to secure a minimum level of air service in local communities. Since 2014, AHN has been de-subsidized due to unmet passenger load requirements, although it completed a $17 million runway extension project in 2015 and is now looking for new commercial airlines to provide service.

REFERENCES

Abotalebi, Elnaz, Darren M. Scott, and Mark R. Ferguson (2019) Why is electric vehicle uptake low in Atlantic Canada? A comparison to leading adoption provinces. *Journal of Transport Geography* **74**, 289-298.

Adanu, Emmanuel Kofi, Alexander Hainen, and Steven Jones (2018) Latent class analysis of factors that influence weekday and weekend single-vehicle crash severities. *Accident Analysis & Prevention* **113**, 187-192.

Adanu, Emmanuel Kofi, and Steven Jones (2017) Effects of Human-Centered Factors on Crash Injury Severities. *Journal of Advanced Transportation* **2017**, 1208170.

Aguiléra, Anne, and Laurent Proulhac (2015) Socio-occupational and geographical determinants of the frequency of long-distance business travel in France. *Journal of Transport Geography* **43**, 28-35.

Ahmed, Tanjeeb, Michael Hyland, Navjyoth J. S. Sarma, Suman Mitra, and Arash Ghaffar (2020) Quantifying the employment accessibility benefits of shared automated vehicle mobility services: Consumer welfare approach using logsums. *Transportation Research Part A: Policy and Practice* **141**, 221-247.

Akar, Gulsah, and Jean-Michel Guldmann (2012) Another Look at Vehicle Miles Traveled:Determinants of Vehicle use in Two-Vehicle Households. *Transportation Research Record* **2322**(**1**), 110-118.

Åkerman, Jonas (2005) Sustainable air transport—on track in 2050. *Transportation Research Part D: Transport and Environment* **10**(**2**), 111-126.

Alemi, Farzad, Giovanni Circella, Patricia Mokhtarian, and Susan Handy (2019) What drives the use of ridehailing in California? Ordered probit models of the usage frequency of Uber and Lyft. *Transportation Research Part C: Emerging Technologies* **102**, 233-248.

Alizadeh, Hamzeh, Bilal Farooq, Catherine Morency, and Nicolas Saunier (2019) Frequent versus occasional drivers: A hybrid route choice model. *Transportation Research Part F: Traffic Psychology and Behaviour* **64**, 171-183.

Allen, Jaime, Juan Carlos Muñoz, and Juan de Dios Ortúzar (2019) Understanding public transport satisfaction: Using Maslow's hierarchy of (transit) needs. *Transport Policy* **81**, 75-94.

Allenby, Greg M, and Peter E Rossi (1998) Marketing models of consumer heterogeneity. *Journal of econometrics* **89**(**1-2**), 57-78.

Alonso-González, María J., Niels van Oort, Oded Cats, Sascha Hoogendoorn-Lanser, and Serge Hoogendoorn (2020) Value of time and reliability for urban pooled on-demand services. *Transportation Research Part C: Emerging Technologies* **115**, 102621.

Alwosheel, Ahmad, Sander van Cranenburgh, and Caspar G. Chorus (2019) 'Computer says no' is not enough: Using prototypical examples to diagnose artificial neural networks for discrete choice analysis. *Journal of Choice Modelling* **33**, 100186.

Amaral, Christopher, and Ceren Kolsarici (2020) The financial advice puzzle: The role of consumer heterogeneity in the advisor choice. *Journal of Retailing and Consumer Services* **54**, 102014.

Amemiya, Takeshi (1985) *Advanced Econometrics*: Harvard university press.

Anderson, Jason, and Salvador Hernandez (2017) Heavy-Vehicle Crash Rate Analysis: Comparison of Heterogeneity Methods Using Idaho Crash Data. *Transportation Research Record* **2637**(**1**), 56-66.

Angueira, Jaime, Ahmadreza Faghih-Imani, Annesha Enam, Karthik C. Konduri, and Naveen Eluru (2015) Exploration of Short-Term Vehicle Utilization Choices in Households with Multiple Vehicle Types. *Transportation Research Record* **2493**(**1**), 39-47.

Angueira, Jaime, Karthik Charan Konduri, Vincent Chakour, and Naveen Eluru (2019) Exploring the relationship between vehicle type choice and distance traveled: a latent segmentation approach. *Transportation Letters* **11**(**3**), 146-157.

Anowar, Sabreena, and Naveen Eluru (2018) Univariate or multivariate analysis for better prediction accuracy? A case study of heterogeneity in vehicle ownership. *Transportmetrica A: Transport Science* **14**(**8**), 635-668.

Anowar, Sabreena, Ahmadreza Faghih-Imani, Eric J. Miller, and Naveen Eluru (2019) Regret minimization based joint econometric model of mode choice and departure time: a case study of university students in Toronto, Canada. *Transportmetrica A: Transport Science* **15**(**2**), 1214-1246.

Anowar, Sabreena, Shamsunnahar Yasmin, Naveen Eluru, and Luis F. Miranda-Moreno (2014) Analyzing car ownership in Quebec City: a comparison of traditional and latent class ordered and unordered models. *Transportation* **41**(**5**), 1013-1039.

Ansari, Asim, Kamel Jedidi, and Sharan Jagpal (2000) A Hierarchical Bayesian Methodology for Treating Heterogeneity in Structural Equation Models. *Marketing Science* **19**(**4**), 328-347.

Araghi, Yashar, Maarten Kroesen, Eric Molin, and Bert Van Wee (2016) Revealing heterogeneity in air travelers' responses to passenger-oriented environmental policies: A discrete-choice latent class model. *International Journal of Sustainable Transportation* **10**(**9**), 765-772.

Ardeshiri, Ali, and Akshay Vij (2019) Lifestyles, residential location, and transport mode use: A hierarchical latent class choice model. *Transportation Research Part A: Policy and Practice* **126**, 342-359.

Arentze, Theo A. (2015) Individuals' social preferences in joint activity location choice: A negotiation model and empirical evidence. *Journal of Transport Geography* **48**, 76-84.

Arunotayanun, Kriangkrai, and John W. Polak (2011) Taste heterogeneity and market segmentation in freight shippers' mode choice behaviour. *Transportation Research Part E: Logistics and Transportation Review* **47**(**2**), 138-148.

Astroza, Sebastian, Venu M. Garikapati, Ram M. Pendyala, Chandra R. Bhat, and Patricia L. Mokhtarian (2019) Representing heterogeneity in structural relationships among multiple choice variables using a latent segmentation approach. *Transportation* **46**(**5**), 1755-1784.

Astroza, Sebastian, Abdul Rawoof Pinjari, Chandra R. Bhat, and Sergio R. Jara-Díaz (2017) A Microeconomic Theory–Based Latent Class Multiple Discrete–Continuous Choice Model of Time Use and Goods Consumption. *Transportation Research Record* **2664**(**1**), 31-41.

Athey, Susan (2019) "The Impact of Machine Learning on Economics." In *The Economics of Artificial Intelligence: An Agenda*, 507-547. University of Chicago Press.

Athey, Susan, and Guido W. Imbens (2019) Machine Learning Methods That Economists Should Know About. *Annual Review of Economics* **11**(**1**), 685-725.

Aultman-Hall, Lisa, Chester Harvey, James Sullivan, and Jeffrey J. LaMondia (2018) The implications of long-distance tour attributes for national travel data collection in the United States. *Transportation* **45**(**3**), 875-903.

Babbie, Earl (2012) *The Practice of Social Research*. 13th ed. Belmont, CA: Wadsworth Publishing Company.

Bae, Bumjoon, Yuandong Liu, Lee D. Han, and Hamparsum Bozdogan (2019) Spatio-temporal traffic queue detection for uninterrupted flows. *Transportation Research Part B: Methodological* **129**, 20-34.

Bago d'Uva, Teresa, Andrew M. Jones, and Eddy van Doorslaer (2009) Measurement of horizontal inequity in health care utilisation using European panel data. *Journal of Health Economics* **28**(**2**), 280-289.

Bailey, Joseph, and Jonn Axsen (2015) Anticipating PEV buyers' acceptance of utility controlled charging. *Transportation Research Part A: Policy and Practice* **82**, 29-46.

Bakk, Zsuzsa, Fetene B. Tekle, and Jeroen K. Vermunt (2013) Estimating the Association between Latent Class Membership and External Variables Using Bias-adjusted Three-step Approaches. *Sociological Methodology* **43**(**1**), 272-311.

Baldacchino, Tara, Elizabeth J. Cross, Keith Worden, and Jennifer Rowson (2016) Variational Bayesian mixture of experts models and sensitivity analysis for nonlinear dynamical systems. *Mechanical Systems and Signal Processing* **66-67**, 178-200.

Barbour, Natalia, Nikhil Menon, Yu Zhang, and Fred Mannering (2019) Shared automated vehicles: A statistical analysis of consumer use likelihoods and concerns. *Transport Policy* **80**, 86-93.

Bauer, Daniel J. (2005) A Semiparametric Approach to Modeling Nonlinear Relations Among Latent Variables. *Structural Equation Modeling: A Multidisciplinary Journal* **12**(**4**), 513-535.

Beck, Matthew J., John M. Rose, and David A. Hensher (2013) Environmental attitudes and emissions charging: An example of policy implications for vehicle choice. *Transportation Research Part A: Policy and Practice* **50**, 171-182.

Becken, Susanne (2002) Analysing International Tourist Flows to Estimate Energy Use Associated with Air Travel. *Journal of Sustainable Tourism* **10**(**2**), 114-131.

Beel, Joeran, Bela Gipp, Stefan Langer, and Corinna Breitinger (2016) Research-paper recommender systems: a literature survey. *International Journal on Digital Libraries* **17**(**4**), 305-338.

Behnood, Ali, and Fred L. Mannering (2016) An empirical assessment of the effects of economic recessions on pedestrian-injury crashes using mixed and latent-class models. *Analytic Methods in Accident Research* **12**, 1-17.

Bell, Robert M., Yehuda Koren, and Chris Volinsky (2007) The bellkor solution to the Netflix prize. *KorBell Team's Report to Netflix*.

Ben-Akiva, Moshe, and Bruno Boccara (1995) Discrete choice models with latent choice sets. *International Journal of Research in Marketing* **12**(**1**), 9-24.

Ben-Akiva, Moshe, Daniel McFadden, Kenneth Train, Joan Walker, Chandra Bhat, Michel Bierlaire, Denis Bolduc, Axel Boersch-Supan, David Brownstone, and David S Bunch (2002) Hybrid choice models: progress and challenges. *Marketing Letters* **13**(**3**), 163-175.

Ben-Akiva, Moshe E, and Steven R Lerman (1985) *Discrete Choice Analysis: Theory and Application to Travel Demand*. Cambridge, MA: MIT Press.

Bentz, Yves, and Dwight Merunka (2000) Neural networks and the multinomial logit for brand choice modelling: a hybrid approach. *Journal of Forecasting* **19**(**3**), 177-200.

Bergantino, Angela Stefania, and Leonardo Madio (2020) Intermodal competition and substitution. HSR versus air transport: Understanding the socio-economic determinants of modal choice. *Research in Transportation Economics* **79**, 100823.

Berliner, Rosaria M., Lisa Aultman-Hall, and Giovanni Circella (2018) Exploring the Self-Reported Long-Distance Travel Frequency of Millennials and Generation X in California. *Transportation Research Record* **2672**(**47**), 208-218.

Bhat, Chandra R (1997) An endogenous segmentation mode choice model with an application to intercity travel. *Transportation Science* **31**(**1**), 34-48.

Bhat, Chandra R., and Naveen Eluru (2009) A copula-based approach to accommodate residential self-selection effects in travel behavior modeling. *Transportation Research Part B: Methodological* **43**(**7**), 749-765.

Bhat, Chandra R., Teresa Frusti, Huimin Zhao, Stefan Schönfelder, and Kay W. Axhausen (2004) Intershopping duration: an analysis using multiweek data. *Transportation Research Part B: Methodological* **38**(**1**), 39-60.

Bierlaire, Michel (2003) "BIOGEME: A free package for the estimation of discrete choice models." Swiss transport research conference.

Bishop, Christopher M (1994) Mixture density networks, available at https://publications.aston.ac.uk/id/eprint/373/1/NCRG_94_004.pdf.

Bishop, Christopher M (1995) *Neural Networks for Pattern Recognition*: Oxford university press.

Bishop, Christopher M (2006) *Pattern Recognition and Machine Learning*. Springer.

Boeri, Marco, Riccardo Scarpa, and Caspar G. Chorus (2014) Stated choices and benefit estimates in the context of traffic calming schemes: Utility maximization, regret minimization, or both? *Transportation Research Part A: Policy and Practice* **61**, 121-135.

Bolck, Annabel, Marcel Croon, and Jacques Hagenaars (2004) Estimating Latent Structure Models with Categorical Variables: One-Step Versus Three-Step Estimators. *Political Analysis* **12**(**1**), 3-27.

Box, George E. P. (1976) Science and Statistics. *Journal of the American Statistical Association* **71**(**356**), 791-799.

Boyer, Ryan C., William T. Scherer, and Michael C. Smith (2017) Trends Over Two Decades of Transportation Research:A Machine Learning Approach. *Transportation Research Record* **2614**(**1**), 1-9.

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

Breiman, Leo, Jerome Friedman, Richard Olshen, and Charles Stone (1984) *Classification and Regression Trees*: Chapman & Hall/CRC.

Brey, Raúl, and Joan L. Walker (2011) Latent temporal preferences: An application to airline travel. *Transportation Research Part A: Policy and Practice* **45**(**9**), 880-895.

Brown, Sarah, Robert B. Durand, Mark N. Harris, and Tim Weterings (2014) Modelling financial satisfaction across life stages: A latent class approach. *Journal of Economic Psychology* **45**, 117-127.

Brownstone, David, David S. Bunch, and Kenneth Train (2000) Joint mixed logit models of stated and revealed preferences for alternative-fuel vehicles. *Transportation Research Part B: Methodological* **34**(**5**), 315-338.

Buddhavarapu, Prasad, James G. Scott, and Jorge A. Prozzi (2016) Modeling unobserved heterogeneity using finite mixture random parameters for spatially correlated discrete count data. *Transportation Research Part B: Methodological* **91**, 492-510.

Byrne, Barbara M, Richard J Shavelson, and Bengt Muthén (1989) Testing for the equivalence of factor covariance and mean structures: the issue of partial measurement invariance. *Psychological bulletin* **105**(**3**), 456.

Cameron, A Colin, and Pravin K Trivedi (2005) *Microeconometrics: Methods and Applications*. New York, U.S. Cambridge University Press.

Cao, Xinyu (2009) Disentangling the influence of neighborhood type and self-selection on driving behavior: an application of sample selection model. *Transportation* **36**(**2**), 207-222.

Cao, Xinyu, Patricia L. Mokhtarian, and Susan L. Handy (2009) Examining the Impacts of Residential Self-Selection on Travel Behaviour: A Focus on Empirical Findings. *Transport Reviews* **29**(**3**), 359-395.

Cervero, Robert, and Jin Murakami (2010) Effects of Built Environments on Vehicle Miles Traveled: Evidence from 370 US Urbanized Areas. *Environment and Planning A* **42**(**2**), 400-418.

Chakour, Vincent, and Naveen Eluru (2014) Analyzing commuter train user behavior: a decision framework for access mode and station choice. *Transportation* **41**(**1**), 211-228.

Chand, Sai, and Vinayak V. Dixit (2018) Application of Fractal theory for crash rate prediction: Insights from random parameters and latent class tobit models. *Accident Analysis & Prevention* **112**, 30-38.

Chen, Faan, Jiaorong Wu, Xiaohong Chen, and Jianjun Wang (2017) Vehicle kilometers traveled reduction impacts of Transit-Oriented Development: Evidence from Shanghai City. *Transportation Research Part D: Transport and Environment* **55**, 227-245.

Chen, Fang-Yuan, Pi-Yuan Hsu, and Ting-Wei Lin (2011) Air Travelers' environmental consciousness: A preliminary investigation in Taiwan. *International Journal of Business and Management* **6**(**12**), 78.

Chiou, Yu-Chiun, Rong-Chang Jou, Chu-Yun Kao, and Chiang Fu (2013) The adoption behaviours of freeway electronic toll collection: A latent class modelling approach. *Transportation Research Part E: Logistics and Transportation Review* **49**(**1**), 266-280.

Choi, Sungtaek, and Patricia L. Mokhtarian (2020) How attractive is it to use the internet while commuting? A work-attitude-based segmentation of Northern California commuters. *Transportation Research Part A: Policy and Practice* **138**, 37-50.

Choo, Sangho, Patricia L. Mokhtarian, and Ilan Salomon (2005) Does telecommuting reduce vehicle-miles traveled? An aggregate time series analysis for the U.S. *Transportation* **32**(**1**), 37-64.

Chorus, Caspar G, Theo A Arentze, and Harry JP Timmermans (2008) A random regret-minimization model of travel choice. *Transportation Research Part B: Methodological* **42**(**1**), 1-18.

Chu, Hsing-Chung (2014) Assessing factors causing severe injuries in crashes of high-deck buses in long-distance driving on freeways. *Accident Analysis & Prevention* **62**, 130-136.

Circella, Giovanni, Lew Fulton, Farzad Alemi, Rosaria M. Berliner, Kate Tiedman, Patricia L. Mokhtarian, and Susan Handy (2016) *What Affects Millennials' Mobility? PART I: Investigating the Environmental Concerns, Lifestyles, Mobility-Related Attitudes and Adoption of Technology of Young Adults in California*. Institute of Transportation Studies UC DAVIS. Available at http://ncst.ucdavis.edu/project/ucd-ct-to-11/.

Circella, Giovanni, Kate Tiedeman, Susan Handy, Farzad Alemi, and Patricia Mokhtarian (2016) *What Affects U.S. Passenger Travel? Current Trends and Future Perspectives*. UC Davis. Available at https://ncst.ucdavis.edu/research-product/what-affects-us-passenger-travel-current-trends-and-future-perspectives.

Clewlow, Regina R, and Gouri Shankar Mishra (2017) Disruptive transportation: The adoption, utilization, and impacts of ride-hailing in the United States. *Institute of Transportation Studies, University of California, Davis*.

Cole, Veronica T., and Daniel J. Bauer (2016) A Note on the Use of Mixture Models for Individual Prediction. *Structural Equation Modeling: A Multidisciplinary Journal* **23**(**4**), 615-631.

Collins, Andrew T., John M. Rose, and David A. Hensher (2013) Specification issues in a generalised random parameters attribute nonattendance model. *Transportation Research Part B: Methodological* **56**, 234-253.

Cragg, John G (1971) Some statistical models for limited dependent variables with application to the demand for durable goods. *Econometrica (pre-1986)* **39**(**5**), 829.

Cybenko, G. (1989) Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems* **2**(**4**), 303-314.

Czepkiewicz, Michał, Jukka Heinonen, Petter Næss, and Harpa Stefansdóttir (2020) Who travels more, and why? A mixed-method study of urban dwellers' leisure travel. *Travel Behaviour and Society* **19**, 67-81.

Czepkiewicz, Michał, Jukka Heinonen, and Juudit Ottelin (2018) Why do urbanites travel more than do others? A review of associations between urban form and long-distance leisure travel. *Environmental Research Letters* **13**(**7**), 073001.

Das, Subasish, Karen Dixon, Xiaoduan Sun, Anandi Dutta, and Michelle Zupancich (2017) Trends in Transportation Research: Exploring Content Analysis in Topics. *Transportation Research Record* **2614**(**1**), 27-38.

Das, Subasish, Anandi Dutta, and Marcus Brewer (2020) Transportation research record articles: a case study of trend mining. The Proceedings of Transportation Research Board Annual Meeting, Washington DC.

Das, Subasish, Xiaoduan Sun, and Anandi Dutta (2016) Text Mining and Topic Modeling of Compendiums of Papers from Transportation Research Board Annual Meetings. *Transportation Research Record* **2552**(**1**), 48-56.

Davidson, Russell, and James G. MacKinnon (2000) Bootstrap tests: how many bootstraps? *Econometric Reviews* **19**(**1**), 55-68.

Davis, Adam W., Elizabeth C. McBride, Krzysztof Janowicz, Rui Zhu, and Konstadinos G. Goulias (2018) Tour-Based Path Analysis of Long-Distance Non-Commute Travel Behavior in California. *Transportation Research Record* **2672**(**49**), 1-11.

Davison, Lisa, Clare Littleford, and Tim Ryley (2014) Air travel attitudes and behaviours: The development of environment-based segments. *Journal of Air Transport Management* **36**, 13-22.

de Souza Silva, Laize Andréa, Maurício Oliveira de Andrade, and Maria Leonor Alves Maia (2018) How does the ride-hailing systems demand affect individual transport regulation? *Research in Transportation Economics* **69**, 600-606.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977) Maximum Likelihood from Incomplete Data Via the EM Algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)* **39**(**1**), 1-22.

DeSarbo, Wayne S., and William L. Cron (1988) A maximum likelihood methodology for clusterwise linear regression. *Journal of Classification* **5**(**2**), 249-282.

Dias, Felipe F., Patrícia S. Lavieri, Taehooie Kim, Chandra R. Bhat, and Ram M. Pendyala (2019) Fusing Multiple Sources of Data to Understand Ride-Hailing Use. *Transportation Research Record* **2673**(**6**), 214-224.

Diebolt, Jean, and Christian P. Robert (1994) Estimation of Finite Mixture Distributions through Bayesian Sampling. *Journal of the Royal Statistical Society. Series B (Methodological)* **56**(**2**), 363-375.

Ding, Chuan, Xiaolei Ma, Yinhai Wang, and Yunpeng Wang (2015) Exploring the influential factors in incident clearance time: Disentangling causation from self-selection bias. *Accident Analysis & Prevention* **85**, 58-65.

Dong, Xiaojing, and Frank S. Koppelman (2014) Comparison of continuous and discrete representations of unobserved heterogeneity in logit models. *Journal of Marketing Analytics* **2**(**1**), 43-58.

Dow, William H., and Edward C. Norton (2003) Choosing Between and Interpreting the Heckit and Two-Part Models for Corner Solutions. *Health Services and Outcomes Research Methodology* **4**(**1**), 5-18.

Dowds, Jonathan, Lisa Aultman-Hall, and Jeffrey J. LaMondia (2020) Comparing alternative methods of collecting self-assessed overnight long-distance travel frequencies. *Travel Behaviour and Society* **19**, 124-136.

Dubey, Subodh, Prateek Bansal, Ricardo A. Daziano, and Erick Guerra (2020) A Generalized Continuous-Multinomial Response Model with a t-distributed Error Kernel. *Transportation Research Part B: Methodological* **133**, 114-141.

Efron, B., and R. Tibshirani (1986) Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy. *Statist. Sci.* **1**(**1**), 54-75.

El Zarwi, Feras, Akshay Vij, and Joan L. Walker (2017) A discrete choice framework for modeling and forecasting the adoption and diffusion of new transportation services. *Transportation Research Part C: Emerging Technologies* **79**, 207-223.

Eluru, Naveen, Morteza Bagheri, Luis F. Miranda-Moreno, and Liping Fu (2012) A latent class modeling approach for identifying vehicle driver injury severity factors at highway-railway crossings. *Accident Analysis & Prevention* **47**, 119-127.

Enzler, Heidi Bruderer (2017) Air travel for private purposes. An analysis of airport access, income and environmental concern in Switzerland. *Journal of Transport Geography* **61**, 1-8.

Erdoğan, Sevgi, Cinzia Cirillo, and Jean-Michel Tremblay (2015) Ridesharing as a Green Commute Alternative: A Campus Case Study. *International Journal of Sustainable Transportation* **9**(**5**), 377-388.

Espino, Raquel, and Concepción Román (2020) Valuation of transfer for bus users: The case of Gran Canaria. *Transportation Research Part A: Policy and Practice* **137**, 131-144.

Faghih-Imani, Ahmadreza, and Naveen Eluru (2020) A finite mixture modeling approach to examine New York City bicycle sharing system (CitiBike) users' destination preferences. *Transportation* **47**(**2**), 529-553.

Fagnant, Daniel J., and Kara M. Kockelman (2018) Dynamic ride-sharing and fleet sizing for a system of shared autonomous vehicles in Austin, Texas. *Transportation* **45**(**1**), 143-158.

Fatmi, Mahmudur Rahman, Subeh Chowdhury, and Muhammad Ahsanul Habib (2017) Life history-oriented residential location choice model: A stress-based two-tier panel modeling approach. *Transportation Research Part A: Policy and Practice* **104**, 293-307.

Fatmi, Mahmudur Rahman, and Muhammad Ahsanul Habib (2016) Modeling Travel Tool Ownership of the Elderly Population: Latent Segmentation-Based Logit Model. *Transportation Research Record* **2565**(**1**), 18-26.

Fatmi, Mahmudur Rahman, and Muhammad Ahsanul Habib (2019) Modeling Vehicle Collision Injury Severity Involving Distracted Driving: Assessing the Effects of

Land Use and Built Environment. *Transportation Research Record* **2673**(**7**), 181-191.

Ferguson, Mark, Moataz Mohamed, Christopher D. Higgins, Elnaz Abotalebi, and Pavlos Kanaroglou (2018) How open are Canadian households to electric vehicles? A national latent class choice analysis with willingness-to-pay and metropolitan characterization. *Transportation Research Part D: Transport and Environment* **58**, 208-224.

Finch, W. Holmes, and Kendall Cotton Bronk (2011) Conducting Confirmatory Latent Class Analysis Using Mplus. *Structural Equation Modeling: A Multidisciplinary Journal* **18**(**1**), 132-151.

Fountas, Grigorios, Panagiotis Ch Anastasopoulos, and Fred L. Mannering (2018) Analysis of vehicle accident-injury severities: A comparison of segment- versus accident-based latent class ordered probit models with class-probability functions. *Analytic Methods in Accident Research* **18**, 15-32.

Frändberg, Lotta, and Bertil Vilhelmson (2003) Personal Mobility: A Corporeal Dimension of Transnationalisation. The Case of Long-Distance Travel from Sweden. *Environment and Planning A: Economy and Space* **35**(**10**), 1751-1768.

Friedman, Jerome H. (2001) Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics* **29**(**5**), 1189-1232.

Fu, Xuemei (2020) How habit moderates the commute mode decision process: integration of the theory of planned behavior and latent class choice model. *Transportation*. https://doi.org/10.1007/s11116-020-10144-6.

Fu, Xuemei, and Zhicai Juan (2017) Accommodating preference heterogeneity in commuting mode choice: an empirical investigation in Shaoxing, China. *Transportation Planning and Technology* **40**(**4**), 434-448.

Gong, Shuangqing, Ali Ardeshiri, and Taha Hossein Rashidi (2020) Impact of government incentives on the market penetration of electric vehicles in Australia. *Transportation Research Part D: Transport and Environment* **83**, 102353.

Goodfellow, Ian, Yoshua Bengio, Aaron Courville, and Yoshua Bengio (2016) *Deep learning*. Vol. 1: MIT press Cambridge.

Gopinath, Dinesh Ambat (1995) "Modeling heterogeneity in discrete choice processes: application to travel demand." Massachusetts Institute of Technology.

Gössling, Stefan, Paul Hanna, James Higham, Scott Cohen, and Debbie Hopkins (2019) Can we fly less? Evaluating the 'necessity' of air travel. *Journal of Air Transport Management* **81**, 101722.

Gössling, Stefan, Andreas Humpe, and Thomas Bausch (2020) Does 'flight shame' affect social norms? Changing perspectives on the desirability of air travel in Germany. *Journal of Cleaner Production* **266**, 122015.

Graham, Anne, and David Metz (2017) Limits to air travel growth: The case of infrequent flyers. *Journal of Air Transport Management* **62**, 109-120.

Green, Paul E, Frank J Carmone, and David P Wachspress (1976) Consumer segmentation via latent class analysis. *Journal of Consumer Research* **3**(**3**), 170-174.

Greene, William, Mark N Harris, Bruce Hollingsworth, and Pushkar Maitra (2014) A latent class model for obesity. *Economics Letters* **123**(**1**), 1-5.

Greene, William H (2012) *Econometric Analysis*. 7th ed. U.K. Pearson Education.

Greene, William H (2016) Nlogit version 6.0 reference guide. Econometric Software. *Inc., Castle Hill*.

Greene, William H., and David A. Hensher (2003) A latent class model for discrete choice analysis: contrasts with mixed logit. *Transportation Research Part B: Methodological* **37**(**8**), 681-698.

Guerra, Erick, and Ricardo A. Daziano (2020) Electric vehicles and residential parking in an urban environment: Results from a stated preference experiment. *Transportation Research Part D: Transport and Environment* **79**, 102222.

Gurumurthy, Krishna Murthy, and Kara M. Kockelman (2018) Analyzing the dynamic ride-sharing potential for shared autonomous vehicle fleets using cellphone data from Orlando, Florida. *Computers, Environment and Urban Systems* **71**, 177-185.

Hackbarth, André, and Reinhard Madlener (2016) Willingness-to-pay for alternative fuel vehicle characteristics: A stated choice study for Germany. *Transportation Research Part A: Policy and Practice* **85**, 89-111.

Haghani, Milad, and Majid Sarvi (2017) Stated and revealed exit choices of pedestrian crowd evacuees. *Transportation Research Part B: Methodological* **95**, 238-259.

Han, Yafei (2019) "Neural-embedded discrete choice models." Massachusetts Institute of Technology.

Han, Yafei, Christopher Zegras, Francisco Camara Pereira, and Moshe Ben-Akiva (2020) A neural-embedded choice model: Tastenet-mnl modeling taste heterogeneity with flexibility and interpretability. *arXiv preprint arXiv:2002.00922*.

Hancock, Thomas O., Stephane Hess, Andrew Daly, and James Fox (2020) Using a sequential latent class approach for model averaging: Benefits in forecasting and behavioural insights. *Transportation Research Part A: Policy and Practice* **139**, 429-454.

Harb, Mustapha, Yu Xiao, Giovanni Circella, Patricia L. Mokhtarian, and Joan L. Walker (2018) Projecting travelers into a world of self-driving vehicles: estimating travel behavior implications via a naturalistic experiment. *Transportation **45(6)**, 1671-1685.*

Hares, Andrew, Janet Dickinson, and Keith Wilkes (2010) Climate change and the air travel decisions of UK tourists. *Journal of Transport Geography* **18**(**3**), 466-473.

Harvey, Chester, Lisa Aultman-Hall, Jeffrey LaMondia, James Sullivan, Elizabeth Greene, and Chloe Ritter (2015) *Conducting a Longitudinal Survey of Overnight Travel: Methods and Preliminary Findings*. Available at https://rosap.ntl.bts.gov/view/dot/28939.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman (2009) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*: Springer Science & Business Media.

Heckman, J., and B. Singer (1984) A Method for Minimizing the Impact of Distributional Assumptions in Econometric Models for Duration Data. *Econometrica* **52**(**2**), 271-320.

Heckman, James (2001) Micro Data, Heterogeneity, and the Evaluation of Public Policy: Nobel Lecture. *Journal of Political Economy* **109**(**4**), 673-748.

Heckman, James, Justin L. Tobias, and Edward Vytlacil (2001) Four Parameters of Interest in the Evaluation of Social Programs. *Southern Economic Journal* **68**(**2**), 211-223.

Heckman, James J. (1979) Sample Selection Bias as a Specification Error. *Econometrica* **47**(**1**), 153-161.

Henao, Alejandro, and Wesley E. Marshall (2019) The impact of ride-hailing on vehicle miles traveled. *Transportation* **46**(**6**), 2173-2194.

Hensher, David (2014) "Attribute Processing as a Behavioural Strategy in Choice Making." In *Handbook of choice modelling*. Edward Elgar Publishing.

Hensher, David A., Camila Balbontin, and Andrew T. Collins (2018) Heterogeneity in decision processes: Embedding extremeness aversion, risk attitude and perceptual conditioning in multiple process rules choice making. *Transportation Research Part A: Policy and Practice* **111**, 316-325.

Hensher, David A., Andrew T. Collins, and William H. Greene (2013) Accounting for attribute non-attendance and common-metric aggregation in a probabilistic decision process mixed multinomial logit model: a warning on potential confounding. *Transportation* **40**(**5**), 1003-1020.

Hensher, David A., and William H. Greene (2003) The Mixed Logit model: The state of practice. *Transportation* **30**(**2**), 133-176.

Hensher, David A., and William H. Greene (2010) Non-attendance and dual processing of common-metric attributes in choice analysis: a latent class specification. *Empirical Economics* **39**(**2**), 413-426.

Hensher, David A., John M. Rose, and William H. Greene (2015) *Applied Choice Analysis: A Primer*. Cambridge, UK: Cambridge University Press.

Hess, Stephane (2014) "Latent class structures: taste heterogeneity and beyond." In *Handbook of Choice Modelling*. Edward Elgar Publishing.

Hess, Stephane, and David Palma (2019) Apollo: A flexible, powerful and customisable freeware package for choice model estimation and application. *Journal of choice modelling* **32**, 100170.

Hess, Stephane, and John W. Polak (2006) Exploring the potential for cross-nesting structures in airport-choice analysis: A case-study of the Greater London area. *Transportation Research Part E: Logistics and Transportation Review* **42**(**2**), 63-81.

Hess, Stephane, Jeremy Shires, and Peter Bonsall (2013) A latent class approach to dealing with respondent uncertainty in a stated choice survey for fare simplification on bus journeys. *Transportmetrica A: Transport Science* **9**(**6**), 473-493.

Hess, Stephane, Greg Spitz, Mark Bradley, and Matt Coogan (2018) Analysis of mode choice for intercity travel: Application of a hybrid choice model to two distinct US corridors. *Transportation Research Part A: Policy and Practice* **116**, 547-567.

Hess, Stephane, Amanda Stathopoulos, Danny Campbell, Vikki O'Neill, and Sebastian Caussade (2013b) It's not that I don't care, I just don't care very much: confounding between attribute non-attendance and taste heterogeneity. *Transportation* **40**(**3**), 583-607.

Hess, Stephane, Amanda Stathopoulos, and Andrew Daly (2012) Allowing for heterogeneous decision rules in discrete choice models: an approach and four case studies. *Transportation* **39**(**3**), 565-591.

Hetrakul, Pratt, and Cinzia Cirillo (2013) Accommodating taste heterogeneity in railway passenger choice models based on internet booking data. *Journal of Choice Modelling* **6**, 1-16.

Hillel, Tim, Michel Bierlaire, Mohammed Z. E. B. Elshafie, and Ying Jin (2021) A systematic review of machine learning classification methodologies for modelling passenger mode choice. *Journal of Choice Modelling* **38**, 100221.

Hoeting, Jennifer A., David Madigan, Adrian E. Raftery, and Chris T. Volinsky (1999) Bayesian Model Averaging: A Tutorial. *Statistical Science* **14**(**4**), 382-401.

Hoijtink, Herbert (2001) Confirmatory Latent Class Analysis: Model Selection Using Bayes Factors and (Pseudo) Likelihood Ratio Statistics. *Multivariate Behavioral Research* **36**(**4**), 563-588.

Holz-Rau, Christian, Joachim Scheiner, and Kathrin Sicks (2014) Travel Distances in Daily Travel and Long-Distance Travel: What Role is Played by Urban Form? *Environment and Planning A: Economy and Space* **46**(**2**), 488-507.

Hornik, Kurt, Maxwell Stinchcombe, and Halbert White (1989) Multilayer feedforward networks are universal approximators. *Neural Networks* **2**(**5**), 359-366.

Hruschka, Harald (2007) Using a heterogeneous multinomial probit model with a neural net extension to model brand choice. *Journal of Forecasting* **26**(**2**), 113-127.

Hruschka, Harald, Werner Fettes, and Markus Probst (2004) An empirical comparison of the validity of a neural net based multinomial logit choice model to alternative model specifications. *European Journal of Operational Research* **159**(**1**), 166-180.

Hruschka, Harald, Werner Fettes, Markus Probst, and Christian Mies (2002) A flexible brand choice model based on neural net methodologyA comparison to the linear utility multinomial logit model and its latent class extension. *OR Spectrum* **24**(**2**), 127-143.

Hurtubia, Ricardo, My Hang Nguyen, Aurélie Glerum, and Michel Bierlaire (2014) Integrating psychometric indicators in latent class choice models. *Transportation Research Part A: Policy and Practice* **64**, 135-146.

Hwang, Yeong-Hyeon, and Daniel R. Fesenmaier (2003) Multidestination Pleasure Travel Patterns: Empirical Evidence from the American Travel Survey. *Journal of Travel Research* **42**(**2**), 166-171.

Hymel, Kent M., Kenneth A. Small, and Kurt Van Dender (2010) Induced demand and rebound effects in road transport. *Transportation Research Part B: Methodological* **44**(**10**), 1220-1241.

Ishaq, Robert, Shlomo Bekhor, and Yoram Shiftan (2014) A latent class model with fuzzy segmentation and weighted variables. *Transportmetrica A: Transport Science* **10**(**10**), 878-893.

Jacobs, Robert A, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton (1991a) Adaptive mixtures of local experts. *Neural computation* **3**(**1**), 79-87.

Jacobs, Robert A., Michael I. Jordan, and Andrew G. Barto (1991b) Task decomposition through competition in a modular connectionist architecture: The what and where vision tasks. *Cognitive Science* **15**(**2**), 219-250.

Jahanshahi, Kaveh, and Ying Jin (2020) Identification and mapping of spatial variations in travel choices through combining structural equation modelling and latent class analysis: findings for Great Britain. *Transportation*.

Jedidi, Kamel, Harsharanjeet S. Jagpal, and Wayne S. DeSarbo (1997a) Finite-Mixture Structural Equation Models for Response-Based Segmentation and Unobserved Heterogeneity. *Marketing Science* **16**(**1**), 39-59.

Jedidi, Kamel, Harsharanjeet S. Jagpal, and Wayne S. DeSarbo (1997b) STEMM: A General Finite Mixture Structural Equation Model. *Journal of Classification* **14**(**1**), 23-50.

Jin, Wen, Yinglu Deng, Hai Jiang, Qianyan Xie, Wei Shen, and Weijian Han (2018) Latent class analysis of accident risks in usage-based insurance: Evidence from Beijing. *Accident Analysis & Prevention* **115**, 79-88.

Jöreskog, K. G. (1971) Simultaneous factor analysis in several populations. *Psychometrika* **36**(**4**), 409-426.

Jou, Rong-Chang, Rong-Yi Jian, and Yuan-Chan Wu (2013) Behavior of Passengers Regarding the Purchase of Business Class Seats on the High-Speed Rail in Taiwan: Application of the Hurdle Model. *International Journal of Sustainable Transportation* **7**(**6**), 468-479.

Kamakura, Wagner A, and Gary Russell (1989) A probabilistic choice model for market segmentation and elasticity structure. *Journal of Marketing Research* **26**(**3**), 379-390.

Ke, Yue, and B. Starr McMullen (2017) Regional differences in the determinants of Oregon VMT. *Research in Transportation Economics* **62**, 2-10.

Keskisaari, Ville, Juudit Ottelin, and Jukka Heinonen (2017) Greenhouse gas impacts of different modality style classes using latent class travel behavior model. *Journal of Transport Geography* **65**, 155-164.

Khan, Nazmul Arefin, Muhammad Ahsanul Habib, and Shaila Jamal (2020) Effects of smartphone application usage on mobility choices. *Transportation Research Part A: Policy and Practice* **132**, 932-947.

Kim, Seheon, Dujuan Yang, Soora Rasouli, and Harry Timmermans (2017) Heterogeneous hazard model of PEV users charging intervals: Analysis of four year charging transactions data. *Transportation Research Part C: Emerging Technologies* **82**, 248-260.

Kim, Sung Hoo, and Jin-Hyuk Chung (2016) Reinterpretation of the Likert scale for public transportation user satisfaction: Pattern recognition approach. *Transportation Research Record: Journal of the Transportation Research Board*(**2541**), 90-99.

Kim, Sung Hoo, Jin-Hyuk Chung, Sunyoung Park, and Keechoo Choi (2017) Analysis of user satisfaction to promote public transportation: A pattern-recognition approach focusing on out-of-vehicle time. *International Journal of Sustainable Transportation* **11**(**8**), 582-592.

Kim, Sung Hoo, Giovanni Circella, and Patricia L. Mokhtarian (2019a) Identifying latent mode-use propensity segments in an all-AV era. *Transportation Research Part A: Policy and Practice* **130**, 192-207.

Kim, Sung Hoo, and Patricia L. Mokhtarian (2018) Taste heterogeneity as an alternative form of endogeneity bias: Investigating the attitude-moderated effects of built environment and socio-demographics on vehicle ownership using latent class modeling. *Transportation Research Part A: Policy and Practice* **116**, 130-150.

Kim, Sung Hoo, P. L. Mokhtarian, and Giovanni Circella (2019b) *The Impact of Emerging Technologies and Trends on Travel Demand in Georgia*: Georgia Department of Transportation.

Kim, Sung Hoo, Patricia L. Mokhtarian, and Giovanni Circella (2020a) How, and for whom, will activity patterns be modified by self-driving cars? Expectations from the state of Georgia. *Transportation Research Part F: Traffic Psychology and Behaviour* **70**, 68-80.

Kim, Sung Hoo, Patricia L. Mokhtarian, and Giovanni Circella (2020b) Will autonomous vehicles change residential location and vehicle ownership? Glimpses from Georgia. *Transportation Research Part D: Transport and Environment* **82**, 102291.

Kim, Sung Hoo, and Patricia L. Mokhtarian (2021) Who (never) makes overnight leisure trips? Disentangling structurally zero trips from usual trip generation processes. *Travel Behaviour and Society* **25**, 78-91.

Kormos, Christine, Jonn Axsen, Zoe Long, and Suzanne Goldberg (2019) Latent demand for zero-emissions vehicles in Canada (Part 2): Insights from a stated choice experiment. *Transportation Research Part D: Transport and Environment* **67**, 685-702.

Kotler, Philip, and Gary Armstrong (2010) *Principles of Marketing*. Pearson education.

Kotsiantis, S. B., I. D. Zaharakis, and P. E. Pintelas (2006) Machine learning: a review of classification and combining techniques. *Artificial Intelligence Review* **26**(**3**), 159-190.

Koutsopoulos, Haris N., and Haneen Farah (2012) Latent class model for car following behavior. *Transportation Research Part B: Methodological* **46**(**5**), 563-578.

Kroesen, Maarten (2015) Do partners influence each other's travel patterns? A new approach to study the role of social norms. *Transportation Research Part A: Policy and Practice* **78**, 489-505.

Kroesen, Maarten (2019) Is active travel part of a healthy lifestyle? Results from a latent class analysis. *Journal of Transport & Health* **12**, 42-49.

Krueger, Rico, Akshay Vij, and Taha H. Rashidi (2018) Normative beliefs and modality styles: a latent class and latent variable model of travel behaviour. *Transportation* **45**(**3**), 789-825.

Kuhn, Kenneth D. (2018) Using structural topic modeling to identify latent topics and trends in aviation incident reports. *Transportation Research Part C: Emerging Technologies* **87**, 105-122.

Lambert, Diane (1992) Zero-Inflated Poisson Regression, With an Application to Defects in Manufacturing. *Technometrics* **34**(**1**), 1-14.

LaMondia, Jeffrey J., Lisa Aultman-Hall, and Elizabeth Greene (2014) Long-Distance Work and Leisure Travel Frequencies:Ordered Probit Analysis Across Non–Distance-Based Definitions. *Transportation Research Record* **2413**(**1**), 1-12.

LaMondia, Jeffrey J, Daniel J Fagnant, Hongyang Qu, Jackson Barrett, and Kara Kockelman (2016) Long-Distance Travel Mode-Shifts Due to Automated Vehicles: A Statewide Mode-Shift Simulation Experiment and Travel Survey Analysis. Transportation Research Board 95th Annual Meeting.

LaMondia, Jeffrey J., Michael Moore, and Lisa Aultman-Hall (2015) Modeling Intertrip Time Intervals between Individuals' Overnight Long-Distance Trips. *Transportation Research Record* **2495**(**1**), 23-31.

Lannoo, Steven, Veronique Van Acker, Roselinde Kessels, Daniel Palhazi Cuervo, and Frank Witlox (2018) Getting Business People on the Coach: A Stated Preference Experiment for Intercity Long Distance Coach Travel. *Transportation Research Record* **2672**(**8**), 165-174.

Lanza, Stephanie T, Linda M Collins, David R Lemmon, and Joseph L Schafer (2007) PROC LCA: A SAS procedure for latent class analysis. *Structural equation modeling: a multidisciplinary journal* **14**(**4**), 671-694.

Lanza, Stephanie T, John J Dziak, Liying Huang, Aaron T Wagner, and Linda M Collins (2015) LCA Stata plugin users' guide (Version 1.2). *University Park: The Methodology Center, Penn State*.

Laudy, Olav, Mark Zoccolillo, Raymond H. Baillargeon, Jan Boom, Richard E. Tremblay, and Herbert Hoijtink (2005) Applications of confirmatory latent class analysis in developmental psychology. *European Journal of Developmental Psychology* **2**(**1**), 1-15.

Lavieri, Patrícia S., and Chandra R. Bhat (2019) Investigating objective and subjective factors influencing the adoption, frequency, and characteristics of ride-hailing trips. *Transportation Research Part C: Emerging Technologies* **105**, 100-125.

Lavieri, Patrícia S, Venu M Garikapati, Chandra R Bhat, Ram M Pendyala, Sebastian Astroza, and Felipe F Dias (2017) Modeling individual preferences for ownership and sharing of autonomous vehicle technologies. *Transportation Research Record: Journal of the Transportation Research Board*(**2665**), 1-10.

Lazarsfeld, Paul F (1950) The logical and mathematical foundation of latent structure analysis. *Studies in Social Psychology in World War II Vol. IV: Measurement and Prediction*, 362-412.

Leisch, Friedrich (2004) Flexmix: A general framework for finite mixture models and latent glass regression in R.

Leung, Siu Fai, and Shihti Yu (1996) On the choice between sample selection and two-part models. *Journal of Econometrics* **72**(**1**), 197-229.

Li, Binbin, Enjian Yao, Toshiyuki Yamamoto, Ying Tang, and Shasha Liu (2020) Exploring behavioral heterogeneities of metro passenger's travel plan choice under unplanned service disruption with uncertainty. *Transportation Research Part A: Policy and Practice* **141**, 294-306.

Li, Gen (2018) Application of Finite Mixture of Logistic Regression for Heterogeneous Merging Behavior Analysis. *Journal of Advanced Transportation* **2018**, 1436521.

Li, Juan, Boyu Jiang, Chunjiao Dong, Jue Wang, and Xuan Zhang (2020) Analysis of Driver Decisions at the Onset of Yellow at Signalized Intersections. *Journal of Advanced Transportation* **2020**, 2023093.

Li, Zhenning, Cong Chen, Qiong Wu, Guohui Zhang, Cathy Liu, Panos D. Prevedouros, and David T. Ma (2018) Exploring driver injury severity patterns and causes in low visibility related single-vehicle crashes using a finite mixture random parameters model. *Analytic Methods in Accident Research* **20**, 1-14.

Liao, Fanchao, Eric Molin, Harry Timmermans, and Bert van Wee (2020) Carsharing: the impact of system characteristics on its potential to replace private car trips and reduce car ownership. *Transportation* **47**(**2**), 935-970.

Lin, Jen-Jia, Nai-Ling Wang, and Cheng-Min Feng (2017) Public bike system pricing and usage in Taipei. *International Journal of Sustainable Transportation* **11**(**9**), 633-641.

Lin, Jen-Jia, Pengjun Zhao, Kazuyuki Takada, Shengxiao Li, Tetsuo Yai, and Chi-Hao Chen (2018) Built environment and public bike usage for metro access: A comparison of neighborhoods in Beijing, Taipei, and Tokyo. *Transportation Research Part D: Transport and Environment* **63**, 209-221.

Linzer, Drew A, and Jeffrey B Lewis (2011) poLCA: An R package for polytomous variable latent class analysis. *Journal of statistical software* **42**(**10**), 1-29.

Little, Roderick JA, and Donald B Rubin (2019) *Statistical Analysis with Missing Data*. Vol. 793: John Wiley & Sons.

Lord, Dominique, and Fred Mannering (2010) The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives. *Transportation Research Part A: Policy and Practice* **44**(**5**), 291-305.

Ma, Lu, Guan Wang, Xuedong Yan, and Jinxian Weng (2016) A hybrid finite mixture model for exploring heterogeneous ordering patterns of driver injury severity. *Accident Analysis & Prevention* **89**, 62-73.

Ma, Tai-Yu, Philippe Gerber, Samuel Carpentier, and Sylvain Klein (2015) Mode choice with latent preference heterogeneity: a case study for employees of the EU institutions in Luxembourg. *Transportmetrica A: Transport Science* **11**(**5**), 441-463.

Machado, Jose Luis, Rocio de Oña, Francisco Diez-Mesa, and Juan de Oña (2018) Finding service quality improvement opportunities across different typologies of public transit customers. *Transportmetrica A: Transport Science* **14**(**9**), 761-783.

Maddala, Gangadharrao S (1986) Disequilibrium, self-selection, and switching models. *Handbook of econometrics* **3**, 1633-1688.

Madden, David (2008) Sample selection versus two-part models revisited: The case of female smoking and drinking. *Journal of Health Economics* **27**(**2**), 300-307.

Mahmud, S. M. Sohel, Luis Ferreira, Md Shamsul Hoque, and Ahmad Tavassoli (2019) Micro-level safety risk assessment model for a two-lane heterogeneous traffic environment in a developing country: A comparative crash probability modeling approach. *Journal of Safety Research* **69**, 125-134.

Maness, Michael, and Cinzia Cirillo (2016) An indirect latent informational conformity social influence choice model: Formulation and case study. *Transportation Research Part B: Methodological* **93**, 75-101.

Mannering, Fred, Chandra R. Bhat, Venky Shankar, and Mohamed Abdel-Aty (2020) Big data, traditional data and the tradeoffs between prediction and causality in highway-safety analysis. *Analytic Methods in Accident Research* **25**, 100113.

Mannering, Fred L., and Chandra R. Bhat (2014) Analytic methods in accident research: Methodological frontier and future directions. *Analytic Methods in Accident Research* **1**, 1-22.

Mannering, Fred L., Venky Shankar, and Chandra R. Bhat (2016) Unobserved heterogeneity and the statistical analysis of highway accident data. *Analytic Methods in Accident Research* **11**, 1-16.

Manning, W. G., N. Duan, and W. H. Rogers (1987) Monte Carlo evidence on the choice between sample selection and two-part models. *Journal of Econometrics* **35**(**1**), 59-82.

Martin, Charles Patrick, and Jim Torresen (2018) RoboJam: A musical mixture density network for collaborative touchscreen interaction. International Conference on Computational Intelligence in Music, Sound, Art and Design.

Masoudnia, Saeed, and Reza Ebrahimpour (2014) Mixture of experts: a literature survey. *Artificial Intelligence Review* **42**(**2**), 275-293.

Masyn, Katherine E (2013) Latent class analysis and finite mixture modeling. *The Oxford handbook of quantitative methods*, 551.

McFadden, Daniel (2001) Economic choices. *American Economic Review* **91**(**3**), 351-378.

McFadden, Daniel, and Kenneth Train (2000) Mixed MNL Models for Discrete Response. *Journal of Applied Econometrics* **15**(**5**), 447-470.

McLachlan, Geoffrey J, and David Peel (2004) *Finite Mixture Models*: John Wiley & Sons.

Mehadil Orvin, Muntahith, and Mahmudur Rahman Fatmi (2020) Modeling Destination Choice Behavior of the Dockless Bike Sharing Service Users. *Transportation Research Record* **2674**(**11**), 875-887.

Meredith, William (1964) Notes on factorial invariance. *Psychometrika* **29**(**2**), 177-185.

Mokhtarian, Patricia L., and Xinyu Cao (2008) Examining the impacts of residential self-selection on travel behavior: A focus on methodologies. *Transportation Research Part B: Methodological* **42**(**3**), 204-228.

Mokhtarian, Patricia L., Susan L. Handy, and Ilan Salomon (1995) Methodological issues in the estimation of the travel, energy, and air quality impacts of telecommuting. *Transportation Research Part A: Policy and Practice* **29**(**4**), 283-302.

Mokhtarian, Patricia L., Ilan Salomon, and Lothlorien S. Redmond (2001) Understanding the Demand for Travel: It's Not Purely 'Derived'. *Innovation: The European Journal of Social Science Research* **14**(**4**), 355-380.

Mokhtarian, Patricia L., and David van Herick (2016) Quantifying residential self-selection effects: A review of methods and findings from applications of propensity score and sample selection approaches. *Journal of Transport and Land Use* **9**(**1**), 9-28.

Molesworth, Brett R. C., and Tay T. R. Koo (2016) The influence of attitude towards individuals' choice for a remotely piloted commercial flight: A latent class logit approach. *Transportation Research Part C: Emerging Technologies* **71**, 51-62.

Molin, Eric, and Kees Maat (2015) Bicycle parking demand at railway stations: Capturing price-walking trade offs. *Research in Transportation Economics* **53**, 3-12.

Molin, Eric, Patricia Mokhtarian, and Maarten Kroesen (2016) Multimodal travel groups and attitudes: A latent class cluster analysis of Dutch travelers. *Transportation Research Part A: Policy and Practice* **83**, 14-29.

Molina, Mario, and Filiz Garip (2019) Machine Learning for Sociology. *Annual Review of Sociology* **45**(**1**), 27-45.

Molnar, Christoph (2020) *Interpretable Machine Learning*. Available at https://christophm.github.io/interpretable-ml-book/.

Monchambert, Guillaume (2020) Why do (or don't) people carpool for long distance trips? A discrete choice experiment in France. *Transportation Research Part A: Policy and Practice* **132**, 911-931.

Motoaki, Yutaka, and Ricardo A. Daziano (2015) A hybrid-choice latent-class model for the analysis of the effects of weather on cycling demand. *Transportation Research Part A: Policy and Practice* **75**, 217-230.

Mouter, Niek, Sander van Cranenburgh, and Bert van Wee (2017) An empirical assessment of Dutch citizens' preferences for spatial equality in the context of a national transport investment plan. *Journal of Transport Geography* **60**, 217-230.

Mullahy, John (1986) Specification and testing of some modified count data models. *Journal of Econometrics* **33**(**3**), 341-365.

Mullainathan, Sendhil, and Jann Spiess (2017) Machine Learning: An Applied Econometric Approach. *Journal of Economic Perspectives* **31**(**2**), 87-106.

Muthén, Bengt O. (1989) Latent variable modeling in heterogeneous populations. *Psychometrika* **54**(**4**), 557-585.

Muthén, Linda K, and Bengt O Muthén (2017) *Mplus: Statistical analysis with latent variables: User's guide*: Muthén & Muthén Los Angeles.

Nayum, Alim, Christian A. Klöckner, and Sunita Prugsamatz (2013) Influences of car type class and carbon dioxide emission levels on purchases of new cars: A retrospective analysis of car purchases in Norway. *Transportation Research Part A: Policy and Practice* **48**, 96-108.

Nazari, Fatemeh, Mohamadhossein Noruzoliaee, and Abolfazl Mohammadian (2018) Shared versus private mobility: Modeling public interest in autonomous vehicles accounting for latent attitudes. *Transportation Research Part C: Emerging Technologies* **97**, 456-477.

Nguyen, Hien D., Luke R. Lloyd-Jones, and Geoffrey J. McLachlan (2016) A Universal Approximation Theorem for Mixture-of-Experts Models. *Neural Computation* **28**(**12**), 2585-2593.

NHTS (2018) *Summary of Travel Trends: 2017 National Household Travel Survey*. NHTS. Available at https://nhts.ornl.gov/assets/2017_nhts_summary_travel_trends.pdf.

Olaru, Doina, Brett Smith, and John H. E. Taplin (2011) Residential location and transit-oriented development in a new rail corridor. *Transportation Research Part A: Policy and Practice* **45**(**3**), 219-237.

Oliva, Ignacio, Patricia Galilea, and Ricardo Hurtubia (2018) Identifying cycling-inducing neighborhoods: A latent class approach. *International Journal of Sustainable Transportation* **12**(**10**), 701-713.

Orrù, Graziella, Merylin Monaro, Ciro Conversano, Angelo Gemignani, and Giuseppe Sartori (2020) Machine Learning in Psychometrics and Psychological Research. *Frontiers in Psychology* **10**(**2970**).

Ortelli, Nicola, Tim Hillel, Francisco Camara Pereira, Matthieu de Lapparent, and Michel Bierlaire (2021) Assisted specification of discrete choice models. *Journal of Choice Modelling*, 100285.

Ottelin, Juudit, Jukka Heinonen, and Seppo Junnila (2014) Greenhouse gas emissions from flying can offset the gain from reduced driving in dense urban areas. *Journal of Transport Geography* **41**, 1-9.

Pan, Long, Enjian Yao, and Don MacKenzie (2019) Modeling EV charging choice considering risk attitudes and attribute non-attendance. *Transportation Research Part C: Emerging Technologies* **102**, 60-72.

Pani, Agnivesh, Sabya Mishra, Mihalis Golias, and Miguel Figliozzi (2020) Evaluating public acceptance of autonomous delivery robots during COVID-19 pandemic. *Transportation Research Part D: Transport and Environment* **89**, 102600.

Park, Byung-Jung, Dominique Lord, and Lingtao Wu (2016) Finite mixture modeling approach for developing crash modification factors in highway safety analysis. *Accident Analysis & Prevention* **97**, 274-287.

Payne, John W, James R Bettman, and Eric J Johnson (1992) Behavioral decision research: A constructive processing perspective. *Annual review of psychology* **43**(**1**), 87-131.

Pearson, Karl (1894) Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London. A* **185**, 71-110.

Peer, Stefanie, Jasper Knockaert, Paul Koster, and Erik T. Verhoef (2014) Over-reporting vs. overreacting: Commuters' perceptions of travel times. *Transportation Research Part A: Policy and Practice* **69**, 476-494.

Perrine, Kenneth A., Kara M. Kockelman, and Yantao Huang (2020) Anticipating long-distance travel shifts due to self-driving vehicles. *Journal of Transport Geography* **82**, 102547.

Piendl, Raphael, Tilman Matteis, and Gernot Liedtke (2019) A machine learning approach for the operationalization of latent classes in a discrete shipment size choice model. *Transportation Research Part E: Logistics and Transportation Review* **121**, 149-161.

Pinjari, Abdul Rawoof, Naveen Eluru, Chandra R. Bhat, Ram M. Pendyala, and Erika Spissu (2008) Joint Model of Choice of Residential Neighborhood and Bicycle Ownership: Accounting for Self-Selection and Unobserved Heterogeneity. *Transportation Research Record* **2082**(**1**), 17-26.

Polzin, Steven E., and Xuehao Chu (2014) Peak Vehicle Miles Traveled and Postpeak Consequences? *Transportation Research Record* **2453**(**1**), 22-29.

Potoglou, Dimitris, Colin Whittle, Ioannis Tsouros, and Lorraine Whitmarsh (2020) Consumer intentions for alternative fuelled and autonomous vehicles: A segmentation analysis across six countries. *Transportation Research Part D: Transport and Environment* **79**, 102243.

Prato, Carlo Giacomo, Katrín Halldórsdóttir, and Otto Anker Nielsen (2017) Latent lifestyle and mode choice decisions when travelling short distances. *Transportation* **44**(**6**), 1343-1363.

Prato, Carlo G., Sigal Kaplan, Alexandre Patrier, and Thomas K. Rasmussen (2019) Integrating police reports with geographic information system resources for uncovering patterns of pedestrian crashes in Denmark. *Journal of Transport Geography* **74**, 10-23.

Qin, Huanmei, Jianqiang Gao, Hongzhi Guan, and Hongbo Chi (2017) Estimating heterogeneity of car travelers on mode shifting behavior based on discrete choice models. *Transportation Planning and Technology* **40**(**8**), 914-927.

Rahmani, Djamel, and Maria L. Loureiro (2019) Assessing drivers' preferences for hybrid electric vehicles (HEV) in Spain. *Research in Transportation Economics* **73**, 89-97.

Ralph, Kelcie, Carole Turley Voulgaris, Brian D. Taylor, Evelyn Blumenberg, and Anne E. Brown (2016) Millennials, built form, and travel insights from a nationwide typology of U.S. neighborhoods. *Journal of Transport Geography* **57**, 218-226.

Ramos, Juan (2003) Using tf-idf to determine word relevance in document queries. Proceedings of the first instructional conference on machine learning.

Razo, Michael, and Song Gao (2013) A rank-dependent expected utility model for strategic route choice with stated preference data. *Transportation Research Part C: Emerging Technologies* **27**, 117-130.

Richmond, Korin (2007) Trajectory mixture density networks with multiple mixtures for acoustic-articulatory inversion. International Conference on Nonlinear Speech Processing.

Rokach, Lior (2010) Ensemble-based classifiers. *Artificial Intelligence Review* **33**(**1**), 1-39.

Rossetti, Tomás, Verónica Saud, and Ricardo Hurtubia (2019) I want to ride it where I like: measuring design preferences in cycling infrastructure. *Transportation* **46**(**3**), 697-718.

Rummel, Rudolf J (1970) *Applied Factor Analysis*. Evanston, IL: Northwestern University Press.

Salon, Deborah (2015) Heterogeneity in the relationship between the built environment and driving: Focus on neighborhood type and travel purpose. *Research in Transportation Economics* **52**, 34-45.

Savolainen, Peter T. (2016) Examining driver behavior at the onset of yellow in a traffic simulator environment: Comparisons between random parameters and latent class logit models. *Accident Analysis & Prevention* **96**, 300-307.

Saxena, Neeraj, Taha Hossein Rashidi, and Joshua Auld (2019) Studying the tastes effecting mode choice behavior of travelers under transit service disruptions. *Travel Behaviour and Society* **17**, 86-95.

Scarpa, Riccardo, and Mara Thiene (2005) Destination choice models for rock climbing in the Northeastern Alps: a latent-class approach based on intensity of preferences. *Land economics* **81**(**3**), 426-444.

Schmidt, Alejandro, Juan de Dios Ortúzar, and Ricardo D. Paredes (2019) Heterogeneity and college choice: Latent class modelling for improved policy making. *Journal of Choice Modelling* **33**, 100185.

Scrucca, Luca, Michael Fop, T Brendan Murphy, and Adrian E Raftery (2016) mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. *The R journal* **8**(**1**), 289.

Seelhorst, Michael, and Yi Liu (2015) Latent air travel preferences: Understanding the role of frequent flyer programs on itinerary choice. *Transportation Research Part A: Policy and Practice* **80**, 49-61.

Shahnaz, Farial, Michael W. Berry, V. Paul Pauca, and Robert J. Plemmons (2006) Document clustering using nonnegative matrix factorization. *Information Processing & Management* **42**(**2**), 373-386.

Shmueli, Galit (2010) To explain or to predict? *Statistical Science* **25**(**3**), 289-310.

Sifringer, Brian, Virginie Lurkin, and Alexandre Alahi (2020) Enhancing discrete choice models with representation learning. *Transportation Research Part B: Methodological* **140**, 236-261.

Simons-Morton, Bruce G., Kyeongmi Cheon, Feng Guo, and Paul Albert (2013) Trajectories of kinematic risky driving among novice teenagers. *Accident Analysis & Prevention* **51**, 27-32.

Singh, Abhilash C., Sebastian Astroza, Venu M. Garikapati, Ram M. Pendyala, Chandra R. Bhat, and Patricia L. Mokhtarian (2018) Quantifying the relative contribution of factors to household vehicle miles of travel. *Transportation Research Part D: Transport and Environment* **63**, 23-36.

Smith, Wendell R (1956) Product differentiation and market segmentation as alternative marketing strategies. *Journal of marketing* **21**(**1**), 3-8.

Sobhani, Anae, Naveen Eluru, and Ahmadreza Faghih-Imani (2013) A latent segmentation based multiple discrete continuous extreme value model. *Transportation Research Part B: Methodological* **58**, 154-169.

Soria, Jason, Ying Chen, and Amanda Stathopoulos (2020) K-Prototypes Segmentation Analysis on Large-Scale Ridesourcing Trip Data. *Transportation Research Record* **2674**(**9**), 0361198120929338.

Srinivasan, Karthik K., G. Maheswara Naidu, and Tejaswi Sutrala (2009) Heterogeneous Decision Rule Model of Mode Choice Incorporating Utility Maximization and Disutility Minimization. *Transportation Research Record* **2132**(**1**), 59-68.

Strandell, Anna, and C. Michael Hall (2015) Impact of the residential environment on second home use in Finland – Testing the compensation hypothesis. *Landscape and Urban Planning* **133**, 12-23.

Sun, Lijun, and Yafeng Yin (2017) Discovering themes and trends in transportation research using topic modeling. *Transportation Research Part C: Emerging Technologies* **77**, 49-66.

Sun, Zhongwei, Theo Arentze, and Harry Timmermans (2012) A heterogeneous latent class model of activity rescheduling, route choice and information acquisition decisions under multiple uncertain events. *Transportation Research Part C: Emerging Technologies* **25**, 46-60.

Swait, Joffre (1994) A structural equation model of latent segmentation and product choice for cross-sectional revealed preference choice data. *Journal of Retailing and Consumer Services* **1**(**2**), 77-89.

Tang, Wei, and Patricia L. Mokhtarian (2009) Accounting for Taste Heterogeneity in Purchase Channel Intention Modeling: An Example from Northern California for Book Purchases. *Journal of Choice Modelling* **2**(**2**), 148-172.

Tawfik, Aly M., and Hesham A. Rakha (2013) Latent Class Choice Model of Heterogeneous Drivers' Route Choice Behavior Based on Learning in a Real-World Experiment. *Transportation Research Record* **2334**(**1**), 84-94.

Teichert, Thorsten, Edlira Shehu, and Iwan von Wartburg (2008) Customer segmentation revisited: The case of the airline industry. *Transportation Research Part A: Policy and Practice* **42**(**1**), 227-242.

Thorhauge, Mikkel, Akshay Vij, and Elisabetta Cherchi (2020) Heterogeneity in departure time preferences, flexibility and schedule constraints. *Transportation*.

Tinessa, Fiore, Vittorio Marzano, and Andrea Papola (2020) Mixing distributions of tastes with a Combination of Nested Logit (CoNL) kernel: Formulation and performance analysis. *Transportation Research Part B: Methodological* **141**, 1-23.

Tirachini, Alejandro, and Mariana del Río (2019) Ride-hailing in Santiago de Chile: Users' characterisation and effects on travel behaviour. *Transport Policy* **82**, 46-57.

Tirachini, Alejandro, and Andres Gomez-Lobo (2020) Does ride-hailing increase or decrease vehicle kilometers traveled (VKT)? A simulation approach for Santiago de Chile. *International Journal of Sustainable Transportation* **14**(**3**), 187-204.

Tirachini, Alejandro, Ricardo Hurtubia, Thijs Dekker, and Ricardo A. Daziano (2017) Estimation of crowding discomfort in public transport: Results from Santiago de Chile. *Transportation Research Part A: Policy and Practice* **103**, 311-326.

Tobin, James (1958) Estimation of Relationships for Limited Dependent Variables. *Econometrica* **26**(**1**), 24-36.

Torres, Cati, Nick Hanley, and Antoni Riera (2011) How wrong can you be? Implications of incorrect utility function specification for welfare measurement in choice experiments. *Journal of Environmental Economics and Management* **62**(**1**), 111-121.

Train, Kenneth (2009) *Discrete Choice Methods with Simulation*: Cambridge University Press.

Tran, Yen, Toshiyuki Yamamoto, and Hitomi Sato (2020) The influences of environmentalism and attitude towards physical activity on mode choice: The new evidences. *Transportation Research Part A: Policy and Practice* **134**, 211-226.

UDOT (2013) *Utah Travel Study*. Utah DOT. Available at https://wfrc.org/MapsData/UtahTravelStudy/UtahTravelStudy_FinalReport_130228.pdf.

Urban, Christopher J, and Kathleen M Gates (2021) Deep learning: A primer for psychologists. *Psychological Methods*.

Van Acker, Veronique, Roselinde Kessels, Daniel Palhazi Cuervo, Steven Lannoo, and Frank Witlox (2020) Preferences for long-distance coach transport: Evidence from a discrete choice experiment. *Transportation Research Part A: Policy and Practice* **132**, 759-779.

van Cranenburgh, Sander, and Ahmad Alwosheel (2019) An artificial neural network based approach to investigate travellers' decision rules. *Transportation Research Part C: Emerging Technologies* **98**, 152-166.

van de Coevering, Paul, Kees Maat, and Bert van Wee (2018) Residential self-selection, reverse causality and residential dissonance. A latent class transition model of interactions between the built environment, travel attitudes and travel behavior. *Transportation Research Part A: Policy and Practice* **118**, 466-479.

van der Pol, Marjon, Gillian Currie, Seija Kromm, and Mandy Ryan (2014) Specification of the Utility Function in Discrete Choice Experiments. *Value in Health* **17**(**2**), 297-301.

van der Waerden, Peter, and Jaap van der Waerden (2018) The Relation between Train Access Mode Attributes and Travelers' Transport Mode-Choice Decisions in the Context of Medium- and Long-Distance Trips in the Netherlands. *Transportation Research Record* **2672**(**8**), 719-730.

van Herick, David, and Patricia L. Mokhtarian (2020) How much does the method matter? An empirical comparison of ways to quantify the influence of residential self-selection. *Travel Behaviour and Society* **18**, 68-82.

Vermunt, Jeroen K., and Jay Magidson (2003) Latent class models for classification. *Computational Statistics & Data Analysis* **41**(**3**), 531-537.

Vermunt, Jeroen K, and Jay Magidson (2016) Technical guide for Latent GOLD 5.0: Basic and advanced. *Belmont Massachusetts: Statistical Innovations Inc*.

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

Vij, Akshay, Sreeta Gorripaty, and Joan L. Walker (2017) From trend spotting to trend 'splaining: Understanding modal preference shifts in the San Francisco Bay Area. *Transportation Research Part A: Policy and Practice* **95**, 238-258.

Vij, Akshay, and Rico Krueger (2017) Random taste heterogeneity in discrete choice models: Flexible nonparametric finite mixture distributions. *Transportation Research Part B: Methodological* **106**, 76-101.

Vij, Akshay, and Joan L. Walker (2014) Preference endogeneity in discrete choice models. *Transportation Research Part B: Methodological* **64**, 90-105.

Vij, Akshay, and Joan L. Walker (2016) How, when and why integrated choice and latent variable models are latently useful. *Transportation Research Part B: Methodological* **90**, 192-217.

Wafa, Zeina, Chandra R. Bhat, Ram M. Pendyala, and Venu M. Garikapati (2015) Latent-Segmentation-Based Approach to Investigating Spatial Transferability of Activity-Travel Models. *Transportation Research Record* **2493**(**1**), 136-144.

Wagenmakers, Eric-Jan, Ruud Wetzels, Denny Borsboom, Han L. J. van der Maas, and Rogier A. Kievit (2012) An Agenda for Purely Confirmatory Research. *Perspectives on Psychological Science* **7**(**6**), 632-638.

Wainer, Howard (1999) Visual Revelations. *CHANCE* **12**(**4**), 44-46.

Walker, Joan L. (2001) "Extended discrete choice models: integrated framework, flexible error structures, and latent variables." Massachusetts Institute of Technology.

Walker, Joan L., and Moshe Ben-Akiva (2011) Advances in discrete choice: mixture models. *A handbook of transport economics* **160**.

Walker, Joan L., and Jieping Li (2007) Latent lifestyle preferences and household location decisions. *Journal of Geographical Systems* **9**(**1**), 77-101.

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

Wang, Shenhao, Baichuan Mo, and Jinhua Zhao (2021) Theory-based residual neural networks: A synergy of discrete choice models and deep neural networks. *Transportation Research Part B: Methodological* **146**, 333-358.

Wang, Shenhao, Qingyi Wang, and Jinhua Zhao (2020b) Deep neural networks for choice analysis: Extracting complete economic information for interpretation. *Transportation Research Part C: Emerging Technologies* **118**, 102701.

Wang, Yu, Yacan Wang, and Charisma Choudhury (2020) Modelling heterogeneity in behavioral response to peak-avoidance policy utilizing naturalistic data of Beijing subway travelers. *Transportation Research Part F: Traffic Psychology and Behaviour* **73**, 92-106.

Washington, Simon, Matthew G Karlaftis, Fred Mannering, and Panagiotis Anastasopoulos (2020) *Statistical and Econometric Methods for Transportation Data Analysis*: CRC press.

Wedel, Michel, and Wagner A Kamakura (2012) *Market Segmentation: Conceptual and Methodological Foundations*. Vol. 8: Springer Science & Business Media.

Wen, Chieh-Hua, and Frank S. Koppelman (2001) The generalized nested logit model. *Transportation Research Part B: Methodological* **35**(**7**), 627-641.

Wen, Chieh-Hua, and Shan-Ching Lai (2010) Latent class models of international air carrier choice. *Transportation Research Part E: Logistics and Transportation Review* **46**(**2**), 211-221.

Wen, Chieh-Hua, Wei-Chung Wang, and Chiang Fu (2012) Latent class nested logit model for analyzing high-speed rail access mode choice. *Transportation Research Part E: Logistics and Transportation Review* **48**(**2**), 545-554.

Wen, Yuan, Don MacKenzie, and David R. Keith (2016) Modeling the Charging Choices of Battery Electric Vehicle Drivers by Using Stated Preference Data. *Transportation Research Record* **2572**(**1**), 47-55.

Westat (2018) *NHTS Main Study Retrieval Questionnaire*. US Federal Highway Administration. Available at https://nhts.ornl.gov/assets/2016/NHTS_Retrieval_Instrument_20180228.pdf.

Wind, Yoram (1978) Issues and Advances in Segmentation Research. *Journal of Marketing Research* **15**(**3**), 317-337.

Wolbertus, Rick, and Bas Gerzon (2018) Improving Electric Vehicle Charging Station Efficiency through Pricing. *Journal of Advanced Transportation* **2018**, 4831951.

Wooldridge, Jeffrey M (2015) Control function methods in applied econometrics. *Journal of Human Resources* **50**(**2**), 420-445.

Wu, Amery D., and Bruno D. Zumbo (2007) Understanding and Using Mediators and Moderators. *Social Indicators Research* **87**(**3**), 367.

Xie, Yuanchang, Kaiguang Zhao, and Nathan Huynh (2012) Analysis of driver injury severity in rural single-vehicle crashes. *Accident Analysis & Prevention* **47**, 36-44.

Xu, Wei, Xin Liu, and Yihong Gong (2003) Document clustering based on non-negative matrix factorization. Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval.

Yan, Xiang, Xinyu Liu, and Xilei Zhao (2020) Using machine learning for direct demand modeling of ridesourcing services in Chicago. *Journal of Transport Geography* **83**, 102661.

Yarkoni, Tal, and Jacob Westfall (2017) Choosing Prediction Over Explanation in Psychology: Lessons From Machine Learning. *Perspectives on Psychological Science* **12**(**6**), 1100-1122.

Yasmin, Shamsunnahar, and Naveen Eluru (2016) Latent segmentation based count models: Analysis of bicycle safety in Montreal and Toronto. *Accident Analysis & Prevention* **95**, 157-171.

Yasmin, Shamsunnahar, Naveen Eluru, Chandra R. Bhat, and Richard Tay (2014) A latent segmentation based generalized ordered logit model to examine factors influencing driver injury severity. *Analytic Methods in Accident Research* **1**, 23-38.

Yen, Steven T., and Jan Rosinski (2008) On the marginal effects of variables in the log-transformed sample selection models. *Economics Letters* **100**(**1**), 4-8.

Young, Mischa, and Steven Farber (2019) The who, why, and when of Uber and other ride-hailing trips: An examination of a large sample household travel survey. *Transportation Research Part A: Policy and Practice* **119**, 383-392.

Yu, Hao, Zhenning Li, Guohui Zhang, and Pan Liu (2019) A latent class approach for driver injury severity analysis in highway single vehicle crash considering unobserved heterogeneity and temporal influence. *Analytic Methods in Accident Research* **24**, 100110.

Yu, Haixiao, and Don MacKenzie (2016) Modeling Charging Choices of Small-Battery Plug-In Hybrid Electric Vehicle Drivers by Using Instrumented Vehicle Data. *Transportation Research Record* **2572(1)**, 56-65.

Yu, Rongjie, Xuesong Wang, and Mohamed Abdel-Aty (2017) A Hybrid Latent Class Analysis Modeling Approach to Analyze Urban Expressway Crash Risk. *Accident Analysis & Prevention* **101**, 37-43.

Yung, Yiu-Fai (1997) Finite mixtures in confirmatory factor-analysis models. *Psychometrika* **62**(**3**), 297-330.

Zahabi, Seyed Amir H., Luis Miranda-Moreno, Zachary Patterson, and Philippe Barla (2015) Spatio-temporal analysis of car distance, greenhouse gases and the effect of built environment: A latent class regression analysis. *Transportation Research Part A: Policy and Practice* **77**, 1-13.

Zeevi, A. J., R. Meir, and V. Maiorov (1998) Error bounds for functional approximation and estimation using mixtures of experts. *IEEE Transactions on Information Theory* **44(3)**, 1010-1025.

Zen, H., and A. Senior (2014) Deep mixture density networks for acoustic modeling in statistical parametric speech synthesis. 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 4-9 May 2014.

Zhang, Junyi, Masashi Kuwano, Backjin Lee, and Akimasa Fujiwara (2009) Modeling household discrete choice behavior incorporating heterogeneous group decision-making mechanisms. *Transportation Research Part B: Methodological* **43(2)**, 230-250.

Zhang, Lei, Jinhyun Hong, Arefeh Nasri, and Qing Shen (2012) How built environment affects travel behavior: A comparative analysis of the connections between land use and vehicle miles traveled in US cities. *Journal of Transport and Land Use* **5(3)**, 40-52.

Zhang, Zhengchao, Congyuan Ji, Yineng Wang, and Yanni Yang (2020) A Customized Deep Neural Network Approach to Investigate Travel Mode Choice with Interpretable Utility Information. *Journal of Advanced Transportation* **2020**, 5364252.

Zhao, Qingyuan, and Trevor Hastie (2021) Causal Interpretations of Black-Box Models. *Journal of Business & Economic Statistics* **39**(**1**), 272-281.

Zhao, Xilei, Xiang Yan, Alan Yu, and Pascal Van Hentenryck (2020) Prediction and behavioral analysis of travel mode choice: A comparison of machine learning and logit models. *Travel Behaviour and Society* **20**, 22-35.

Zhou, Bin, and Kara M. Kockelman (2008) Self-Selection in Home Choice:Use of Treatment Effects in Evaluating Relationship Between Built Environment and Travel Behavior. *Transportation Research Record* **2077**(**1**), 54-61.

Zhou, Heng, Richard Norman, Jianhong Xia, Brett Hughes, Keone Kelobonye, Gabi Nikolova, and Torbjorn Falkmer (2020) Analysing travel mode and airline choice using latent class modelling: A case study in Western Australia. *Transportation Research Part A: Policy and Practice* **137**, 187-205.

Zou, Yajie, Hang Yang, Yunlong Zhang, Jinjun Tang, and Weibin Zhang (2017) Mixture modeling of freeway speed and headway data using multivariate skew-t distributions. *Transportmetrica A: Transport Science* **13**(**7**), 657-678.

Zou, Yajie, and Yunlong Zhang (2011) Use of Skew-Normal and Skew-t Distributions for Mixture Modeling of Freeway Speed Data. *Transportation Research Record* **2260**(**1**), 67-75.

Zou, Yajie, Yunlong Zhang, and Dominique Lord (2013) Application of finite mixture of negative binomial regression models with varying weight parameters for vehicle crash data analysis. *Accident Analysis & Prevention* **50**, 1042-1051.

Zou, Yajie, Yunlong Zhang, and Dominique Lord (2014) Analyzing different functional forms of the varying weight parameter for finite mixture of negative binomial regression models. *Analytic Methods in Accident Research* **1**, 39-52.