NORTHWESTERN UNIVERSITY

Accounting and Controlling for Heterogeneity in Behavior and Survey Response: Application in Non-Profit Fundraising and Commute Mode Choice

A DISSERTATION

SUBMITTED TO THE GRADUATE SCHOOL
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

for the degree

DOCTOR OF PHILOSOPHY

Field of Civil and Environmental Engineering

By

Jingyuan Bao

EVANSTON, ILLINOIS

June 2021

ABSTRACT

Accounting and Controlling for Heterogeneity in Behavior and Survey Response:
Application in Non-Profit Fundraising and Commute Mode Choice

Jingyuan Bao

This dissertation present a Compound Poisson Mixture Regression model of the distribution of transaction frequency and monetary value, and apply it to study donations at a private university in the Midwestern United States. The model captures the joint effect of covariates, recognizing that both response variables emanate from one statistical unit – a donor. Moreover, the mixture regression framework provides a rigorous and appealing approach to account for heterogeneity and other features in the data. In particular, the framework captures latent, group-level factors through coefficients that vary across the different population segments.

The data in the study are from donation records for the 17 year period between 2000-2016, and an alumni survey conducted in the Fall of 2017. The empirical results highlight features of the proposed model, and lead to insights with potential to improve fundraising efforts. Specifically, the results show that the proposed model captures behavioral differences manifested as heterogeneity in either donation amounts, frequencies,

or both response variables. Interestingly and in spite of the inclusion of subjective factors assessed through the survey, the results suggest that between-segment differences are not explained by the available data, i.e., the between-segment heterogeneity is unobserved. The results show that covariates, including a number of subjective factors, i.e., connectedness/psychological distance, perceptions of donation impact, and willingness to volunteer, display stratified marginal effects on either transaction amounts, frequencies, or compound effects on both response variables. We discuss how characterization of such effects supports development of targeted fundraising/marketing strategies.

In order to deal with heterogeneous issues arising from the Compound Poisson Mixture Regression model, and to provide a practical way to control rating scale bias in a broader field, we present a method to estimate and control for individuals' rating scale biases appering in responses to surveys about their experiences, attitudes, feelings and perceptions. The approach is based on the Rasch model, and is motivated by the increasing use of survey data in marketing research. Without relying on additional objective information for anchoring purposes, the proposed approach utilizes only survey data itself to provide individual-question level bias correction, with impacts of both individual rating scales and specific questions accounted for. We apply the method to study data from an alumni survey at a private university in the Midwestern United States. Specifically, we use the bias-corrected parameters to estimate the relationships between attitudes and donation behavior. The results show that the bias-corrected survey data significantly improves model accuracy. Moreover, we observe that the marginal effects of survey variables from the bias-corrected model turn out to be different with model with original survey data in certain variables, which indicates that rating scale biases may impact insights related to the effects of alumni attitude. While the (practical) effectiveness of the proposed bias correction method is illustrated, we discuss limitations in the Rasch Model-based method.

To further generalize accounting for heterogeneity in transportation field, this dissertation presents a segmentation analysis of households in the Chicago Metropolitan Area based on reported travel outcomes. The data are from the travel tracker survey conducted between 2007-2008 by the Chicago Metropolitan Agency for Planning. In our analysis, we assume that unobserved, group-level factors play a pivotal role in determining/explaining the heterogeneity observed across the population in terms of mode choice and distance traveled. As a benchmark, we consider a segmentation model relying exclusively on distance traveled by personally-owned vehicle or taxi, an approach used the literature. The results suggest additional information on trips of other modes is useful and validates our joint segmentation approach. Our analysis of the Chicago data suggests that the population consists of 4 segments of households. Aggregate analysis of the travel outcomes in each ZIP code highlights complicated inter-dependencies among travel behavior, residential location, and public transport coverage. Nevertheless, disaggregate analysis (of the correlations in the cluster membership probabilities) suggests that socioeconomic and demographic factors play stronger role in travel outcomes, than do build environment factors. The discussion concludes the actual relationship between urban form and travel behavior is not as simple as it seems in analysis of their statistical relationship, and relevant policies are also supported by our findings.

Acknowledgements

I would like to express my gratitude to my advisor Prof. Pablo L. Durango-Cohen. To me, he is a teacher, a colleague and also a friend. I cannot imagine the 5 years at Northwestern without his generous help, guidance, and support.

In addition, I would like to thank Prof. Ji-Ping Wang, who provided guidance and precious advice in my learning and research in statistical modeling. I would also like to thank Prof. Amanda Stathopoulos. She has been an amazing professor throughout my time at Northwestern, sharing every valuable insight in my research.

My appreciation also goes to my friends and colleagues, especially in the transportation program. Your company has made my time at Northwestern easier and warmer.

Last but not the least, I would like to thank my family. Although we have been apart for most of the time in these years, I always feel you by my side.

Table of Contents

List of Tables

List of Figures

CHAPTER 1

Introduction

1.1. Motivation

Higher education institutions in the United States and elsewhere increasingly rely on support from individual donors. According to Council for Aid to Education (2020), colleges and universities in the US raised \$19.50 billion from individuals in 2019, accounting for 38.3% of the total annual voluntary support, and representing a 13.7% increase over the amount raised in 2014. Sargeant (2013) estimates that a typical charity loses 50% of its annual donors after their first donation, and up to an additional 30% after their second. Moreover, the costs of recruiting a new donor are estimated to be about 3 times those of maintaining an existing one (Bennett, 2006). Therefore, it is important to characterize the marginal effect of various factors that influence recurring donations, to better tailor soliciting strategies for institutions.

As one type of the important factors that influencing individual donation behavior, psychological factors – including feelings, experience, and perception factors – have received significant attentions. Survey data have become a popular and important tool in marketing research for collecting such information from customers. Advantages of surveys, compared to recording long-term transaction data and interviewing customers, may include lower costs, faster data collection, and higher response rates (Ilieva et al., 2002). Also, as explained by Ping Jr (2004) and Yang et al. (2010), marketing research relies on survey data when factors related to customer behavior are unobserved, e.g., attitudes, experiences, feelings, perceived relationships, etc. In such surveys, customers are often asked to select a single number within a discrete and ordinal set – known as Likert scale – to rate their level of agreement with a given question/statement. Rating scores, in turn, are used to make inferences that, ultimately, may support marketing strategies that target (potential) customers. However, even when common errors and biases in data collection are controlled for, the issue of accounting for rating scale biases often persists. An individual's *rating scale* or *response style* is tied to the process by which they convert subjective assessments into numerical scores representing levels within the given scales/sets. The issue is that the same rating score from different individuals may actually reflect different underlying feelings or beliefs, while at the same time, individuals reporting different ratings may actually share similar feelings. It follows that ignoring rating scale biases may undermine the reliability of inferences and subsequent analysis Greenleaf (1992).

In travel behavior analysis, commute trips garner the lion's share of the attention for mode choice modeling, since traffic during weekday's peak hours causes the most severe congestion issues (Bhat, 1997b). In addition to mode choice, recurring trips are more likely to influence residential location, vehicle ownership, and other decisions tied to travel outcomes. According to Elldér (2014), while weekday commute and service trips are dependent on residential location, leisure trips on weekends have great variation within a neighborhood. Vehicle miles traveled (VMT) is often used used as a measure of travel behavior. This measure is also widely used in a variety of transportation and planning functions, e.g., estimation of vehicle emissions and energy consumption, public transit planning, etc. VMT often refers to the total distance traveled by Privately Owned

Vehicle (POV) or by Taxi. Trip distances covered by other modes are often excluded (Chatman, 2003), or added to VMT by POV and Taxi together (Hong et al., 2014). Other literature (Lin and Long, 2008; Paulssen et al., 2014) tried to introduce mode share to reflect people's mode choice, but VMT and mode share have not been both used together to represent commute travel behavior.

In this dissertation, motivated by the aforementioned limitations existing in marketing research and travel behavior analysis, we present the following models to better account for and control the heterogeneity that widely exist.

1.2. Dissertation outline

The remainder of the dissertation is organized as follows:

Chapter 2 presents the Compound Poisson Mixture Regression model for recurring donations. We focus on modeling the joint distribution of donation frequency and amount, and capturing heterogeneity in donation behavior distribution and effects of individual characteristics on donation behavior.

Chapter 3 further dives in the problem of heterogeneous individual rating scale and rating scale bias that exists in subjective survey data. We propose a rating scale bias correction method based on Rasch Model, and present an application in alumni survey data to validate effectiveness of the proposed method.

Chapter 4 presents a joint segmentation analysis in travel behavior to capture the heterogeneity in multi-dimensional travel behavior variables, and shows how travel patterns differ across different segment. Spatial distribution analysis is also presented to demonstrate the travel pattern characteristics.

CHAPTER 2

An analysis of factors influencing recurring donations in a university setting: A Compound Poisson Mixture Regression model

2.1. Introduction

Higher education institutions in the United States and elsewhere increasingly rely on support from individual donors. According to Council for Aid to Education (2020), colleges and universities in the US raised $19.50 billion from individuals in 2019, accounting for 38.3% of the total annual voluntary support, and representing a 13.7% increase over the amount raised in 2014. Sargeant (2013) estimates that a typical charity loses 50% of its annual donors after their first donation, and up to an additional 30% after their second. Moreover, the costs of recruiting a new donor are estimated to be about 3 times those of maintaining an existing one (Bennett, 2006). While aggregate data for universities are not, as far as we are aware, available, these trends also apply to them,[1] and highlight the importance of understanding/managing recurring donations as part of effective and efficient fundraising efforts. We also note that the motivation to retain/add donors transcends fundraising *per se* because influential information brokers, i.e., US News and

[1] The results in Durango-Cohen et al. (2013a), in a similar setting to the present study, show attrition rates exceeding 50% after 2 years. In the same vein, McAlexander and Koenig (2001); Quigley Jr. et al. (2002) state that maintaining the loyalty of existing alumni donors at universities is easier than enlisting new ones.

World Report, use the percentage of alumni donating to a university's annual fund in their rankings of colleges and universities as a measure of student satisfaction.

The aforementioned trends motivate the work herein, which aims to characterize the marginal effect of various factors that influence recurring donations. Specifically, we formulate a Compound Poisson Mixture Regression (CPMR) model of the joint distribution of donation frequency and monetary value. The model provides an intuitive approach to characterize the compound effect of covariates on the 2 response variables, recognizing that they emanate from one statistical unit – a donor. Mixture modeling provides a flexible and rigorous approach to account for unobserved heterogeneity across a population. Unobserved heterogeneity refers to systematic differences among individuals stemming from factors that are unobserved because they are unobservable, e.g., feelings, motivation, or because data are (inadvertently) missing. In a MR model, the assumption is that unobserved heterogeneity is manifested through marginal effects stratified across the segments that comprise the population. From a marketing standpoint, the approach is appealing because it supports identification of the key factors that determine (variation in) each segment's responses, and thus, the development of tailored fundraising strategies.

We use the CPMR model to analyze donations at a private university in the Midwestern United States. The data are from 2 sources: alumni donation records for the 17 year period between 2000-2016, and an alumni survey conducted in the Fall of 2017. Information held by the university includes transaction records, as well as information on socioeconomic, education, and demographic (SEED) factors used in the analysis. We partnered with the university to design and conduct an alumni survey to elicit information about their experiences as students/graduates, as well as their feelings towards and their perceived relationship with the institution. We refer to these covariates as experience, feelings, and relationship (EFR) factors. The empirical results highlight features of the proposed model, and lead to insights with potential to improve fundraising efforts. Specifically, the results show that the proposed model captures behavioral differences manifested as heterogeneity in either donation amounts, frequencies, or both response variables. Interestingly and in spite of the inclusion of EFR factors, the empirical results suggest that these between-segment differences are not explained by the available data, i.e., the between-segment heterogeneity is unobserved. Between-segment differences are explained by the covariates' heterogeneous marginal effects across the population. This includes a number of EFR factors, i.e., connectedness/psychological distance, perceptions of donation impact, willingness to volunteer, that display significant effects on either transaction amounts, frequencies, or compound effects on both response variables. Importantly and as discussed later, characterization of such effects supports development of targeted marketing strategies. The results are, of course, functions of the data, but they reinforce the importance of the survey to explain the donor population's responses and the choice of model, which are important contributions of the present study.

The remainder of the paper is organized as follows. Section 2.2 positions our work with respect to the charitable-giving literature, and with respect to market segmentation models with applications to university fundraising. We also highlight the contributions of this study. In Section 2.3, we describe the data used in the study, including preprocessing. In Section 2.4, we present a formulation of the CPMR model. We also describe model estimation, selection, and validation. Empirical results and discussion, focusing on the covariates' heterogeneous marginal effects are presented in Section 2.5.

Finally, conclusions, limitations of the present study, and potential directions appear in Section 2.6.

2.2. Literature review

The paper presents a model of the joint distribution of transaction frequency and monetary value, which we use to characterize the effect of various factors on recurring donations to a university. In this section, we position the work with respect to benchmarks appearing in the literature on university fundraising and on market segmentation. Comprehensive reviews of empirical analyses in the charitable giving literature and of the university fundraising literature appear in Bekkers and Wiepking (2007, 2011) and Lindahl and Conley (2002), respectively. Wedel and Kamakura (2000) is a seminal reference on segmentation modeling and analysis.

The model presented here differs from counterparts appearing in the fundraising literature in terms the representation of donation behavior along multiple dimensions. It is also relevant to highlight the focus on characterizing factors that contribute to recurring/repeated transactions, which is supported by the available data.

The majority of empirical studies of university fundraising/alumni-giving use monetary value as the response variable of interest (see, e.g., Leslie and Ramey (1988); Clotfelter (2003); Pedro et al. (2018)); however, the importance of studying the factors that affect donation propensity/likelihood/timing/frequency has been recognized and is increasing. Holmes (2009) is, perhaps, the most comprehensive study to consider the 2 response variables. In particular, she presents tobit and probit models of donations at a liberal arts college. Interestingly, the results show that, within the same donor population, the determinants of the 2 response variables are different, i.e., higher incomes and residence in a state that allows charitable tax deductions have a large marginal effect on donation amounts, whereas athletic and academic prestige have large effects the number of donations. While there are studies of interactions between the 2 responses, cf. Lindahl and Winship (1992) who present a logit model of the likelihood of a donation exceeding a threshold, models in the literature formulated to capture the effect of factors on their joint distribution are rare.[2] That is, recognizing that both variables correspond to responses by the same statistical unit. This is an important contribution because evidence from synergistic fields, such as consumer behavior, shows that analogous variables, purchase quantities and timing/frequency, exhibit strong relationships (Simonson, 1990; Boatwright et al., 2003; Jen et al., 2009). The segmentation analysis of Schröder et al. (2019) also shows the benefits of combining multidimensional responses, i.e., online browsing behavior and purchasing, to identify the factors that influence consumer behavior.

As other charitable giving settings, the university fundraising literature reflects significant interest on identification of factors that influence donations. At the individual level, oft-used SEED variables include age, income, degree type, and gender (Okunade, 1996; Monks, 2003; Weerts and Ronca, 2007; Dvorak and Toubman, 2013). Individual level EFR factors, such as participation in social activities, class reunion attendance, feelings of gratitude, etc. are often assessed through surveys and have also been shown to play an important role in alumni-giving (Clotfelter, 2003; Gaier, 2005; Merchant et al., 2010;

[2] Moon and Azizi (2013) is one of the few exceptions. They present an auto-Logistic and an auto-Gaussian model, to predict who will donate and how much they will donate, respectively. The two models are integrated using the Tobit type 2 framework, which allows for simultaneous estimation of the 2 components. The model is used to study donations to a veterinary school. The results link treatments received by pets to subsequent donations.

Aaker et al., 2010; McDearmon, 2013; Pedro et al., 2018). Studies of multiple universities have led to the identification of significant institutional level factors such as type of institution (public vs. private), athletic success, number of students, student body composition (gender, full/part-time type, etc.), cost of tuition, average student debt, etc. (Turner et al., 2001; Gunsalus, 2005; Terry and Macy, 2007; Holmes et al., 2008; Meer and Rosen, 2009). As is explained in the context of synergistic charitable giving settings, cf. Ein-Gar and Levontin (2013); Levontin et al. (2015) and the references therein, little is known about which (aforementioned) factors influence individuals to donate repeatedly. This is explained by both the lack of (sufficient) historical data, and by the literature's focus on (average) donation amounts. As in the consumer behavior literature, where Reibstein (2002) and others have shown that the antecedents of first-time/sporadic and recurring behaviors can differ, the analysis herein, therefore, addresses a gap in the university fundraising literature – one that is of practical significance due to the importance of (increasing) donor retention.

We rely on the taxonomy of segmentation models presented in Wedel and Kamakura (2000) to position the model presented herein with respect to those appearing in the university fundraising literature and other similar applications. Among others, they rely on the following attributes: segmentation approach, segmentation basis, and type of statistical model used to explain responses. With respect to segmentation approach, models are classified as either *a priori* or *post-hoc*. In *a priori* segmentation models, the number and types of segments are determined in advance of the analysis, whereas in *post-hoc* segmentation they are determined as a result of the analysis based on goodness-of-fit or other criteria. The conditions used to assign individuals to segments are referred to as segmentation bases. Segmentation bases are either *observed* when they correspond to observed/measured SEED, EFR, institutional, or response variables; or *unobserved*, frequently attributed either to missing data or to latent variables. Finally, with respect to the capability to explain responses, models are classified as *predictive* or *descriptive*. Predictive models relate explanatory variables to the outcomes of a set of dependent variables. In contrast, descriptive models represent the (joint) distribution of the variables without distinction between outcome and explanatory variables. Based on this taxonomy, the proposed model and approach can be categorized as *post-hoc*, *predictive*, and where the segmentation basis can be *unobserved*.

Segmentation analyses are widely used in the university fundraising/alumni-giving to explain heterogeneity in donor populations. Work in the literature typically relies on observed bases to predetermine population segments *a priori*. The goals are either to identify traits of individuals sharing similar donation behavior, i.e., RFM statistics,[3] or to describe/predict donation behavior of individuals sharing SEED, EFR, or other characteristics. In terms analysis of repeated transactions, Wunnava and Lauze (2001) and Sun et al. (2007), segment populations into groups of occasional and consistent donors, or into groups of donors and nondonors. They profile the ensuing segments to draw insights about factors associated with the different responses. Predictive models relying on *a priori* segmentation have also appeared in the literature. For example, the seminal study of Leslie and Ramey (1988) shows that regression coefficients for various segments – alumni, non-alumni, business, and non-business organizations – display differences in terms of

[3] RFM statistics refer to *recency*, i.e., number of periods/years since last donation, *frequency*, i.e., number of years donated in a given analysis period, and *monetary value*, i.e., average, total, maximum donation over an analysis period.

their magnitudes, signs, and levels of significance. *Post hoc* segmentation approaches relying on observed bases are also common. Representative studies relying on contemporary methods include Weerts and Ronca (2009) and Le Blanc and Rucks (2009). Weerts and Ronca (2009) use the Classification and Regression Tree (CART) method, and Le Blanc and Rucks (2009) present a cluster analysis approach. The CART method divides the explanatory variables into 2 sets: the first set is used to segment the population, and the second set is used to fit regressions that describe within segment variation. Cluster analysis relies on optimization problems with objectives to maximize within-segment homogeneity or between-segment heterogeneity. The criteria, including the variables of interest, are specified in advance. The advantage of relying on observed segmentation bases is that results can be intuitive and actionable – even for individuals entering the population (see, e.g., the synergistic study of Vafainia et al. (2019)). The disadvantage is that the results can be biased where, as is often the case, heterogeneity is driven by unobserved factors.

Finite-Mixture/Latent-Class Modeling has emerged in the last 3 decades as a rigorous framework to explain unobserved heterogeneity in populations of interest (McLachlan and Peel, 2004). The underlying assumptions are that populations are composed of a finite set of segments (in unknown proportions), and that heterogeneous responses are explained by latent, group-level effects. Compared to conventional approaches to account for unobserved heterogeneity, i.e., fixed and random effects models, the framework is appealing because, *post hoc*, the ensuing segmentation is actionable. The framework's flexibility to accommodate different structures representing behavioral responses is another feature that makes it appealing. Here, we exploit this flexibility to represent donations from a population as a finite mixture of Compound Poisson Processes. CP models have been used to model a wide range of phenomena where event magnitude and frequency are relevant. Examples include rainfall (Öztürk, 1981; Dunn, 2004), hurricane likelihood and intensity (Katz, 2002), and to model the distribution of insurance claims (Jørgensen and Paes De Souza, 1994; Smyth and Jørgensen, 2002). Not only does the model avoid bias introduced by conducting separate analysis, but it also provides a basis to estimate compound effects of covariates on the 2 response variables. We are not aware of applications of CP models where it is relevant to account for heterogeneous effects of covariates.

The CPMR model builds on recent work formulating descriptive segmentation models that capture the joint distribution of donation frequency and monetary value. Durango-Cohen et al. (2013a) presents a Markov Chain mixture model, i.e., a dynamic model, of how donations evolve over time. Rather than trends or other features, the results show that the patterns that govern each of the segments are determined by the frequency of donations. Durango-Cohen et al. (2013b) presents a Bernoulli-Gaussian model, which is similar to the work herein. Donation sequences are assumed to be generated by (a finite set of) Bernoulli processes, where the monetary values follow a Gaussian Distribution. Thus, the number of donations over a given number of periods/years follow a Binomial Distribution. Here, we approximate the Binomial Distribution with a Poisson Distribution, which makes the model more intuitive and comparable to analysis presented in other disciplines. It also improves model estimation. More importantly, the model herein differs in that our objective is to estimate the effect of various factors on the response variables. A practical contribution of the present study is that, in addition, to SEED variables, the study includes the effect of subjective factors related to experiences as students and alumni of the university, as well as feelings towards and the relationship with the institution. Data to assess the latter were obtained from an alumni survey conducted in partnership with the university. The inclusion of covariates in the model means that, technically, rather than reflecting differences in the response variables, the segmentation captures between-segment differences in the marginal effects of the explanatory variables. This is a significant contribution from a marketing perspective because, as is discussed, it supports and provides estimates of the impacts of tailored strategies.

2.3. Data

The data used in this study are from 2 sources: (i) alumni donation records for the 17 year period between 2000-2016, and (ii) a survey conducted in the Fall of 2017. Donation records are held by the university, and include information on socioeconomic, education and demographic factors. We partnered with the university to design and conduct an alumni survey to elicit information about their experience as students and graduates, as well as their feelings towards and perceived relationship with the institution. Among the 2,859 survey respondents, 1,042 donated at least once in the analysis period. Our final data set is from records and survey responses for 771 alumni, who graduated in 2014 or earlier, whose donations averaged $1,000 or less per year, and who responded to 11 or more (out of 17) survey questions. The data used in our analysis, including the construction of the predictor and response variables, as well as explanation regarding discarded records, are described further in the remainder of the section.

The 2 response variables in our analysis are number of years donated in the analysis period, and average monetary value ($/donation). For example, an individual with donations in 4 out of the 17 years in the period 2000 – 2016 with respective totals $100, $200, $100, and $150, and with no donations recorded in the other 13 years, donated at a frequency of 4 times per 17-years with average value of $(100+200+100+150)/4 = \$137.5$ per donation. For individuals graduating after the year 2000, the number of donations over a 17 year period was estimated as $\hat{T}(i) = \left\| T(i)\frac{17}{2017-G_i} \right\|$, where G_i is i's graduation year, $T(i)$ is the actual number of years when donations were received, and $\|\cdot\|$ is the rounding operator. To avoid overestimating the number of donations of recent graduates, alumni graduating after 2014 were excluded.

The 5 SEED factors used in our study are Years Since Graduation (YSG), Household Income Level, Gender, Field of Study, and Degree. Some variables were constructed from donor records held by the university. Specifically:

- YSG correspond to the difference between 2016 and a donor's graduation year.
- Household Income Level was obtained from the alumni survey, where respondents were asked to select 1 of the following 8 income levels: 1 – Less than $25,000; 2 – $25,000 to $34,999 ; 3 – $35,000 to $49,999 ; 4 – $50,000 to $74,999 ; 5 – $75,000 to $99,999 ; 6 – $100,000 to $149,999 ; 7 – $150,000 to $199,999 ; 8 – $200,000 or higher.
- For the Field of Study variable, the types "Certificate", "JD", and "Unknown" were combined into an "Others" type because they have very few observations. Also, we use the first degree earned at the university and the School for individuals

who received multiple degrees from the university.[4] For example, an individual with a BS degree in Science and an MS in Engineering has corresponding Degree and School types of "Undergraduate" and "Science".

The 17 questions used to rate EFR factors, and the corresponding response ranges, are listed below. The choice of questions, and the format in which they were displayed drew from the literature. Examples are (i) the questions on student experience and satisfaction, which were adopted from Kramer and Yoon (2007); Jung and Yoon (2013), and (ii) the 7-point Likert scale for the "Similarity" variable was mapped to Venn-like diagrams that represent different degrees of similarity between alumni and university by different extents of overlapping between two circles, adopted from Aron et al. (1992).

- Connected: How connected do you feel to the institution? (From "1 – Not very connected" to "7 – Very connected")
- FeelingsUniv: How would you describe your feelings toward the university today? (From "1 – Very cold" to "7 – Very warm")
- FeelingsAA: How would you describe your feelings toward the university's alumni association? (From "1 – Very cold" to "7 – Very warm")
- Competent: How competent do you perceive the university to be? (From "1 – Very incompetent" to "7 – Very competent")
- Similarity: Please rate the similarities between you and the university. (From "1 – Not very similar" to "7 – Very similar")
- UseWordUs: To what extent do you use the word "us" to describe you and the university community? (From "1 – Never" to "7 – Always")

[4] As discussed below, multiple degree interactions were explored, but turned out to be insignificant.

- DonationImpact[5]:

  For donors: How impactful do you feel your donation(s) has/have been? (From "1 – Not impactful at all" to "7 – Very impactful")

  For non-donors: What do you think the impact of a donation from you would be? (From "1 – Not impactful at all" to "7 – Very impactful")

- StuExpAcademic: How would you assess your experience as a student? Overall academic experience. (From "1 – Completely dissatisfied" to "7 – Completely satisfied")
- StuExpAcademicDept: How would you assess your experience as a student? Overall academic experience in your department or program. (From "1 – Completely dissatisfied" to "7 – Completely satisfied")
- StuExpLife: How would you assess your experience as a student? Non-academic or student life experience. (From "1 – Completely dissatisfied" to "7 – Completely satisfied")
- StuExpOverall: How would you assess your experience as a student? Overall experience at the university. (From "1 – Completely dissatisfied" to "7 – Completely satisfied")
- UnivOweSuccess: Please rate your level of agreement with the statement – I owe a portion of my career success to the university. (From "1 – Strongly disagree" to "7 – Strongly agree")

[5] This variable is obtained by combining the responses of two questions below, for donors and non-donors respectively. Namely, we take responses of first question for donors, and second question for non-donors.

- UnivRecommend: Please rate your level of agreement with the statement – I would recommend the university to friends or family. (From "1 – Strongly disagree" to "7 – Strongly agree")
- MonthWebsite: In the last three months, how often have you visited any university website? (1 – Never, 2 – Rarely, 3 – Once a month, 4 – Once a week, 5 – Twice a week)
- YearContact: Over the past two years, how often were you contacted by the university? (1 – Never, 2 – Once a year, 3 – Twice a year, 4 – 3-5 times per year, 5 – 6-11 times per year, 6 – Monthly)
- ContactSatisfied: How satisfied are you with the amount of contact from the university? (From "-3 – Far too little" to "3 – Far too much". As shown in Table 2.3, the range was mapped to 1–7 range.)
- WishVolunteer: Would be interested in volunteering at an alumni event? (0 – No, 1 – Yes)

As noted, records for donors who skipped 6 or more survey questions were excluded from the study (9.72% of donor respondents). Although data imputation, described below, was used to deal with missing values, we set the threshold of 6 to balance the tradeoffs between the quality of the imputation and the bias introduced by excluding records. To assess the nature of this bias, we ran a significance test on the numbers and average amounts of donations of the included and excluded individuals. Because neither response variable follows a Normal Distribution, we use the Mann–Whitney U Test, a non-parametric test, to establish if the data from the 2 groups follow the same distributions. The null hypothesis is that the 2 populations follow equal distribution, which is

Table 2.1. Significance Test Results of Donation Amount and Frequency

	Included Population Mean	Excluded Population Mean	Mann–Whitney U Test p-Value
Proportion	90.28%	9.72%	-
Donation Amount	209.56	199.01	0.230
Donation Number	6.91	5.03	0.003

rejected with low p-values. Table 2.1 displays the means and p-values for both donation numbers and average values. The results indicate that the $\approx$ \$10 difference in average values is not significant. In contrast and on average, $\approx$ 2 fewer donations were received from alumni who skipped 6 or more survey questions. A difference that is statistically significant, and suggests that results should not be extrapolated.

Statistics describing the response and SEED variables of the 771 individuals included in the analysis appear in Table 2.2. The statistics differ from those in Table 2.1 because donors with average amounts of \$1,000 or more were removed after those with 6 or more missing responses.

For the 771 donors in the analysis, there are still some missing values from unanswered questions. We address the issue by conducting a data imputation using the R Package, *Mice*, where missing values are obtained by sampling from conditional distributions obtained from observed data. More details regarding the data imputation methodology and our processing are presented in A and in Buuren and Groothuis-Oudshoorn (2010). Table 2.3 shows statistics of the EFR variables in the final data set after imputation.

Table 2.2. Summary of Response and SEED Variables

Variable	Min	Mean	Max	S.D
Donation Amount	4.33	128.37	941.18	148.11
Donation Frequency	1	6.91	17	5.39
Year Since Graduation (YSG)	2	33.15	75	17.92
Household Income Level	1	5.86	8	1.67
Gender Distribution (Male)		79.51%		
Degree Type:				
Undergraduate		77.56%		
Master		19.58%		
Doctoral		2.72%		
Others		0.13%		
Field of Study:				
Engineering		51.88%		
Law		0.13%		
Architecture		9.99%		
Science		16.21%		
Design		3.63%		
Human Sciences		5.97%		
Applied Technology		1.30%		
Business		10.89%		

2.4. Model formulation and estimation

In this section, we describe the CPMR model used in our analysis. We begin by presenting basic properties of CP processes, and explain why they are appealing to model donation behavior. We then describe how the model can be generalized using the mixture modeling framework. Finally, we provide an overview of the estimation and model/feature selection process used herein.

Table 2.3. Summary of EFR Variables

Variable	Mean	S.D.
Connected	3.96	1.56
FeelingsUniv	4.54	1.31
FeelingsAA	4.49	1.36
Competent	5.70	1.25
Similarity	3.57	1.58
UseWordUs	3.06	1.81
DonationImpact	4.02	1.49
StuExpAcademic	5.75	1.04
StuExpAcademicDept	5.91	1.12
StuExpLife	4.45	1.69
StuExpOverall	5.54	1.14
UnivOweSuccess	5.63	1.39
UnivRecommend	5.71	1.31
MonthWebiste	1.64	0.80
YearContact	4.03	1.31
ContactSatisfied	4.19	0.79
WishVolunteer	0.11	0.31

2.4.1. Compound Poisson Distribution

Compound Poisson processes are used to describe the distribution of the sum of *iid* random variables, where the number of events/transactions follows a Poisson process. Letting $T(i)$ represent the number of times/years that individual i donated in the analysis period, and ξ_i^t be the monetary value of individual i's tth donation, we have that the sum of individual i's donations, $\chi(i)$ is given by

$$\chi(i) = \sum_{t=1}^{T(i)} \xi_i^t \tag{2.1}$$

where the number of donations, $T(i)$, is assumed to follow a Poisson Distribution, i.e., $T(i) \sim \text{Poisson}(\lambda), i = 1, \ldots, N$, where $\xi_i^t, t = 1, \ldots, T(i); i = 1, \ldots, N$ are assumed to be *iid*, and further, are assumed to be independent of $T(i), i = 1, \ldots, N$. Thus, the expectation and variance of $\chi(i)$ are given by

(2.2)

$$E[\chi(i)] = E_{T(i)}\left[E_{\chi(i)|T(i)}[\chi(i)|T(i)]\right] = E_{T(i)}\left[T(i) \cdot E[\xi_i^t]\right] = E[T(i)] \cdot E[\xi_i^t] = \lambda \cdot E[\xi_i^t]$$

$$\begin{aligned} Var[\chi(i)] &= E_{T(i)}\left[Var_{\chi(i)|T(i)}(\chi(i)|T(i))\right] + Var_{T(i)}\left[E_{\chi(i)|T(i)}(\chi(i)|T(i))\right] \\ &= E_{T(i)}\left[T(i) \cdot Var(\xi_i^t)\right] + Var\left(T(i) \cdot E[\xi_i^t]\right) \\ &= E[T(i)] \cdot Var(\xi_i^t) + E[\xi_i^t]^2 \cdot Var(T(i)) = E[T(i)] \cdot \left(Var(\xi_i^t) + E[\xi_i^t]^2\right) \\ &= \lambda \cdot E[(\xi_i^t)^2] \qquad (2.3) \end{aligned}$$

where $E[T(i)] = Var(T(i)) = \lambda$ is from the Poisson Distribution.

2.4.2. Mixture Regression model

We refer to the number of donations in the analysis period as the Frequency variable, Y_f, and the average monetary value per year as the Amount variable, Y_a. To truncate and scale the donation amounts, in our model, Y_a is assumed to follow a Log-Normal Distribution.[6] Y_f is assumed to follow a Poisson Process. The explanatory variables, capturing individual i's SEED and EFR characteristics, associated with the 2 response

[6] Other distributions, including the Normal Distribution, were considered, but resulted in inferior goodness-of-fit. We also note that the Normal Distribution does not comply with the assumption that the random variable is positive.

variables are captured in 2 overlapping vectors, $\boldsymbol{x}_{f,i}$ and $\boldsymbol{x}_{a,i}$. The probability distribution of the number of i's donations follows the generalized Poisson Linear Regression Model:

$$f_{freq}(y_{f,i}|\boldsymbol{x}_{f,i},\boldsymbol{\beta}) = \frac{e^{-\exp(\boldsymbol{x}'_{f,i}\boldsymbol{\beta})}\exp(\boldsymbol{x}'_{f,i}\boldsymbol{\beta})^{y_{f,i}}}{y_{f,i}!} \tag{2.4}$$

where $y_{f,i}$ is number of years that individual i donated, and the Poisson linear regression coefficients are captured in the vector $\boldsymbol{\beta}$.

For the Amount part, the Log-Normal assumption means that $y^l_{a,i} = \log(y_{a,i})$ follows a Normal distribution. Thus, the probability distribution of i's average donations per year is given by a Normal linear regression model as follows:

$$f_{amt}(y^l_{a,i}|\boldsymbol{x}_{a,i},\boldsymbol{\gamma}) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left[\frac{-(y^l_{a,i}-\boldsymbol{x}'_{a,i}\boldsymbol{\gamma})^2}{2\sigma^2}\right] \tag{2.5}$$

where $y^l_{a,i}$ is the log of i's average yearly donation amounts, and $\boldsymbol{\gamma}$ is the Normal linear regression coefficient vector. Due to the independence assumption, the joint probability distribution of the average monetary value and donation frequency for individual i is given by:

$$\begin{aligned} f(\boldsymbol{y}_i|\boldsymbol{x}_i,\boldsymbol{\theta}) &= f_{freq}(y_{f,i}|\boldsymbol{x}_{f,i},\boldsymbol{\beta})f_{amt}(y^l_{a,i}|\boldsymbol{x}_{a,i},\boldsymbol{\gamma}) \\ &= \frac{e^{-\exp(\boldsymbol{x}'_{f,i}\boldsymbol{\beta})}\exp(\boldsymbol{x}'_{f,i}\boldsymbol{\beta})^{y_{f,i}}}{y_{f,i}!}\frac{1}{\sqrt{2\pi}\sigma}\exp\left[\frac{-(y^l_{a,i}-\boldsymbol{x}'_{a,i}\boldsymbol{\gamma})^2}{2\sigma^2}\right] \end{aligned} \tag{2.6}$$

where $\boldsymbol{y}_i = \{y_{f,i}, y^l_{a,i}\}$, $\boldsymbol{x} = \{\boldsymbol{x}_{f,i}, \boldsymbol{x}_{a,i}\}$, and $\boldsymbol{\theta} = \{\boldsymbol{\beta}, \boldsymbol{\gamma}\}$.

The CP regression model can be generalized further into a mixture model. In a mixture model, the underlying (donor) population is assumed to be composed of a finite number, K, of collectively exhaustive and mutually exclusive segments. Each individual, in turn, is assumed to belong to exactly one segment, although segment membership is unobserved. In the absence of prior information, the overall probability distribution of $\boldsymbol{y}_i$ is

$$\begin{aligned} f(\boldsymbol{y}_i|\boldsymbol{x}_i,\boldsymbol{\pi},\boldsymbol{\Theta}) &= \sum_{k=1}^{K} \pi_k f_k(\boldsymbol{y}_i|\boldsymbol{x}_i,\boldsymbol{\theta}_k) = \sum_{k=1}^{K} \pi_k f_{freq,k}(y_{f,i}|\boldsymbol{x}_{f,i},\boldsymbol{\beta}_k) f_{amt,k}(y^l_{a,i}|\boldsymbol{x}_{a,i},\boldsymbol{\gamma}_k) \\ &= \sum_{k=1}^{K} \pi_k \frac{e^{-\exp(\boldsymbol{x}'_{f,i}\boldsymbol{\beta}_k)} \exp(\boldsymbol{x}'_{f,i}\boldsymbol{\beta}_k)^{y_{f,i}}}{y_{f,i}!} \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left[\frac{-(y^l_{a,i}-\boldsymbol{x}'_{a,i}\boldsymbol{\gamma}_k)^2}{2\sigma_k^2}\right] \end{aligned} \tag{2.7}$$

where π_k represents the proportion of the population in segment k, $k = 1, 2, .., K$, with $\pi_k \geq 0$ and $\sum_{k=1}^{K} \pi_k = 1$; and $\boldsymbol{\Theta} = \{\boldsymbol{\beta_1}, ..., \boldsymbol{\beta_K}; \boldsymbol{\gamma_1}, ..., \boldsymbol{\gamma_K}\}$. Here $\boldsymbol{\beta}_k$ and $\boldsymbol{\gamma}_k$ are the Poisson and Normal linear regression coefficient vectors for segment k. Notice that all individuals belonging to a segment share the same regression coefficients. We let the latent variable z_i, defined over $\{1, 2, ..., K\}$, represent i's segment membership, and write the probability of individual i belonging to segment k as

$$p_{ik} = P(z_i = k|\boldsymbol{y}_i,\boldsymbol{x}_i,\boldsymbol{\pi},\boldsymbol{\Theta}) = \frac{\pi_k f_k(\boldsymbol{y}_i|\boldsymbol{x}_i,\boldsymbol{\theta}_k)}{\sum_{k=1}^{K} \pi_k f_k(\boldsymbol{y}_i|\boldsymbol{x}_i,\boldsymbol{\theta}_k)} \tag{2.8}$$

Each individual i, therefore, has K segment membership probabilities $p_{i1}, p_{i2}, ..., p_{iK}$. For analysis, as is done below, it is often convenient to assign individuals to a segment with the largest probability, i.e., $\hat{z}_i = \arg\max_k(p_{ik})$.

2.4.3. Model estimation, selection and diagnosis

Here, we describe the estimation of the CPMR model presented in the previous section. We also highlight analysis to support model and feature selection. We used the Expectation-Maximization Algorithm for model estimation. The specific instance appears in B. Grün and Leisch (2008) describes the R Package, *Flexmix*, which we use.

In terms of the model specification, we first checked the multicollinearity among the explanatory variables. Intuitively, there is likely to be multicollinearity, especially among the EFR variables. For example, an individual's academic experience may be a key factor in, and therefore positively correlated to, their overall student experience. To check, we calculate the Variance Inflation Factor (VIF) of all covariates – both SEED and EFR variables.[7] The results in Table 2.4 show that all VIFs are less than 10, which indicates that the data do not exhibit multicollinearity, and thus, all variables are candidates for the model.

The formulation of a MR model requires feature/variable selection, and specification of a number of segments. The latter means that application of stepwise regression for feature selection is not straightforward. Thus, we use an iterative approach of adding segments and keeping variables in each of the 2 sub-models: Frequency and Amount, when their coefficients are significant for at least one segment. This explains why the the vectors $\boldsymbol{x}_{f,i}$ and $\boldsymbol{x}_{a,i}$ have different components. In addition to the raw variables, a

[7] The VIF is a common statistical measure of multicollinearity. It can be calculated by fitting a linear regression of predictive variable x_j on all other predictive variables. Denote R_j^2 to be the R^2 of the regression, and the VIF of x_j is given by $\text{VIF}_j = \frac{1}{1-R_j^2}$.

Table 2.4. Variance Inflation Factors

	Connected	FeelingsUniv	FeelingsAA	Competent
VIF	2.17	2.48	2.21	1.75
	Similarity	UseWordUs	DonationImpact	StuExpAcademic
VIF	1.78	1.91	1.43	2.85
	StuExpAcademicDept	StuExpLife	StuExpOverall	UnivOweSuccess
VIF	2.61	1.83	3.50	1.72
	UnivRecommend	MonthWebiste	YearContact	ContactSatisfied
VIF	2.32	1.33	1.18	1.14
	WishVolunteer	YSG	HouseholdIncomeLevel	Gender
VIF	1.07	1.42	1.13	1.27
	Degree	School		
VIF	3.81	4.39		

number of interactions were considered with only YSG*WishVolunteer being significant in the Amount part.[8] The final variables in the model are:

- Amount part:

$$\begin{aligned} \boldsymbol{x}_a'\boldsymbol{\gamma} =& \gamma_0 + \gamma_1\text{YSG} + \gamma_2\text{HouseholdIncomeLevel} \\ & + \gamma_3\text{Connected} + \gamma_4\text{DonationImpact} + \gamma_5\text{StuExpLife} \\ & + \gamma_6\text{MonthWebsite} + \gamma_7\text{YSG*WishVolunteer} \end{aligned} \tag{2.9}$$

[8] Other interactions considered include YSG^2, and the effect of multiple degrees earned at the university.

- Frequency part:

$$\begin{aligned}\boldsymbol{x}_f'\boldsymbol{\beta} =&\beta_0 + \beta_1\text{YSG} + \beta_2\text{HouseholdIncomeLevel} + \beta_3\text{Connected}\\ &+ \beta_4\text{UseWordUs} + \beta_5\text{DonationImpact} + \beta_6\text{StuExpAcademic}\\ &+ \beta_7\text{StuExpOverall} + \beta_8\text{WishVolunteer} + \beta_9\text{SchoolLaw}\\ &+ \beta_10\text{SchoolArchitecture} + \beta_{11}\text{SchoolScience} + \beta_{12}\text{SchoolDesign}\\ &+ \beta_{13}\text{SchoolHumanScience} + \beta_{14}\text{SchoolAppliedTech} + \beta_{15}\text{SchoolBusiness}\end{aligned} \tag{2.10}$$

where the categorical variable "School" is represented by 7 dummy variables, with the type of "Engineering" being the benchmark type.

Selecting the number of population segments involves trading off goodness-of-fit, model complexity, and other considerations. In addition to the log-likelihood (LL), we consider criteria that account for both model complexity and goodness-of-fit. The criteria are Akaike Information Criteria (AIC), Bayesian Information Criteria (BIC), and Consistent Akaike Information Criteria (CAIC).[9] Table 2.5 and Figure 2.1 show the number of segments K and corresponding statistics for the model specification presented above. We observe that the log-likelihood increases sharply from 1 segment to 2 segments. The rate of increase slows for 3 or more segments. The AIC shows a similar improvement trend. The BIC and CAIC display elbow points at 2 segments. Based on the statistics, as well as the results presented in Section 2.5, i.e., balanced segments, we select a 3 segment model instance for further analysis.

[9] The goodness-of-fit statistics are calculated as $AIC(K) = -2LL + 2p_K$, $BIC(K) = -2LL + p_K \ln(N)$, and $CAIC(K) = -2LL + p_K(\ln(N) + 1)$, where LL is the log-likelihood, p_K is the number of parameters with K segments, and N is the number of individuals.

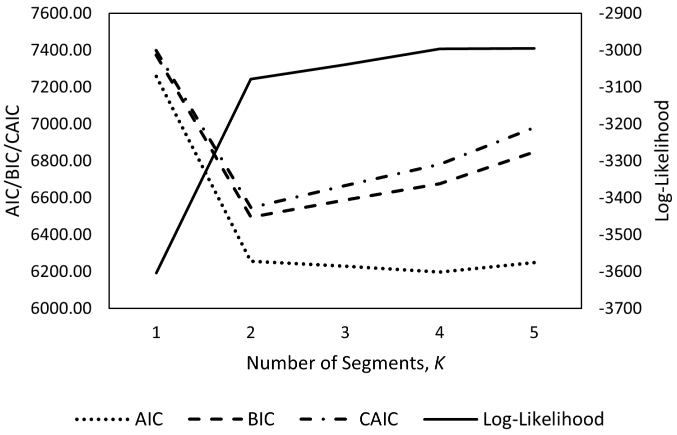


Figure 2.1. Log-Likelihood, AIC, BIC and CAIC as Function of Number of Segments K

Table 2.5. Model Selection Statistics

K	p	LL	AIC	BIC	CAIC
1	25	-3603.94	7257.882	7374.074	7399.93
2	51	-3077.73	6257.45	6494.48	6547.23
3	77	-3038.25	6230.49	6588.36	6668.02
4	103	-2995.55	6197.09	6675.80	6782.35
5	129	-2995.28	6248.56	6848.12	6981.55

2.5. Results and discussion

We begin this section by profiling the 3 segments in terms of the responses they display, i.e., donation amounts and frequencies, their composition given by SEED and EFR variables, and the marginal effects of the explanatory variables. Throughout, we compare the results obtained with the mixture model to those obtained for a population-level model, and discuss possible implications for fundraising.

Statistics describing the response variables are shown in Table 2.6. The segments are presented in decreasing order of average donation amounts, with Segment 1 accounting for about 25% of the population, Segment 2 for ≈50%, and Segment 3 for ≈25%. No segment appears to be unusually small, which is a possible sign of "overfitting" the data. The histograms in Figures 2.2a and 2.2b provide visual representations of the donation amount and frequency distributions for each of the segments. We observe that the 3 segments display different donation behavior. Specifically:

- Segments 2 and 3 display similar distributions of average donation amounts, whereas the distribution for individuals in Segment 1 displays a skew towards higher values, which explains why the average donation amounts for individuals in the segment are higher than those of individuals in the other segments.
- In terms of frequency, individuals in Segment 1 donate frequently, i.e., at a rate of 9 or more times the 17 year analysis period, whereas individuals in Segments 2 and 3 donate 1–4, and 4–14 times, respectively.[10] Interestingly, we observe that the mean donation frequency across the population is not representative of the mean frequency in any one of the 3 segments. Across the population we see that there is a clear distinction between sporadic donors in Segment 2, and recurring donors in the other segments – with each group comprising ≈ 50% of the population.

These results highlight key attributes of the CPMR model:

(1) Using a multidimensional response vector allows the model to capture behavioral differences manifested as heterogeneity in either donation amounts, frequencies, or both response variables. Models of one response variables are less capable, e.g.,

[10] Intervals are approximately ±1 standard deviation from the segment's mean donation amount.

an ordinary MR model of average donation amounts cannot reveal the aforementioned differences between segments 2 and 3.

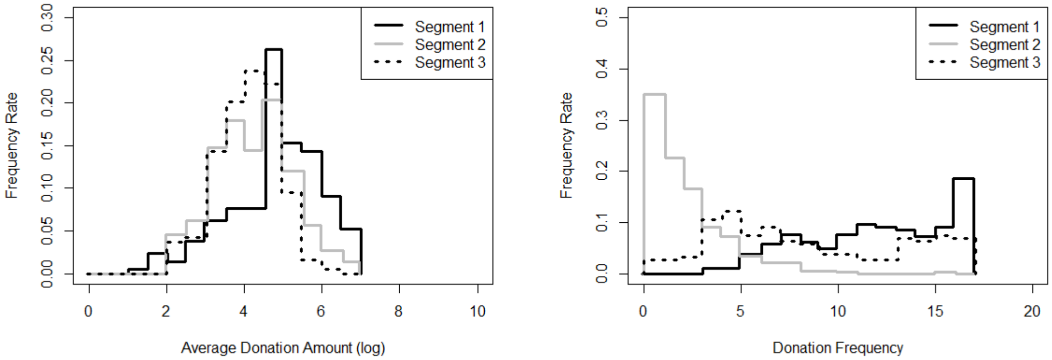
(2) The MR framework provides an appealing and rigorous approach to address technical issues arising in models of multidimensional response variables. In particular, the results in Table 2.6 and Figure 2.2 show that, the segmentation captures the correlation between the 2 response variables across the population. Specifically, individuals with high/(low) donation frequencies and large/(small) average donation amounts are grouped in Segment 1 (Segment 2).[11] We emphasize that this is a feature of the data that cannot be captured by separate analysis of the response variables, or that in the absence of of a multi-segment model, would require the specification of a structural equations model.

Table 2.6. Donation statistics: Population and segment levels

	Population	Segment 1	Segment 2	Segment 3
Proportion	100%	27.11%	48.38%	24.51%
Amount Part Statistics:				
Log Mean[12]	4.35	4.82	4.22	4.10
Exponential of Log Mean	77.80	124.40	68.13	60.14
S.D.	148.11	198.86	126.73	64.05
Frequency Part Statistics:				
Mean	6.91	12.45	2.73	9.02
S.D.	5.39	3.61	2.06	4.78
Hoeffding H-Statistic 95% Confidence Interval	[0.008,0.032]	[-0.006,0.006]	[0.000,0.023]	[-0.003,0.041]

[11] Notice that within each segment, the parameters of the Poisson and Log-Normal distributions display independence. The between-segment dependence exists, not in the distribution parameters, but in the observed $y_{f,i}$ and $y_{a,i}$ instead.

[12] Here, "log mean" specifically stands for mean of $y^l_{a,i} = \log y_{a,i}$, denoted by $\bar{y}^l_{a,i} = \frac{1}{N}\sum_{i=1}^{N} \log y_{a,i}$, since the mixture regression model is actually of $y^l_{a,i}$ instead of $y_{a,i}$. The "exponential of log mean" stands



(a) Log of Average Donation Amount

(b) Donation Frequency

Figure 2.2. Histograms of Donation Frequency and Amount

To examine the hypothesis of independent $y_{a,i}$ and $y_{f,i}$ within each segment, we conduct a Hoeffding's test of independence (Hoeffding, 1948). The test relies on a statistic, H, measuring the deviation from independence. Under the null hypothesis, 2 variables are independent when $H = 0$. To examine the hypothesis, 95% confidence interval of H-statistic is calculated for both the entire population and each segment. The results are in Table 2.6. Donation amount and frequency display statistically-significant dependence across the population, whereas the independence hypothesis is accepted for each of the 3 segments because the 95% confidence intervals include 0 in all cases.

It is of interest to examine the segment composition to tie differences in donation behavior to the explanatory variables. Table 2.7 summarizes the SEED and EFR factors for the population and each of the segments. The main observation is that, even though EFR variables are included in the analysis, the between-segment heterogeneity is

for $\exp(\bar{y}^l_{a,i})$, where we convert the mean of $y^l_{a,i}$ back to donation amount which is more intuitive. Also, since $y_{a,i}$ with Log-Normal distribution is highly skewed, taking exponential of log mean is less biased and more representative than directly taking the mean of $y_{a,i}$. Notice that the S.D. in amount part still stands for standard deviation of original $y_{a,i}$, without the log or exponential transformation.

unobserved, which in turn, further justifies the use of a *post-hoc* segmentation approach without a predetermined segmentation basis. The evidence is that the segment profiles are (almost) indistinguishable. The most noticeable differences are that Segment 1 has (slightly) higher than average proportions of Business School graduates, and (slightly) lower than average proportion of graduates from Other fields of study.

Finally, it is of interest to tie variation in the response variables to stratified effects of the explanatory variables across the population. Table 2.8 shows the regression coefficients obtained for the population and for the 3-segment model. The Amount and Frequency sub-models are separated because each includes different explanatory variables. The table shows that the level of significance, magnitude, and sign of the coefficients, capturing SEED and EFR effects, differ across the population segments.

Use of the log function to scale the donation amounts hinders interpretation of the coefficients in Table 2.8 for the Amount sub-model. Thus, we tabulate the corresponding percentage increases in donation amounts for unit increases in the explanatory variables. They appear in Table 2.9. The significance level marks are maintained for reference.[13]

The results in Tables 2.8 and 2.9 highlight differences in the factors that drive donation behavior within each segment. They also show that certain variables have compound effects on total receipts over the analysis period. The effect of the variables along with possible implications for fundraising are discussed below:

[13] The percentage increase is derived from marginal increase in log of amount. With estimated regression coefficient β_p of the p-th explanatory variable, we know that $\log y_{a,+1} - \log y_{a,0} = \beta_p$ with other explanatory variables fixed, where $y_{a,+1}$ stands for Amount after unit increase in the p-th variable, and $y_{a,0}$ stands for Amount before the unit increase. Based on the equation, it can be derived that: $y_{a,+1}/y_{a,0} = e^{\beta_p}$, i.e., from $y_{a,0}$ to $y_{a,+1}$ the Amount increase by a proportion of $e^{\beta_p} - 1$, which is the percentage number in Table 2.9. Rather than a 1 year change for YSG, we considered a 10 year increment. Thus, the results in Table 2.9 are for one additional decade since graduation, DSG, and the interaction term DSG*WishVolunteer. For all other variables, the increments are for 1 unit.

Table 2.7. Statistics of SEED and EFR factors: Population and 3 Segments

	Population	Segment 1	Segment 2	Segment 3
SEED Variable Mean:				
YSG	33.15	33.26	31.60	36.11
Household Income Level	5.86	5.72	5.85	6.02
Gender (% of Male)	79.51%	80.38%	78.02%	81.48%
Degree Type:				
Undergraduate	77.56%	77.51%	78.82%	75.13%
Master	19.58%	20.10%	18.77%	20.63%
Doctoral	2.72%	2.39%	2.14%	4.23%
Field of Study:				
Engineering	51.88%	53.11%	51.47%	51.32%
Architecture	9.99%	9.57%	10.46%	9.52%
Science	16.21%	15.31%	15.82%	17.99%
Business	10.89%	14.83%	9.65%	8.99%
Others	11.02%	7.18%	12.60%	12.17%
Survey Question Mean:				
Connected	3.96	4.10	3.84	4.04
FeelingsUniv	4.54	4.49	4.52	4.66
FeelingsAA	4.49	4.44	4.49	4.55
Competent	5.70	5.55	5.75	5.76
Similarity	3.57	3.63	3.46	3.72
UseWordUs	3.06	2.94	3.07	3.20
DonationImpact	4.02	4.24	3.82	4.17
StuExpAcademic	5.75	5.69	5.75	5.83
StuExpAcademicDept	5.91	5.89	5.87	6.00
StuExpLife	4.45	4.49	4.44	4.43
StuExpOverall	5.54	5.48	5.54	5.60
UnivOweSuccess	5.63	5.69	5.56	5.69
UnivRecommend	5.71	5.77	5.66	5.75
MonthWebsite	1.64	1.65	1.63	1.65
YearContact	4.03	4.08	4.05	3.94
ContactSatisfied	4.19	4.12	4.25	4.14
WishVolunteer	0.11	0.14	0.10	0.09

- Age, as captured in the YSG/DSG variable, has a significant, positive effect on donation amount. Across the population, an additional DSG is associated with

Table 2.8. Estimated Coefficients of the Population and 3-Segment Mixture Regression Model

	Variable	Population	3-Segment Mixture Regression Model		
		Model	Segment 1	Segment 2	Segment 3
Amount Part	(Intercept)	2.061***	1.758***	2.691***	0.988**
	YSG	0.023***	0.032***	0.018***	0.019***
	HouseholdIncomeLevel	0.146***	0.165***	0.122***	0.203***
	Connected	0.077**	0.021	0.048	0.189***
	DonationImpact	0.128***	0.199***	0.116**	0.050
	StuExpLife	-0.055**	-0.077*	-0.070*	0.012
	MonthWebsite	0.055	0.210*	-0.035	0.102
	YSG*WishVolunteer	0.007*	0.004	-0.001	0.019**
Frequency Part	(Intercept)	1.484***	2.624***	0.071	1.104***
	YSG	0.009***	-0.001	0.007**	0.023***
	HouseholdIncomeLevel	-0.060***	-0.063***	-0.055*	-0.011
	Connected	0.093***	0.078**	0.093**	0.027
	UseWordUs	-0.087***	-0.043*	-0.115***	-0.006
	DonationImpact	0.141***	0.070***	0.155***	0.059·
	StuExpAcademic	-0.076***	-0.073*	-0.088	-0.013
	StuExpOverall	0.056**	0.039	0.185**	0.010
	WishVolunteer	-0.188***	-0.200**	-0.217	-0.616·
	Field of Study:				
	Law	0.323	-0.407	1.190*	0.683
	Architecture	-0.221***	-0.017	-0.160	-0.798***
	Science	-0.071·	0.035	-0.283*	-0.096
	Design	-0.454***	-0.668·	-0.498*	0.016
	Human Sciences	-0.170**	-0.210	-0.216	-0.035
	Applied Technology	-0.063	-0.258	0.563*	-0.414
	Business	-0.104*	-0.073	-0.199	-0.522*

*** Significant at $p < 0.001$

** Significant at $0.001 \leq p < 0.01$

* Significant at $0.01 \leq p < 0.05$

· Significant at $0.05 \leq p < 0.10$

a 25.7% increase in average donation amounts. Segment 3 is especially interesting because it displays a 21.3% increase in average donation amounts, and 0.23

Table 2.9. Marginal Percentage Increases of Donation Amount with Unit Increases in Explanatory Variables

Variable	Population	3-Segment Mixture Regression Model		
	Model	Segment 1	Segment 2	Segment 3
DSG	25.7%***	37.8%***	19.6%***	21.3%***
HouseholdIncomeLevel	15.7%***	17.9%***	12.9%***	22.5%***
Connected	8.0%**	2.2%	4.9%	20.9%***
DonationImpact	13.6%***	22.1%***	12.3%**	5.1%
StuExpLife	-5.4%**	-7.4%*	-6.8%*	1.2%
MonthWebsite	5.6%	23.4%*	-3.4%	10.8%
DSG*WishVolunteer	6.8%*	4.1%	-0.8%	21.5%**

additional donations over the analysis period – a 40-year age difference means about 1 additional donation. A possible explanation for the increase is that children leaving the household – a factor not included in the present study – result in additional disposable income. Older children may result in additional availability, which may also explain the large interaction effect between DSG and willingness to volunteer for/through the university. This interaction means that, among alumni who wish to volunteer – 11% across the population –, an additional DSG is associated with a 25.7+6.8 = 32.5% increase in average donation amounts for the population. Among low-value, recurring donors in Segment 3, the corresponding percentage is 42.8%. Interestingly, the population displays a small but significant reduction in the number of donations among individuals with desire to volunteer (i.e.: WishVolunteer = 1). This effect is concentrated among high-value, frequent donors – Segment 1. To an extent, this may reflect a perceived substitution effect between money and time, where people may be more interested in committing time than money, or perhaps donate more frequently to activities that they are actively engaged in. On the whole, it seems there may be positive returns from creating opportunities for older donors to volunteer for/through the university.

- Not surprisingly, households with higher incomes donate larger amounts. For the population as a whole, a difference of 1 in an individual's self-reported income level is associated with 15.7% larger donation amounts. Interestingly, this effect is less pronounced among individuals in Segment 2 – sporadic donors. We also note that higher incomes are associated with fewer donations – the effect is not statistically-significant for Segment 3. Here, the tradeoff is negligible because the reduction is small (0.06 fewer donations over the analysis period for the population), but understanding the underlying cause may inform fundraising efforts. Households with higher incomes tend to be approached by more organizations, so the reduced number of donations could be the result of competition. Strategies such as modifying the solicitation schedule may prove effective.
- Perceptions of connection to the university (i.e.: the Connected variable) and of the impact of donations (i.e.: the DonationImpact variable) display significant compound effects on average donation amounts and frequencies. For example, considering 2 groups of 6 individuals with DonationImpact levels displaying a difference of 1, we note that 1 additional donation over the analysis period would be expected from the group with the higher rating. Also, their average donation amounts are expected to be 13.6% higher. Assuming that the group with the lower rating is consistent with the average statistics in Table 2.6, we have that, over a 17 year period, the university can expect to receive an additional

$6 \times (1 + 0.136) \times 77.80 \times (6.91 + 1) - 6 \times 77.80 \times 6.91 = \$969$ from the group with the higher rating. Interestingly, the stratification of these effects across the population provides the most important insight into the behavioral differences between sporadic donors in Segment 2 and the recurring donors in Segments 1 and 3. Specifically, we observe that, in Segments 1 and 3, respectively, the variables DonationImpact and Connected play critical roles in explaining within-segment variation in donation amounts. This means that average donation amounts of individuals in these segments display great sensitivity to these factors. Donation amounts of sporadic donors in Segment 2 are influenced, but (far) less sensitive to changes in these variables. Both variables display small, but significant positive effects on donation frequencies for all segments – a result that is consistent with results in other charitable giving settings (Levontin et al., 2015). From a fundraising standpoint, the university can influence perceptions of connection by, for example, distributing (print or digital) newsletters, and other materials. Messaging that emphasizes the impact of donations would appear to be effective. Including links to the university's websites may be an especially powerful way to engage individuals in Segment 1, who seem to respond favorably to the content therein. An additional website visit per month is associated with a 23.4% increase in average donation amounts.

- The experience variables, i.e., StuExpAcademic, StuExpLife, and StuExpOverall, display heterogeneous effects on donation amounts and frequencies. At times, these effects are small, insignificant, or counterintuitive, i.e., with negative signs. Generally, affecting these variables requires large, long-term investments, and the results suggest that the effect on fundraising is mixed/unclear. Because these results are unusual, discussion of the effect of experience variables appears in C.

- Among the education variables, with the exception of Law students (in Segment 2), the fields of study have large and significant negative effects on donation frequency. As explained in Section 2.4.3, these effects capture the marginal effect relative to the benchmark, i.e., engineering students. This means that, among individuals in Segment 2, Law students donated 1.2 more times on average than engineering students over the analysis period. While our example was for the exception, the fact that these effects display negative signs suggests the need to develop strategies that address nonengineer's concerns and interests.
- We would be remiss not to mention that, in our data set, a number of variables display insignificant marginal effects across all population segments, i.e., FeelingsUniv, FeelingsAA, Competent, Similarity, UseWordUs, StuExpAcademciDept, UnivOweSuccess, UnivRecommend, YearContact, and ContactSatisfied. This means that variation in average donation amounts or frequencies is not explained by (systematic) differences in the survey responses used to assess the variables. Referring to Table 2.7, we observe that in most cases, cases we see high levels of satisfaction/approval (with relatively little variation across the donor population as measured by their coefficients of variation). This is important because feelings of gratitude (i.e.: UnivOweSuccess) and donors' perceptions about an organization's competence (i.e.: Competent) have been proposed as factors that influence recurring donations (Levontin et al., 2015).

2.6. Conclusion

We present an innovative Compound Poisson Mixture Regression model of transaction frequency and monetary value, and use it to study donations at a private university in the Midwestern United States. The model is appealing because it captures the joint effect of covariates, and thus, avoids biases introduced in separate analyses of the response variables. Moreover, the mixture regression framework provides a rigorous and appealing approach to account for heterogeneity and other features (i.e.: correlation between the response variables) in the data. In particular, the framework relies on the assumption that unobserved/latent, group-level factors are manifested through coefficients that are allowed to vary across the different segments comprising a population. The approach, therefore, can be used to evaluate factors that drive response variables across segments, and thus, to develop marketing strategies that exploit such differences.

The data for the study come from 2 sources: donation records for the 2000-2016 period, which included information on socioeconomic, education, and demographic (SEED) factors, as well as an alumni survey conducted in the Fall of 2017, which was designed to elicit information about donor experience as a student and university alumnus, their feelings towards and perception of the relationship with the university. The former is important because there is a dearth of empirical studies analyzing repeated transactions, i.e., longitudinal data. The latter is important because data related to EFR factors are often missing.

The results highlight features of the proposed model, and lead to insights with potential to improve fundraising efforts. Specifically,

- We show that the proposed model captures behavioral differences manifested as heterogeneity in either donation amounts, frequencies, or both response variables. In particular, the results show that the donor population can be reasonably split into 3 segments, with Segments 1 and 3, respectively, being composed of high and low value, recurring donors, and Segment 2 of low-value, sporadic donors. Donation amounts for individuals in Segments 2 and 3 are similar, which means that models of 1-dimensional response variables, e.g., a MR model of average donation amounts, are not capable of accounting for the difference between individuals in Segments 2 and 3.

  In spite of the inclusion of EFR factors, the empirical results suggest that the between-segment differences between the response variables are not explained by the available data, i.e., the between-segment heterogeneity is unobserved. The evidence is that the segment profiles, presented in Table 2.7, are indistinguishable. In turn, this emphasizes the importance of building a *post hoc* segmentation framework that does not rely on a predetermined segmentation basis.

- Table 2.8 shows that the level of significance or magnitude of the coefficients, capturing SEED and EFR effects, differ across the population segments and for each of the 2 response variables. The implication is that, depending on the segment they belong to, individuals display heterogeneous responses, i.e., sensitivities, to differences/changes in the factors, which can support (tailored) fundraising strategies.

  For example, as in synergistic studies, we find that age (i.e.: YSG/DSG) and household income have significant effects on average donation amounts. Our model, however, reveals that these effects are stratified, and particularly strong among low-value, recurring donors with interest in volunteering for/through the university. Among donors in this group an additional decade since graduation leads to a 42.8% increase in average donation amounts. It follows that offering such opportunities to donors in this segment may prove to be an effective fundraising strategy.

Interestingly, we observe that the behavioral differences between sporadic donors in Segment 2 and the recurring donors in Segments 1 and 3 are tied to the marginal effects of the variables describing donors' perceptions of the impact of their donations and of their psychological distance. These results are aligned with the literature on recurring donations in charitable-giving settings, and while they are a function of the data, they reinforce the importance of the survey to explain behavior, which is an important contribution of the present study.

In terms of directions for research, we note that, while adequate, diagnosis shows opportunities to improve/generalize the proposed model. For example, the estimation results suggest that the assumption of transactions following a Poisson Distribution may be strong. The mean and variance of the segment-level distributions are not equal. Fine-tuning the model specification by using the Negative Binomial Distribution, or other strategy to generalize the model may be needed to address the latter.

CHAPTER 3

Controlling for response styles to infer organization performance from subjective survey data: Using a Rasch Model to refine the the predictors of donation behavior

3.1. Introduction

Survey data have become a popular and important tool in marketing research for collecting information from customers. Advantages of surveys, compared to recording long-term transaction data and interviewing customers, may include lower costs, faster data collection, and higher response rates (Ilieva et al., 2002). Also, as explained by Ping Jr (2004) and Yang et al. (2010), marketing research relies on survey data when factors related to customer behavior are unobserved, e.g., attitudes, experiences, feelings, perceived relationships, etc. In such surveys, customers are often asked to select a single number within a discrete and ordinal set – known as Likert scale – to rate their level of agreement with a given question/statement. Rating scores, in turn, are used to make inferences that, ultimately, may support marketing strategies that target (potential) customers.

Among the issues that have been raised/analyzed in the analysis of survey data, we note the importance of accounting for intentional or inadvertent errors (Gaba and Winkler, 1992), as well as a number of factors that may result in biases, e.g., underreporting due to respondent's decline in recording accuracy (Yang et al., 2010), data manipulation

(Friedman and Amoo, 1999), etc. Even when common errors and biases in data collection are controlled for, the issue of accounting for rating scale biases often persists. An individual's *rating scale* or *response style* is tied to the process by which they convert subjective assessments into numerical scores representing levels within the given scales/sets. The issue is that the same rating score from different individuals may actually reflect different underlying feelings or beliefs, while at the same time, individuals reporting different ratings may actually share similar feelings. It follows that ignoring rating scale biases may undermine the reliability of inferences and subsequent analysis Greenleaf (1992).

To address the issue, here, we present a method to estimate and control for individual rating scale biases in survey data. Specifically, we use a Partial Credit Model, a generalization of the Rasch Model, to estimate the effect of both individual rating scales, and of question performance from survey responses. The ensuing score residuals, i.e., the difference between rating scores and their expectations obtained from the model, are used as bias-corrected rating scores. Subtracting the expectations from the original rating scores removes systematic rating scale biases, while preserving variation in an individual's underlying assessment of their feelings/beliefs. The method, therefore, provides a correction approach that relies exclusively on the survey data. This is in contrast to approaches appearing in the literature that use additional, typically objective information for "anchoring" purposes.[1] An additional advantage of the proposed approach is related

[1] Use of "anchoring" variables dates back to the Behaviorally-Anchored Rating Scale method of Schwab et al. (1975), where a set of directly-observed behavior variables are rated, and subsequently used to correct for unobserved biases in in subjective rating data. Greenleaf (1992), one of the important benchmarks for our work, also uses a set of objective survey responses that are assumed to be bias-free to estimate and correct for rating scale biases in subjective survey responses.

to generalizations of the Rasch Model, which not only provide flexibility in modeling individual rating scales and question performance, but also can capture segment or population level effects in cases where different levels of aggregation are desired. The approach is appealing because it yields bias-free survey data to improve robustness and stability of (further) analysis/development, to capture group-level effects, and, ultimately, to support the development of marketing strategies.

To illustrate the proposed method, we apply the Partial Credit Model to analyze donation/fundraising data at a private university in the Midwestern United States. Our analysis combines data from an alumni survey conducted in the Fall of 2017 with donation records for the 17 year period between 2000-2016. The objective is to establish relationships between donation behavior, i.e., transaction frequency, monetary value, and within-donation-sequence variation, and donor experiences as students/graduates, as well as their feelings towards and their perceived relationship with (groups at) the institution. In addition to estimating a Partial Credit Model to obtain the effects of individual rating scale biases and of question performance, we also consider a variety of models that describe variables representing donation behavior as a function of explanatory variables from the survey data. The results show that, compared to the original data, the bias-corrected survey data improves the predictive power of all of the models considered. Importantly, the results show that the marginal effects of certain variables change after bias-correction in both sign and level of significance. The results, therefore, reinforce the importance of controlling for rating scale bias in survey data, and the effectiveness of proposed correction method.

The remainder of the paper is organized as follows. Section 3.2 position our work with respect to literature of controlling the rating scale bias and getting corrected measure of subjective attitude. In Section 3.3, we introduce Rasch model and our proposed rating scale bias correction method based on Rasch model, and discuss comparison with previous literature in model details. In Section 3.4, we first describe the data used for application of alumni survey and donation, and then present results of bias-correction and predictive models on alumni donation with alumni survey data. Discussions on performance of bias-correction are also presented. Finally, conclusions and limitations appear in Section 3.5.

3.2. Literature review

This paper proposes a method based on the Rasch Model to control for rating scale biases in survey data. We consider a university fundraising application to illustrate the importance of correcting such biases. In this section, we position our work with respect to literature on estimating and correcting for rating scale biases. We also review applications of using rating scores in marketing research, with a focus on university fundraising and overall charitable-giving.

Survey data on attitudes, feelings, and perceptions have been widely used in marketing research. Huang and Sudhir (2021), for example, use a survey to estimate causal effects of service satisfaction on customer loyalty. Use of survey data extends to a wide variety of charitable-giving applications. For example, in the context of university fundraising, Gaier (2005) establish a significant relationship between undergraduate student experience, as reported in an alumni survey, and the likelihood of donations. In a general charitable-giving study, Aaker et al. (2010) use survey data on donor's impression of competence and sense of closeness toward organization to show their positive impacts on donation behavior.

However, most studies using survey data do not account for the effect of rating scale biases, which can alter the distribution of subjective variables, and thus, the properties, i.e., precision, accuracy, marginal effects, etc., of statistical models that employ these variables. For example, Greenleaf (1992) present a synthetic example to illustrate how rating scale biases can distort the results of cluster analyses based on survey responses. They use the example to argue there is a need to accurately control for rating scale biases in order to strengthen the reliability of analyses and inference based on survey data.

In statistics/psychometrics, the analysis of individual rating scales falls under Item Response Theory (IRT). In IRT, outcome scores appearing in survey responses are assumed to be result of 2 factors: individual rating scales, which are called Person parameters; and question performance/difficulty, which are referred to as Item parameters. While IRT models are commonly used for modeling test results (Lord, 1980), where Person parameters represent individual abilities and Item parameters represent question difficulty, they are also applicable in the analysis of survey response data, especially ordinal rating questions (Piquero et al., 2000). In such cases, Person parameters are interpreted as generosity or tendency to give higher rating scores to questions, and Item parameters interpreted as question performance influencing individuals to give higher rating scores. In IRT modeling, Rasch Models (Rasch, 1961) are common tools in the analysis survey or questionnaire responses, or test results. In its basic form, the dichotomous Rasch Model, i.e., when the response variable is binary, captures the probability of an individual answering correctly or giving a positive answer to a question. The model uses a hierarchical structure as

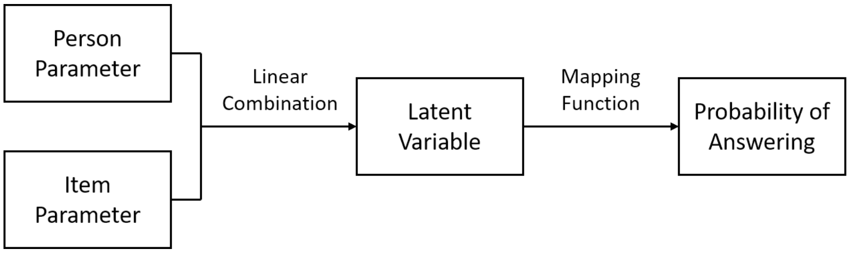


Figure 3.1. Hierarchical Structure of Rasch Model

shown in Figure 3.1, where it introduces a latent variable as linear combination of Person and Item parameters to capture both effects, and uses a mapping function to map the latent variable to probability of giving positive answer. In the basic dichotomous case, latent variable is the difference between Person and Item parameter, and a binary logistic function is used as mapping function. Several extensions have appeared to capture more complicated situations with different forms of questions, especially generalizing to polytomous response cases and removing specific assumptions. The Rating Scale Model (Andrich, 1978), for example, is a generalization of Rasch Model to the polytomous cases. To formulate the multi-category response, it assumes that the difference between categories of answers is equal within and also across items (questions), and adopts an ordinal multinomial logit function for the polytomous case. The Partial Credit Model (Masters, 1982) provides a further generalization where Category parameters of different items are allowed to differ.

Approaches other than the Rasch Model have been proposed, in the marketing research context, to account for individual rating scale biases. Most studies use a similar structure as Rasch Models, with extensions on modeling the individual rating scale and different mapping functions to map Person and Item parameters to rating score. For example, Javaras and Ripley (2007) propose a Multidimensional Unfolding Model that captures the Person parameter as a function of linear regression of individual characteristic factors (e.g., age, gender, country, etc.), and instead of logistic function as used by Rasch Model, it uses a series of ordered threshold points as the mapping function, that difference between Person and Item parameter falling between two certain threshold points leads to corresponding category of the item. Similar with Rasch Model, the estimated Item parameter across the population is used for inference on individual's common attitude as "adjusted response", and estimated coefficients of linear regression function show how rating scale is affected by individual characteristics. Grün and Dolnicar (2016) adopt a similar model, but introduce a finite mixture model framework to allow group-level heterogeneity in true beliefs and linear regression coefficients in Person parameter. Van Rosmalen et al. (2010) propose a Latent-Class Bilinear Multinomial Logit model that uses a similar finite mixture model framework, while adopting multinomial logit function as the mapping function, unlike the ordered thresholds adopted in Javaras and Ripley (2007) and Grün and Dolnicar (2016). However, Van Rosmalen et al. (2010) use a nominal multinomial logit function, and thus the model does not deal with ordinal data. These studies focus on making inferences based on rating scale adjustments by estimated population or group-level beliefs. To get bias-corrected survey data for further analysis, the traditional approach is to standardize rating scores of each individual by their mean and standard deviation of response scores. However, Fischer (2004) argue against this method, for potential negative effects such as occurrence of spurious method factors in factor analysis, and the complication of interpretation. Greenleaf (1992) propose a hypothesis testing approach to examine if bias exists in yeasaying (individual mean) and standard deviation in response style. If bias is detected, a linear model is developed to quantify biases from both yeasaying and standard deviation, and linear combination of original score and style-eliminated score is used to get bias-corrected data. This seems to be the the first study providing bias-corrected survey data to be used for subsequent analysis. However, the approach requires "anchoring" variables to serve as response variable in the linear model, which are usually objective facts or behavior of individuals that are assumed to be bias-free, in order to control for the heterogeneous response styles. Such requirement significantly limits application of survey data bias-correction in cases where additional information is not available. Importantly, this method ignores the effect of specific questions in the survey, while the rating scale bias of an individual can also vary across different questions.

The method proposed in this paper uses a Rasch Model to conduct bias-correction. As opposed to studies that estimate population or group-level true beliefs that are free of the impact of rating scale bias, this study presents a practical method to provide bias-corrected rating score data for each question by each individual. These data can subsequently be utilized for further analysis and inference. In contrast to Greenleaf (1992), the proposed method doesn't rely on availability of additional objective information that are assumed to be bias-free to quantify existing biases. That is, the proposed approach relies exclusively on rating score data from survey for self-correction. In addition, instead of using simple individual mean value, impact of specific questions are taken into consideration for the bias to provide bias correction on an individual-question level. A practical contribution of the paper is that, an application in marketing is conducted, which shows both the effectiveness of the propose method and also unreliability of inference obtained from biased survey data.

From a marketing perspective, this is significant in improving efficiency and accuracy of developing tailored marketing strategies.

3.3. Methodology

We begin this section by describing the Partial Credit Model, a generalization of the Rasch Model to the case of polytomous responses, which we adopt in our analysis. We then describe the use of the score residuals as statistics that control for rating scale biases in survey responses. Throughout, we further position our work with respect to previous literature to highlight advantages of the proposed approach.

3.3.1. Partial Credit Model

In the IRT framework, the assumption is that rating scores in survey responses are the result of 2 factors: (i) individual rating scales, and (ii) item/question characteristics, e.g., difficulty. Person and Item parameters in the Rasch Model are used to capture the effect of these 2 factors. In the model, widely used to describe test results, the dichotomous random variable $X_{ik} \in \{0, 1\}$ is used to represent outcomes. $X_{ik} = 1$ corresponds to the event that the ith individual responds to the kth question correctly $i = 1, 2, ..., N$ and $k = 1, 2, ..., K$, and $X_{ik} = 0$ denotes an incorrect answer. The probability of $X_{ik} = 1$ is is given by:

$$P\left(X_{ik} = 1\right) = \frac{\exp(\theta_i - \delta_k)}{1 + \exp(\theta_i - \delta_k)} \tag{3.1}$$

where θ_i is individual i's Person parameter, and δ_k is question k's Item parameter. It follows that $P(X_{ik} = 0) = 1 - P(X_{ik} = 1)$.

The Partial Credit Model is an extension of the Rasch Model to polytomous cases. In terms of notation, $M_k + 1$ corresponds to the number of categories or rating levels available for item k. These levels are numbered from 0 to M_k (for a total of $M_k + 1$), and thus, $X_{ik} \in \{0, 1, ..., M_k\}$. The assumption in the Partial Credit Model is that the probability of observing a given score depends on the specific category/level in the specific item/question. That is, the marginal effects of of switching to an adjacent category/level is not constant, and thus, there is a need to include the Item-Category parameters β_{km} in the model. Now, the probability of individual i giving answer m to item k is given by:

$$P\left(X_{ik} = m\right) = \frac{\exp\left(m\theta_i + \beta_{km}\right)}{\sum_{l=0}^{M_k} \exp\left(l\theta_i + \beta_{kl}\right)} \tag{3.2}$$

where, again, θ_i is individual i's Person parameter, and now, β_{km} is an Item-Category parameter for category/level m in question k.[2] Note that Partial Credit Model handles the multi-category response as ordinal data (i.e., category 0 to M_k are considered to be ordered). This is derived by considering each shift from category $m - 1$ to m as a dichotomous case similar to Equation 3.1, and with constraint of all probabilities sum to 1, solving the equations leads to the final formulation as shown by Equation 3.2. Details of derivation can be found in Masters (1982).

[2] The signs that precede δ_k in Equation 3.1 and β_{km} in Equation 3.2 can be reversed, depending on the interpretation of Item-Category parameter, e.g., negative signs are used when the parameter reflects a question's difficulty. We use different signs in the expressions to ensure consistency with the original models presented in Rasch (1993) and Masters (1982).

To estimate the Person and Item-Category parameters, two types of methods – Marginal Maximum Likelihood (MML) estimation and Conditional Maximum Likelihood (CML) estimation – are commonly used. In this study, we relied on the CML estimation to estimate those parameters. Details can be found in Mair and Hatzinger (2007) describing the R package *eRM*, which we use.

3.3.1.1. Threshold and Location Parameters of Partial Credit Model. For a given item, k and performance level/category, m, the Threshold parameter, T_{km}, is defined as the Person parameter that makes adjacent categories, $m-1$ and m, equally probable, i.e., $P\left(X_{ik}=m-1\right) = P\left(X_{ik}=m\right)$. This means that for individual i, if $\theta_i > T_{km}$, $P\left(X_{ik}=m\right) > P\left(X_{ik}=m-1\right)$, and vice versa. Evaluating (3.2) when $P\left(X_{ik}=m\right)$ and $P\left(X_{ik}=m-1\right)$ yields:

$$T_{km} = \beta_{k,m-1} - \beta_{km} \tag{3.3}$$

For item k, T_{km} corresponds to the break-even Person parameter between categories $m-1$ and m. In terms of the Person parameter, which, for example, reflects overall satisfaction, each threshold captures the increased difficulty of receiving a score of m instead of $m-1$ for $m = 1, 2, ..., M_k$. It is often used because it is more intuitive than the Item-Category parameter, β_{km}. As presented in (3.4), the Location parameter, L_k, is defined as the average of the Threshold parameters for an item k. It provides an overall measure of an item's/question's characteristics.

$$L_k = \frac{1}{M_k} \sum_{m=1}^{M_k} T_{km} \tag{3.4}$$

Comparison of the Location and Threshold parameters across items yield a relative representation of the difficulty of a question, i.e., the Threshold parameters capture the difficulty of achieving certain proficiency levels. Items with higher Location and Threshold parameters are more difficult, since higher Person parameters are needed to achieve corresponding proficiency levels, i.e., scores. Therefore, when visualizing and interpreting the result of Partial Credit Model and other Rasch Models, Location and Threshold parameters are more common than Item or Item-Category parameters, δ_k or β_{km}.

3.3.1.2. Diagnosis of Partial Credit Models. Following (3.2), estimation of Person and Item-Category parameters allows for the calculation of the probability mass functions of the random variables X_{ik}, i.e., $P\left(X_{ik} = m\right), m = 0, \ldots, M_k, k = 1, \ldots, K, i = 1, \ldots, N$. The associated expectations, E_{ik}, and variances W_{ik}, are given by:

$$E_{ik} = \sum_{m=0}^{M_k} m \cdot P(X_{ik} = m) \tag{3.5}$$

$$W_{ik} = \sum_{m=0}^{M_k} (m - E_{ik})^2 \cdot P(X_{ik} = m) \tag{3.6}$$

The *score residuals*, defined in (3.7), correspond to the difference between the observed and expected rating scores. That is, how far does the observation deviate from its expectation.

$$y_{ik} = X_{ik} - E_{ik} \tag{3.7}$$

Diagnosis of Partial Credit Models relies on the unweighted and variance-weighted *mean square residuals* (MSQ) for item/question k. The former is referred to as the Outfit, and the latter as the Infit MSQ. Both statistics have a mean of 1. To further standardize the outfit and infit MSQ, Wright and Masters (1982) presents a transformation to t-statistics which have means of 0 and standard errors of 1. They are referred to as Outfit t and Infit t, respectively. These MSQs and t statistics are the metrics to assess goodness-of-fit of Rasch Models. In particular, when the model fits the data well, the Outfit and Infit MSQs are expected to be close to 1, and the Outfit and Infit ts should be close to 0. Infit statistics are preferred to the corresponding Outfit statistics because they take account for individual weights. Two common rules-of-thumb are to consider Infit MSQs between 0.7 and 1.3, and Infit ts between -2 and 2 as indications of adequate model fits. Smith et al. (2008) conducted an experiment to compare the 2 statistics. The results show that for polytomous Rasch Models, t-statistics were highly sensitive to sample size, while MSQ statistics remain relatively stable. Thus, in Section 3.4.2 below, we use Infit MSQs statistic in the interval $[0.7, 1.3]$ as indications of a well fit model.[3]

[3] Items/Questions with large MSQs or t statistics are are said to "underfit" the data. Items/Questions with small MSQs or t statistics are referred to as "overfit" the data. Even though observed responses appear to be (highly) consistent with the underlying model assumptions, a lack of variation in the responses to

In addition to model diagnosis, we use the score residuals, y_{ik}, as each individual's corrected rating for item/question k; that is, to replace the original response data. The idea is that individual i's rating on item k, X_{ik}, can be decomposed into two parts. The first is the expected score, E_{ik}, which captures the question characteristics, described by the parameters β_{km}, and the individual's rating scale, described by θ_i and representing the individual's underlying satisfaction with the organization. The second part is the score residual, y_{ik}. A score residual of 0 means that individual i's response to item/question k is as expected, i.e., neutral, whereas positive and negative score residuals, respectively, stem from higher or lower than expected ratings. Ignoring inherent randomness, positive/negative score residuals represent areas/items/questions where an individual's satisfaction with the organization over/under-performs expectations.

With an adequate model, because β_{km} are constant for all individuals, variation in X_{ik} stems from (i) individual rating scales representing overall satisfaction, (ii) factors not captured in the model explaining performance in a given dimension that differs from expectations, and (iii) inherent randomness. For a large sample (iii) is expected to be close to 0.

3.4. Application to university fundraising

We begin this section by describing the data used in our analysis: both the alumni survey, as well as the donation records. We then describe the development of a Partial Credit Model, including assessment and discussion of the parameter estimates and goodness-of-fit statistics. In Section 3.4.3, we use the Person and Item-Category parameters from

a specific question may be an indication of an inappropriate question setting, or of the influence of other external factors.

the Partial Credit model to correct the rating scale biases in the survey responses. The corrected data are used to estimate a number of predictive models describing donation behavior. The models display superior predictive capabilities to those obtained with the original survey responses.

3.4.1. Data

The data used in this study are from 2 sources: (i) alumni donation records held by the university for the 17 year period between 2000–2016; and (ii) a survey conducted in the Fall of 2017. The survey's objective was to elicit information about alumni experiences as students at and as graduates of the university, as well their perceived relationship with the institution. Our final data set consists of donation records and survey responses for 1,934 alumni – obtained from 2,859 survey respondents – who graduated in 2014 or earlier. Of the respondents, 849 alumni donated at least once during the 2000-20016 analysis period. Such individuals are referred to as donors, whereas all others are referred to as non-donors. The data used in our analysis, including the construction of predictor and response variables, as well as further explanations about discarded records, are described in the remainder of the section.

A total of 4 response variables are used in our study to describe donation behavior: (i) donor/non-donor; (ii) donation frequency; (iii) average donation amount (iv) donation behavior/pattern assignment. For all survey respondents, binary response variable of donor/non-donor are used, where 1 represents donor and 0 represents non-donor. For donors, the 3 response variables used are donation frequency, average donation amount ($/year), and donation pattern. For example, an individual with donations in 4 out of the 17 years in the period 2000 – 2016 with respective totals \$100, \$200, \$100, and \$150, and with no donations recorded in the remaining 13 years, the corresponding donation frequency and average amount are, respectively, 4 and $(100+200+100+150)/4 = \$137.5$. For individuals graduating after the year 2000, the donation frequency was rescaled by estimating the number of donations over a 17 year period, N_i' as $N_i' = \left\| N_i \frac{17}{\min\{17, 2017-G_i\}} \right\|$, where G_i is the graduation year, N_i is the actual number of donations received, and $\|\cdot\|$ is the rounding operator. To avoid overestimating the donation frequency of recent graduates, alumni graduating after 2014 were excluded.

For the donation behavior assignment variable, we adopted the segmentation assignment result by applying the Markov chain mixture model by Durango-Cohen et al. (2013a). The Markov Chain Mixture Model (MCMM) considers each individual's sequence of annual donation amounts, and assigns them to segments characterized by the variation/pattern in their sequences. Details of the MCMM are presented in Appendix D. We fitted a MCMM to the donation data and assigned each donor to one of 3 segments, labeled as "Low Variance" (LV), "High Variance" (HV), and "Transient" (TS), respectively, depending on the variation displayed in their contribution sequences. The matrices describing the each segment's dynamics are presented in Appendix D. The LV and HV segments are for recurring donors, whereas the TS segment includes one-time and sporadic donors. The LV segment includes individuals whose donations display consistency in terms of the amounts they donate each year, i.e., they tend to remain at the same level, whereas year-to-year donation amounts for individuals in the HV segment vary significantly, i.e., they transition between levels. The High Variance (HV) segment shows a much higher likelihood that alumni transition between different donation levels, as compared to LV and TS segments. The transition probabilities of each segment are presented in Table D.2 in Appendix D. Durango-Cohen et al. (2013a) argue that alumni in HV segment may be sensitive/responsive to appeals/solicitations. Such classification into the 3 segments is called behavior assignment variable in this study, which better describes the donation behavior pattern of alumni.

The explanatory variables in the models are from the alumni survey, which included 14 questions requiring alumni to assess their level of agreement with the each item's description/text. The questions, including the response scales as appearing in the survey, are presented in Appendix E. To simplify and make the analysis consistent with the description in Section 3.3, the scales were mapped to levels/categories $0, 1, \ldots, 6$. Missing responses for the 1,934 alumni in the analysis, as reported in Table 3.1, were dealt with by conducting data imputation using the R package, *mice*. Missing values were generated by sampling from conditional distributions obtained from the data. Additional details of the data imputation methodology and our processing are presented in Appendix A and in Buuren and Groothuis-Oudshoorn (2010). Table 3.1 also shows the minimum, median, mean, maximum, and standard deviation of each of the survey responses in the final dataset after imputation. Table 3.2 presents statistics describing the donation behavior of the 849 donors from the 1,934 respondents. The statistics include the proportion of donors assigned to each of the 3 behavioral segments.

Table 3.1. Summary of Alumni Survey Variables

Variable	Min	Median	Mean	Max	S.D.	Missing %
StuExpAcademicDept	0	5	4.63	6	1.29	0.00%
StuExpAcademic	0	5	4.48	6	1.22	0.00%
Competent	0	5	4.40	6	1.38	0.52%
UnivRecommend	0	4	4.27	6	1.52	0.00%
StuExpOverall	0	4	4.19	6	1.29	0.00%
UnivOweSuccess	0	4	4.12	6	1.66	0.00%
FeelingsUniv	0	3	3.25	6	1.44	0.67%
ContactSatisfied	0	3	3.21	6	0.97	0.31%
FeelingsAA	0	3	3.19	6	1.44	0.98%
StuExpLife	0	3	3.15	6	1.73	0.00%
DonationImpactful	0	3	2.67	6	1.60	0.00%
Connected	0	3	2.60	6	1.70	0.00%
Similarity	0	2	2.22	6	1.62	1.24%
UseWordUs	0	2	1.92	6	1.78	0.31%

Table 3.2. Summary of Donation Variables of Donor Alumni

Variable	Min	Median	Mean	Max	S.D
Donor Proportion			43.9%		
Average Donation Amount	3.67	92.00	201.18	4000.00	394.91
Log Average Donation Amount	1.30	4.52	4.49	8.29	1.21
Donation Frequency	1	5	6.91	17	5.36
Behavior Assignment – Low Variance	–	–	14.5%	–	–
Behavior Assignment – High Variance	–	–	25.5%	–	–
Behavior Assignment – Transient	–	–	60.0%	–	–

3.4.2. Model estimation and bias correction

For the estimated Partial Credit model, the 4 goodness-of-fit statistics of each survey variable are shown in Table 3.3. As discussed in Section 3.3.1, variables with Infit MSQ between 0.7 and 1.3 are considered to be well-fitted. They are marked with an asterisk symbol (*) in Table 3.3. The 4 variables FeelingsUniv, StuExpOverall, UnivRecommend,

Table 3.3. Item Fit Statistics of Survey Variables

Variable	Outfit MSQ	Infit MSQ	Outfit t	Infit t	
Connected	0.898	0.883	-3.40	-4.08	*
FeelingsUniv	0.695	0.699	-11.08	-10.93	
FeelingsAA	0.826	0.822	-5.96	-6.10	*
Competent	0.945	0.954	-1.62	-1.38	*
Similarity	0.918	0.916	-2.61	-2.82	*
UseWordUs	1.016	1.005	0.43	0.16	*
DonationImpactful	1.094	1.07	2.97	2.26	*
StuExpAcademic	0.757	0.775	-7.84	-7.11	*
StuExpAcademicDept	0.867	0.896	-3.76	-2.90	*
StuExpLife	1.243	1.194	7.25	6.15	*
StuExpOverall	0.603	0.599	-14.21	-14.37	
UnivOweSuccess	0.957	0.958	-1.22	-1.28	*
UnivRecommend	0.67	0.685	-10.82	-10.75	
ContactSatisfied	1.794	1.74	11.81	13.36	

and ContactSatisfied are identified as misfit items, and thus, are excluded from subsequent analysis.

Figure 3.2 shows the Person-Item Map of the Partial Credit Model, which summarizes the estimation results. The top part of the figure is histogram showing distribution of estimated Person parameters, θ_i. The bottom part of the figure presents the Threshold and Location parameters of all well-fitted items from Table 3.3. The open dots with number labels are the corresponding m-th Threshold parameters. The solid dots are the Item Location parameters for each of the 10 items. The unitless x-axis presented in the figure is shared by Person, Threshold, and Location parameters, θ_i, T_{km}, and L_k, respectively. Ticks under histogram of Person parameters are reference points of Threshold and Location parameters, for the convenience of comparing between top and bottom part.

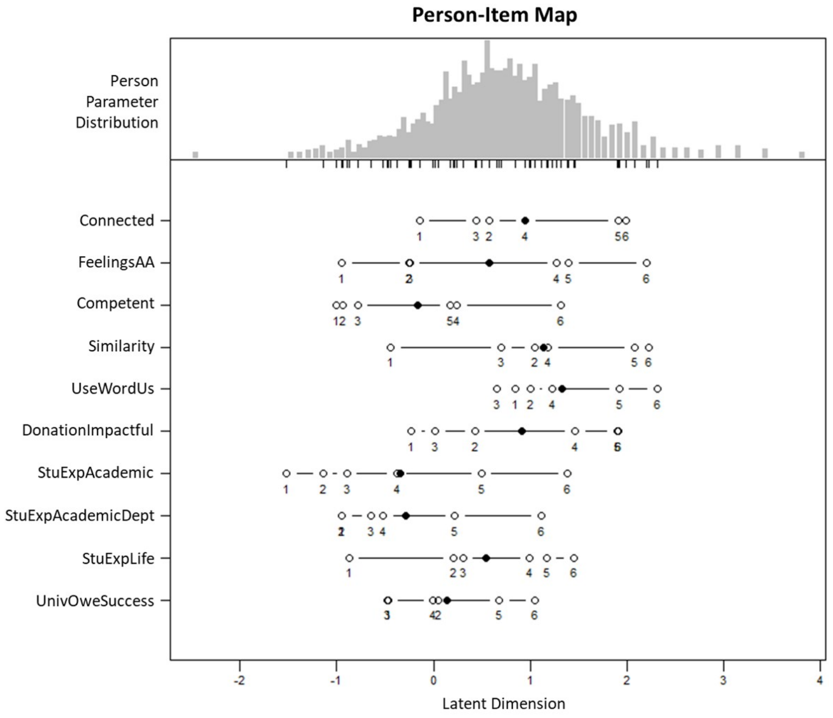
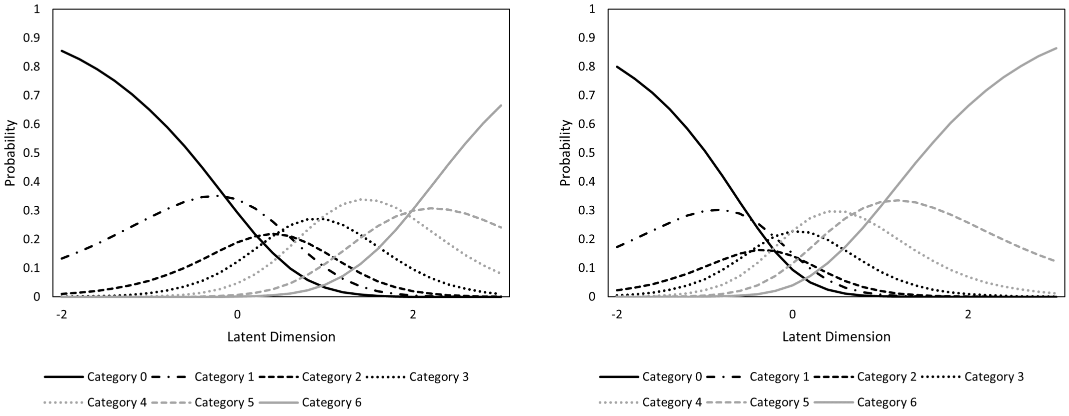


Figure 3.2. Person-Item Map of Partial Credit Model

Each item's Threshold and Location parameters captures the institution's performance as gauged by survey respondents. As discussed in Section 3.3.1.1, higher Threshold and Location parameters indicate greater difficulty in achieving scores at certain levels, i.e., they reflect worse performance relative to other items. The Location parameters in Figure 3.2 reflect higher relative performance levels on questions related to providing a good (overall and departmental) academic experience, to creating an impression of competence among alumni, and to crediting (professional) success to the university. In terms of the items where respondents were less satisfied with the university's performance, we note their perceived level of connection to the university, similarity with the university, and how often they use the word "us" to describe their relationship with the university. Here in Figure 3.2 some items show Threshold parameters in wrong orders, for example, order of “1-3-2-4-5-6” for Connected and order of “1-3-4-2-5-6”. This is due to probability of certain category in an item being too small. Figure 3.3 shows the Item Characteristic Curves (ICC) plots of the two items mentioned above, where x-axis is the latent dimension of Person parameter and y-axis is probability of each category with corresponding value of Person parameter. The intersection points between adjacent curves are the Threshold parameters presented above. In Figure 3.3a, the disorder in item “Connected” is because probability of category 2 is too low, that the category with highest probability transition to category 3 from category 1, which leads to the reversed order of Threshold parameters. The relative positions of curves of categories, as can be observed, are in the ascending order. In this case, the correct order of category with highest probability as Person parameter increases should be “0-1-3-4-5-6”, with category 2 skipped. Same for item “UnivOweSuccess” in Figure 3.3b, probability of category 2 is too low, which causes same issue of disorder of Threshold parameters, and thus category 2 should be skipped in the actual order. In such cases, Threshold parameter of category 2 of both items are no longer meaningful, since respondents generally don’t give an answer of category 2. Masters (1982) also discuss such cases in detail with an example, arguing that this is usually caused by the fact that such category with low probability has a difficulty being too close to the previous category, and thus respondent with higher Person parameter would skip it and jump to the next category.

Statistics describing the Partial Credit Model’s score residuals, y_{ik}, given in (3.7), are presented in Table 3.4. Recall that the score residuals no longer fall in the 0–6 range. These correspond to the the bias-corrected survey response data. Table 3.4 also presents



(a) Item "Connected"

(b) Item "UnivOweSuccess"

Figure 3.3. ICC Plots of Example Items

the distribution of Person parameters, which capture individual rating scales.[4] Comparison of the statistics presented in Table 3.1 to those in Table 3.4 highlights the role of individual rating scales and of item/question characteristics in obfuscating the effects of university performance on satisfaction. For example, the variable UnivOweSuccess has a median score of 4 and a median bias-corrected score of 0.16. Using these statistics as benchmarks, the original data suggest that the university's performance resulted in greater satisfaction along 3 other dimensions (with higher median scores): Competent, StuExpAcademic, and StuExpAcademicDept. However, once the scores are adjusted to control for biases, we conclude that UnivOweSuccess is the variable where university performance had the largest effect on satisfaction. The higher median scores of the other 3 variables

[4] Construction of the bias-corrected data removes the effect of individual rating scales. Individual rating scales are included as explanatory variables in subsequent analysis because they are likely to have significant effects on donation behavior. The Partial Credit Model decomposes the effects of bias-corrected variables and of individual rating scales on the original response data. into We include Person parameters because, the bias-corrected variable eliminate the information of individual rating scale, which might have significant effect on individual behavior and is preserved in the Person parameter variable.

Table 3.4. Summary of Bias-Corrected Variables

Variable	Min	Median	Mean	Max	S.D.
PersonPar	-2.47	0.70	0.70	3.81	0.70
UnivOweSuccess	-5.41	0.16	0.02	4.50	1.24
Competent	-4.89	0.13	0.01	3.47	1.09
StuExpAcademicDept	-5.06	0.11	0.02	4.81	1.00
DonationImpactful	-4.79	0.08	-0.01	5.26	1.28
StuExpAcademic	-3.93	0.07	0.01	3.40	0.89
FeelingsAA	-3.70	0.02	0.00	3.52	1.04
Connected	-4.37	0.02	-0.01	4.34	1.22
StuExpLife	-4.68	-0.05	-0.01	4.40	1.44
Similarity	-4.24	-0.11	-0.02	4.45	1.19
UseWordUs	-4.15	-0.17	-0.03	4.27	1.33

are explained not only by the university's performance, but also by factors outside of the university's control: individual rating scales and question/item characteristics.

3.4.3. Bias correction performance

To showcase the performance of the proposed approach as a bias-correction method, we consider 4 types of prediction models each estimated with the 2 data sets (original and biased-corrected) as explanatory variables. The models and response variables as follows: (1) a Logistic Regression Model with a binary response variable representing each survey respondent as either a donor or non-donor; (2) a Poisson Regression Model where the response variable corresponds to the number/frequency of donations over the 17-year analysis period; (3) a Regression Model where the response variable corresponds to the log of average annual donation amounts; and (4) a Logistic Regression Model where the response variable corresponds to each individual's segment assignment based on the behavior displayed in their donation sequence. Each of the 4 models was estimated both with the original and bias-corrected data sets. Below, we refer to the two models sharing same response variable as a model type. Based on data processing result from Section 3.4.1, type 1 models have population size of 1,934, corresponding to all survey respondents. Types 2, 3 and 4 models have population sizes of 849, corresponding to the donors in the dataset.

The results are presented in the remainder of the section. For each model type, we randomly split the population in to calibration and validation sets in proportions of 80%/20%. We train models on corresponding training set and calculate validation performance metrics on validation set, which represents predicting power of model. For the purpose of model selection, we use forward step-wise regression method based on AIC to determine explanatory variables to include in the models. Performance metrics used for each model type are introduced below. In the models, we included quadratic terms of variables to capture non-linearity in the marginal effects. Model training and validation results are shown in Tables 3.5 to 3.8, where the significance of estimated regression coefficients are represented by: "***" Significance at $p < 0.001$; "**" Significance at $0.001 \leq p < 0.01$; "*" Significance at $0.01 \leq p < 0.05$; "·" Significance at $0.05 \leq p < 0.10$.

3.4.3.1. Classification Model: Donor v. Non-Donor. We estimate a Logistic Regression model for classification of Donor v. Non-Donor response variable. Since the Donor and Non-Donor labels are relatively balanced (44%/56%), we use validation accuracy to assess the models' predictive capabilities. The results are shown in Table 3.5.

From Table 3.5 we see that the model relying on corrected data shows both higher training and validation accuracies. The corrected model is more accurate even though it includes 2 fewer variables (including the quadratic terms) than the original model.

Table 3.5. Result of Logistic Regression on Donor v.s. Non-Donor

Variable	Original		Corrected	
	Coefficient	Std. Error	Coefficient	Std. Error
(Intercept)	-3.737***	0.335	-1.479***	0.152
PersonPar	–	–	3.137***	0.324
PersonPar^2	–	–	-0.973***	0.164
Connected	0.626***	0.140	–	–
Connected^2	-0.094***	0.024	-0.196***	0.044
FeelingsAA	0.095·	0.055	–	–
Similarity	0.469**	0.144	–	–
Similarity^2	-0.053*	0.027	-0.096*	0.044
UseWordUs	-0.233***	0.046	-0.307***	0.052
DonationImpactful	0.283***	0.048	0.263***	0.052
StuExpAcademic	0.147*	0.062	–	–
UnivOweSuccess	–	–	0.278***	0.059
UnivOweSuccess^2	0.038***	0.006	–	–
Training Accuracy	69.8%		70.9%	
Validation Accuracy	68.2%		70.3%	

The variables' marginal effects are presented in Table 3.5. We observe FeelingsAA and StuExpAcademic are not significant in the corrected model. UnivOweSuccess no longer displays a quadratic effect in the bias-corrected model. Connected and Similarity show a light horizontal shift across the bias-corrected and uncorrected models.[5]

3.4.3.2. Poisson Regression Model of Donation Frequency. We develop a Poisson Regression model for regression of donation frequency response, since Poisson Regression is widely used for response variable of count data. We use validation log Mean Square Error (MSE) as the performance metric. Log MSE stands for the MSE of log of response variable. Specifically, denote the donation frequency response variable of individual i as y_i^f and prediction as $\hat{y}_i^f$, the log MSE is calculated by $\sum_{i=1}^{N}(\log(\hat{y}_i^f) - \log(y_i^f))^2$. The log

[5] In Figure 3.4, the y-axis is the linear predictor of generalized linear model, i.e., the individual i's linear response $\mu_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}$, where p is number of predictive variables.

Table 3.6. Result of Poisson Regression on Donation Frequency

Variable	Original		Corrected	
	Coefficient	Std. Error	Coefficient	Std. Error
(Intercept)	1.403***	0.241	0.973***	0.079
PersonPar	–	–	0.968***	0.125
PersonPar^2	–	–	-0.264***	0.056
Connected	0.095***	0.016	–	–
FeelingsAA	-0.227**	0.073	-0.098***	0.022
FeelingsAA^2	0.027**	0.010	0.007	0.014
Competent	-0.211*	0.093	–	–
Competent^2	0.022*	0.011	0.050*	0.019
Similarity	0.105*	0.048	–	–
Similarity^2	-0.012	0.008	–	–
UseWordUs	-0.045	0.036	-0.179***	0.016
UseWordUs^2	-0.022**	0.007	–	–
DonationImpactful	0.433***	0.053	0.182***	0.018
DonationImpactful^2	-0.039***	0.008	0.007	0.011
StuExpAcademic	–	–	-0.202***	0.031
StuExpAcademic^2	–	–	-0.020	0.029
StuExpAcademicDept	-0.074**	0.023	–	–
UnivOweSuccess^2	0.020***	0.002	–	–
Training Log MSE	0.624		0.581	
Validation Log MSE	0.869		0.752	

is taken because the Poisson distribution is skewed, and taking log will make the MSE less biased by large donation frequency value. Result of Poisson Regression model on donation frequency response variable is shown in Table 3.6.

From Table 3.6 we see that the corrected model shows both lower training log MSE and validation log MSE, even with 3 fewer variables in the model (including the quadratic terms) than original model. For marginal effects of variables, Connected, Similarity, StuExpAcademicDept and UnivOweSuccess all become insignificant in the corrected model. StuExpAcademic becomes significant in the corrected model. UseWordUs becomes linear, while Competent has a slight shift after the bias correction. DonationImpactful, on the other hand, has a reversed sign for quadratic term and thus show completely opposite marginal effects on donation frequency.

3.4.3.3. Regression Model on Average Donation Amount. We develop a Normal Regression model for regression of average donation amount response. Since average donation amount variable is highly skewed, we take log of average donation amount and use log value as response variable in Normal Regression model. As shown in Table 3.2, the log average donation amount is much less skewed. We use validation MSE of log average donation amount. Namely, denote the log average donation amount variable of individual i as $y_i^{a,l} = \log y_i^a$ and prediction as $\hat{y}_i^{a,l}$, the MSE is calculated by $\sum_{i=1}^{N}(\hat{y}_i^{a,l} - y_i^{a,l})^2$. Result of Normal Regression model on Log Average Donation Amount response variable is shown in Table 3.7.

From Table 3.7 we see that the corrected model shows both lower training MSE and validation MSE, with one less variables in the model than original model. For marginal effects of variables, Competent and StuExpAcademic become insignificant, while UseWordUs becomes significant after bias-correction. FeelingsAA becomes non-linear, while Similarity and StuExpLife have horizontal shift in the bias-corrected model.

3.4.3.4. Classification Model on Behavior Assignment. We develop a Logistic Regression model for Behavior Assignment response variable. Based on the 3 segments of Behavior Assignment, we aggregate the LV and HV segment here to form a binary Behavior Assignment variable of LV/HV donor v.s. TS donor. Since LV and HV type are desirable to target compared with TS type, this better reflects the alumni population to target. The aggregated proportions are 40.0% of LV/HV (labeled as 1) and 60.0% of

Table 3.7. Result of Normal Regression on Log Average Donation Amount

Variable	Original		Corrected	
	Coefficient	Std. Error	Coefficient	Std. Error
(Intercept)	4.802***	0.750	4.107***	0.108
PersonPar	–	–	0.439*	0.188
PersonPar^2	–	–	-0.117	0.094
Connected	–	–	0.137***	0.033
Connected^2	0.024***	0.006	–	–
FeelingsAA	-0.109**	0.039	-0.041	0.043
FeelingsAA^2	–	–	0.042	0.034
Competent	-0.009	0.041	–	–
Similarity	0.140	0.097	–	–
Similarity^2	-0.031·	0.017	-0.023	0.016
UseWordUs	–	–	-0.036	0.030
DonationImpactful	0.400***	0.096	0.188***	0.031
DonationImpactful^2	-0.039*	0.015	-0.029	0.018
StuExpAcademic	-0.655·	0.354	–	–
StuExpAcademic^2	0.074·	0.039	–	–
StuExpAcademicDept	0.335	0.207	0.019	0.050
StuExpAcademicDept^2	-0.032	0.025	-0.024	0.021
StuExpLife	-0.353**	0.108	–	–
StuExpLife^2	0.043**	0.015	0.046**	0.017
Training MSE	0.995		0.671	
Validation MSE	1.285		1.189	

TS (labeled as 0). Although the binary Behavior Assignment proportions are not highly imbalanced, the model might favor TS donors more than LV/HV donors in accuracy. Therefore, we use the validation Area Under Curve (AUC) of ROC curve[6] and AUC of

[6] For a binary classification model, given the prediction probability $p(\hat{y}_i = 1|\mathbf{x}, \boldsymbol{\beta})$ for all individuals, a hard classification into category can be obtained by setting a specific probability threshold, i.e. $\hat{y}_i = I(p(\hat{y}_i = 1|\mathbf{x}, \boldsymbol{\beta}) > \hat{p})$. The ROC curve is the 2-d plot of True Positive Rate and False Positive Rate with different probability threshold $\hat{p}$, and the Area Under Curve (AUC) can be calculated for the ROC curve. AUC has a range of 0–1, where larger AUC indicates a better classification performance. Usually AUC of ROC curve is more robust than accuracy in imbalanced cases.

Table 3.8. Result of Logistic Regression on Binary Behavior Assignment

Variable	Original		Corrected	
	Coefficient	Std. Error	Coefficient	Std. Error
(Intercept)	0.857	0.841	-0.936***	0.277
PersonPar	–	–	0.972***	0.185
Connected	0.435***	0.083	0.165·	0.088
FeelingsAA	-0.589·	0.311	–	–
FeelingsAA^2	0.061	0.043	–	–
Competent	-0.783*	0.345	–	–
Competent^2	0.088*	0.042	0.098	0.081
Similarity	0.489*	0.198	-0.010	0.081
Similarity^2	-0.101**	0.036	–	–
UseWordUs	-0.269***	0.066	-0.536***	0.076
DonationImpactful	0.455***	0.072	0.409***	0.082
DonationImpactful^2	–	–	-0.016	0.045
StuExpAcademic	–	–	-0.339*	0.147
StuExpAcademic^2	–	–	-0.132	0.140
StuExpAcademicDept	-0.257**	0.097	–	–
StuExpAcademicDept^2	–	–	-0.124	0.093
StuExpLife	-0.191**	0.058	-0.178*	0.071
UnivOweSuccess^2	0.059***	0.009	-0.129·	0.070
Training Accuracy	69.8%		71.2%	
Validation Accuracy	65.6%		75.5%	
Validation AUC ROC	0.683		0.793	
Validation AUC PR	0.542		0.650	

PR curve[7]. Accuracy is also calculated as reference. The result of Logistic Regression on binary Behavior Assignment response variable is shown in Table 3.8.

From Table 3.8 we see that validation performance improvement of bias-corrected model is even more significant than the previous 3 model types in all performance metrics, while the corrected model includes same number of variables as original model. Also,

[7] The difference between ROC curve and PR curve is that, while ROC curve is the plot of True Positive Rate and False Positive Rate, PR curve plots the True Positive Rate versus Precision. AUC of PR curve is calculated in the same way as ROC curve. Also, PR curve focus more on the performance of classification on the positive category, i.e., the LV/HV category.

we can argue that the corrected model has better prediction power in both the LV/HV and TS donors, since accuracy is inclined to TS donors because of its outnumbering while PR curve focuses on the performance on identifying LV/HV donors. For marginal effects of variables, FeelingsAA becomes insignificant while StuExpAcademic becomes significant in the corrected model. Effect of Similarity becomes linear, while effects of Donation-Imapctful and StuExpAcademicDept become non-linear. Competent has horizontal shift in the bias-corrected model. The biggest change comes from UnivOweSucess, where the sign of coefficient changes after the bias correction.

3.4.3.5. Model comparison summary. In this section, we discuss the comparison between original and corrected model in terms of predicting performance and marginal effects of variables.

Table 3.9 summarizes the improvements of corrected model on the validation performance metrics. For classification models, the improvement is calculated by the difference between original and corrected models. For regression models, the improvement is calculated by the ratio of difference to original model metric value[8]. We can see that the bias correction makes largest improvement in the Poisson Regression model on donation frequency and Logistic Regression model on Behavior Assignment. While the Logistic Regression model on Donor/Non-Donor has the smallest improvement, it still shows an improvement larger than 2%.

Also, we see multiple changes in marginal effects of survey variables after bias correction. Therefore, interpretations of marginal effects of survey variables are also impacted by rating scale bias, which leads to important change in university's policy to target

[8] Specifically, for Poisson Regression model $13.4\% = (0.869 - 0.752)/0.869$, and for Normal Regression model $7.5\% = (1.285 - 1.189)/1.285$.

Table 3.9. Validation Result Improvement by Correcting Rating Scale Bias

Model	Metric	Improvement
Logistic Regression on Donor/Non-Donor	Accuracy	2.1%
Poisson Regression on Donation Frequency	Log MSE	13.4%
Normal Regression on Donation Amount	MSE	7.5%
Logistic Regression on Behavior Assignment	Accuracy AUC ROC AUC PR	9.8% 11.0% 10.8%

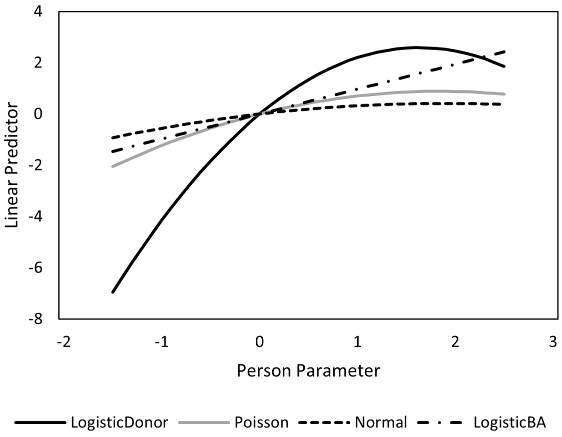


Figure 3.4. Marginal Effects of Person Parameters in Different Models

alumni-giving. For Person parameter, it has a clear trend of linear predictor falling down slightly after Person parameter going above a certain threshold in the first 3 types of models. This indicates that, although we observe an overall trend that higher rating scale leads more alumni donation, some alumni with low or no donation also show high rating scale in the survey. This might be due to the fact that these survey results were not taken seriously, where random high scores were answered to quickly finish the survey.

3.5. Conclusion

In this study, we present a method to control individual rating scale existing in subjective survey response data, and correct bias caused by heterogeneous rating scale based on Rasch model. The method is appealing because, as opposed to previous literature in controlling rating scale and making inference on population or group-level true belief, it provides bias-free survey data that preserves individual true beliefs on different survey questions, and is available for further analysis and modeling. Compared with literature that aims to provide similar bias-corrected survey data, this method does not rely on any additional objective information as “anchoring” variables, and thus is desired in most cases where additional information is not available. As a general approaching dealing with survey response data, this method can be applied in marketing research, and potentially even broader fields where subjective survey data is applied.

We applied the bias correction method in a non-profit fundraising setting. The data comes from 2 sources: an alumni survey conducted in the Fall of 2017, which was designed to elicit information about donor experience as a student and university alumnus, their feelings towards and perception of the relationship with the university; and alumni donation record for the 2000–2016 period. A Partial Credit model – one type of generalized Rasch model – is estimated on the alumni survey data, and rating scale bias is eliminated to obtain bias-corrected survey data. To validate effectiveness of bias correction, we develop predictive models with different response variables representing different alumni donation behavior, and explanatory variables of original and bias-corrected survey data respectively. Results show that, the bias-corrected survey data renders better predicting performance in estimating individual alumnus donation behavior, and different marginal effects of variables compared with original survey data. This demonstrates the higher predicting power of data generated by bias correction, and the importance to correct rating scale bias in order to obtain unbiased insights from analysis, which leads to better targeting strategies in marketing.

Limitations exist in the proposed method. First, goodness-of-fit of questions in Rasch model is required, and thus bias-corrected data of both underfit and overfit questions cannot be obtained, due to violation of assumptions by Rasch model. This narrows the applicability of the method on some proportion of data in specific cases. Besides, additional bias may arise in estimated rating scale from respondents that arbitrarily answer high or low response scores, which leads to inaccurate estimation of bias-corrected data of these individuals. Thus, the method is not yet robust to and cannot distinguish unreliable data source.

CHAPTER 4

A joint segmentation analysis of commute mode choice and VMT in the Chicago Metropolitan Area

4.1. Introduction

We use travel tracker survey data from the Chicago Metropolitan Agency for Planning (CMAP) to analyze travel outcomes of households with different characteristics, including their spatial distribution.

In travel behavior analysis, commute trips garner the lion's share of the attention for mode choice modeling, since traffic during weekday's peak hours causes the most severe congestion issues (Bhat, 1997b). In addition to mode choice, recurring trips are more likely to influence residential location, vehicle ownership, and other decisions tied to travel outcomes. According to Elldér (2014), while weekday commute and service trips are dependent on residential location, leisure trips on weekends have great variation within a neighborhood. Vehicle miles traveled (VMT) is often used used as a measure of travel behavior. This measure is also widely used in a variety of transportation and planning functions, e.g., estimation of vehicle emissions and energy consumption, public transit planning, etc. VMT often refers to the total distance traveled by Privately Owned Vehicle (POV) or by Taxi. Trip distances covered by other modes are often excluded (Chatman, 2003), or added to VMT by POV and Taxi together (Hong et al., 2014). Other literature (Lin and Long, 2008; Paulssen et al., 2014) tried to introduce mode share to reflect people's mode choice, but VMT and mode share have not been both used together to represent commute travel behavior.

One of the important issues in transportation data analysis, as explained by Washington et al. (2010), is to account for heterogeneity. Among the population, both individual-level and group-level heterogeneity widely exist. While individual-level heterogeneity represents the difference between each individual in the population, group-level heterogeneity represents the different characteristics between groups of population, in which we assume homogeneity exists. In travel behavior analysis, capturing the group-level also contributes to policy development and resource allocation, since various groups are identified and travel behaviors within a group follow a certain pattern. To account for group-level heterogeneity, market segmentation techniques are frequently used. Market segmentation techniques partition heterogeneous population into smaller homogeneous groups – denoted by "clusters" in market segmentation, in order to identify groups where potential similarities in behavior exist (Wedel and Kamakura, 2012). The formation of these groups helps policies to be better designed to impact certain cluster members, as opposed to impacting the whole population less effectively (Teichert et al., 2008).

Following McFadden and Reid (1975) and Domencich and McFadden (1975), two main types of analysis – aggregate and disaggregate analysis – are adopted for travel behavior. In aggregate analysis, households are often grouped based on their residence location, and linked with travel behavior in a certain zone. Boarnet and Crane (2001) performed aggregate travel behavior analysis by regressing trip behavior variables on zonal land use measure, and provided important strategy for urban design and transportation planning.

On the other hand, disaggregate analysis refers to analyzing the relationship between individual's demographic information and travel behavior variables. Critically, McFadden (2000) note that "zones don't travel; people travel!". That is, in travel behavior analysis, we should focus on individual's characteristics rather than zonal average statistics, because we can't guarantee that households or individuals living in a certain zone share similar characteristics. Naess (2011) study how urban form factors affect travel behavior falls into the category of disaggregate analysis. The main difference originates from the assumption on diversity of data. Aggregate analysis assumes that households in the same aggregation behave more consistently, and aggregate measure is assumed to be representative. Disaggregate analysis assumes heterogeneity generally exists among population, where each sample represents only its own information.

In this paper, based on the household level travel tracker data, we employ a segmentation framework to study impacts of factors on travel behavior, providing an analysis technique complementary to conventional techniques by modeling multidimensional vehicle miles traveled (VMT) vector. Specifically, we formulate and estimate a finite mixture model that describes the joint distribution of VMT for POV/Taxi and for other modes[1], which includes information on both travel distance and mode split. The combination of 2 important travel outcomes – trip distance and mode split – supports identification of key factors influencing household travel behavior, including socioeconomic and demographic factors, built environment factors, and importantly, latent factors related to individual preference. We consider household size and income, and number of workers, students, vehicles and licenses in a household for socioeconomic and demographic (SED) factors. For

[1] According to travel tracker data, most of other mode trips belong to public transit.

built environment (BE) factors, distance to CDB, land use mix, population and household density are considered.

Following the segmentation result of identification of different clusters, we conduct both aggregate analysis and disaggregate analysis. We use the correlation between the segment membership probabilities and a household's SED and BE factors to establish the significance of factors contributing to travel outcome classes. A spatial analysis that relates segment membership and its geographical distribution is presented to show the complicated interdependencies among travel outcomes, residential location, and public transport coverage. Our analysis, suggests that a household's SED factors play a more significant role in travel outcomes than BE environment factors.

The remainder of the paper is organized as follows: In Section 4.2, we position and contrast our work with respect to the literature. The data used in the study is described in Section 4.3. In Section 4.4, we describe the details of the framework used for segmentation. The results of our analysis appear in Section 4.5. Discussions, limitations and conclusions are discussed in Section 4.6 and Section 4.7.

4.2. Literature review

In this section, we position our work with respect to the travel behavior and market segmentation literatures. As discussed in Domencich and McFadden (1975), several points were raised about travel behavior analysis including diversity of modes and effects of variables on travel decision. Specifically, the authors argued that it is "necessary to give careful attention to the specification" of mode combination, and important factors should be accounted for in the whole process. Furthermore, following Washington et al. (2010), heterogeneity among the population needs to be captured in the model. Therefore, based on guidance of the literatures, we contrast our work to segmentation studies on travel behavior in 4 aspects: (1) representation of travel outcomes; (2) effects of factors on travel outcomes; (3) heterogeneity in travel outcomes; (4) segmentation in analysis of travel outcomes. Then we list the contributions of this paper specifically in those aspects.

4.2.1. Representation of travel outcomes

VMT is often used as representation of travel outcomes in macro-level transportation analysis in order to find its association with BE and SED factors. In most cases, VMT is either calculated by summing up all POV and Taxi trip distances and excluding trips by other modes, or by aggregating all trips by any mode. For example, Chatman (2003) excluded individuals who report bus service in order to generate a truncated data set from original data, and Hong et al. (2014) estimated a Bayesian hierarchical model to examine the effect of BE factors on individual VMT, where total VMT of all trips for each individual was calculated and used.

The problem with using single VMT value for travel behavior modeling is that people's travel mode choice and travel pattern are not reflected. In addition, the exclusion of other modes simply ignore people's travel behavior of transit. To cope with these problems, mode share has also come into play. Paulssen et al. (2014) adopted categorical response variable with 3 levels of “Drive only”, “Drive+transit”, and “Transit only” to developed a hierarchical mixed logit model on travel mode choice and latent variables, without travel distance being considered. Lin and Long (2008) conducted a log-likelihood clustering to identify types of neighborhoods. Four travel measures including total VMT and mode share were considered. However, total VMT and mode share were used in two parts of analysis separately instead of together.

In this study, as opposed to single VMT value and mode share, we include both VMT for POV/Taxi and other modes to fully reflect the travel behavior pattern. Specifically, we use a multi-dimensional vector to represent VMT for both modes instead of a scalar, and assume a multi-variate distribution for the multi-dimensional response variable.

4.2.2. Effects of factors on travel outcomes

The relationships between BE and travel behavior have been widely discussed in the literature (Handy, 1996; Boarnet and Crane, 2001; Ewing and Cervero, 2010). Many studies have shown the significant effect of built environment factors. Naess (2011) made a comparison between metropolitan-level and neighborhood-level urban form effects on travel behaviors, based on a case study in the Copenhagen Metropolitan Area. The former was found to have stronger effects on travel behavior. Hong et al. (2014) presented several issues on conventional statistical analysis on travel behavior and concluded that built environment have significant effects on VMT, though travel attitude and spatial relationship do as well. However, some literature cast doubts on the direct effect of built environment. Boarnet and Sarmiento (1998) built an non-work trips generation model and found that there is no significant link between land use policy and non-work trips. Handy et al. (2005) questioned whether built environments have a causal relationship with travel behavior. A multi-variate analysis on cross-sectional data suggested that SED and attitudes factors, instead of built environment, account for most effects on travel behavior.

In this paper, we consider both SED factors of household, and built environment factors as well. The result of correlation analysis is used to examine the effects of SED and built environment factors to see which have the more significant impact on household travel behavior in the Chicago Metropolitan Area case.

4.2.3. Heterogeneity in travel behavior

Heterogeneity in travel behavior can be summarized by observed and unobserved factors. Generally, the effect of observed factors on response variable of interest can be captured by including corresponding explanatory variables. From a model standpoint, most literature tend to use regression models to link travel behavior measures and explanatory variables. In order to incorporate heterogeneity in regression process, the intercept variable is used. It assumes that the heterogeneity effects are constant and are caused by omitted variables collectively (Washington et al., 2010). There are 2 ways to incorporate this intercept variable – fixed and random effect – which refer to group effect of heterogeneity that is correlated or uncorrelated with explanatory variables. Bhat and Guo (2004) came up with a mixed spatially correlated logit model with a correlated spatial effect term in choice modeling analysis. For the same purpose, Hong et al. (2014) included a varying intercept uncorrelated with explanatory variables and a spatial effect term correlated, in order to capture the heterogeneity caused by spatial auto-correlation. However, two limitations exist in these models. First, it does not reflect the fact that people sharing similar characteristics usually form homogeneous groups that represent different types of travel behavior. The between-group heterogeneity might be even more important and representative. Second, the regression relationships between travel behavior (i.e., VMT) and the explanatory variables only explain how people's characteristics affect the distance of travel, but not the travel pattern and mode choice. For example, a regression model that shows positively significant relationship between VMT and demographic factors provides very limited information, and also has limited capability of capturing complex impacts of factors on travel behavior.

The justification for considering heterogeneity has come in several forms in the literature. In general, the inclusion of heterogeneity in the modeling framework serves to reduce bias in estimation/prediction (Bhat, 1998; Bhat and Guo, 2007). This is especially the case for mode or residential choice models, where some means of accounting for heterogeneity has become common practice (Bhat and Guo, 2004; Koppelman and Sethi, 2005; Greene et al., 2006). Examples that are more relevant to this study are those where heterogeneity is specifically considered in the context of how household demographic and built environment variables interplay to influence travel behavior. For example, Bhat and Guo (2007) examined how various built environment factors influence residential choice and levels of auto ownership. To capture the effect of heterogeneity on sensitivity to built environment attributes, parameters representing heterogeneity caused by both of observed and unobserved effect are directly built into utility functions. Pinjari et al. (2008) applied a similar approach to analyzing the relationship between residential neighborhood type choice and bicycle ownership, while considering the impact of self-selection.

In this paper, we conduct a segmentation analysis on households to segment them into homogeneous clusters in order to capture the group-level heterogeneity that exists in the population. Also, instead of directly linking travel behavior to household characteristics, we generate intermediate latent variables – cluster membership probability – to describe travel behavior pattern, and then link cluster membership probability with household characteristic factors to investigate their impact on people's travel behavior.

4.2.4. Segmentation in analysis of travel behavior

Segmentation involves grouping observations in a heterogeneous population, with the objective of maximizing within-cluster homogeneity and between-cluster heterogeneity, i.e., dividing subjects into different clusters with within-group similarity and inter-group heterogeneity. In this section, we discuss two types of categorizations of segmentation methods: categorized by procedure, *a priori* or *post-hoc* methods; and by variables basis, explanatory variables or response variables.

According to Green (1977), segmentation methods are classified as either *a priori* or *post-hoc* method. For *a priori* methods, the type and number of clusters is determined prior to data collection and analysis. An example is classifying households into clusters based on their geographic location, for example, traffic analysis zone (TAZ) or zip code. Lindsey et al. (2011) performed segmentation of VMT and emission level into predetermined cells based on their spatical location to investigate the effect of residential location on emission. Kressner and Garrow (2012) adopted a segmentation algorithm that divides households into 26 clusters prior to data collection based on their socioeconomic and demographic information. However, *a priori* segmentation requires prior assumption that heterogeneity is associated with the variables, which is not always the case. On the other hand, *post-hoc* segmentation forms homogeneous groups according to a set of measured characteristics from data, and thus can better capture unobserved goup-level heterogeneity that cannot be priorly determined. Bhat (1997a) developed an endogenous segmentation method, which was modeled by cluster-wise multinomial-logit model. The paper showed that this method provides a better fit, and also "intuitively more reasonable results". Lin and Long (2008) performed a K-means clustering on Census Transportation planning data to group people into different neighborhood with similar SED and land use characteristics.

By using different methods, market segmentation is built upon a certain basis of either explanatory or response variables. Most of the literatures applied segmentation with basis of explanatory variables, which usually include SED variables. Pinjari et al. (2008) framed the notion of "sustainable lifestyle" and performed segmentation with basis of corresponding explanatory variables in order to study people's behavior change on tourism. Lin and Long (2008), discussed above, performed a K-means clustering on SED and land use characteristic variables. On the other hand, response variables are rarely used in segmentation in transportation studies. An instance is that Ma and Goulias (1996) clustered individual's travel pattern and activity pattern into homogeneous behavioral groups, and analyzed the relationship between travel pattern and individual characteristics. However, instead of assuming that heterogeneity between clusters is based on explanatory variables, segmentation on response variables focuses on the nature of the actual outcomes of observations, and thus might better capture the unobserved heterogeneity in people's travel behavior.

In this paper, we apply a *post-hoc* segmentation based on response variables. Specifically, a finite mixture model is developed on the vector of VMT for POV/Taxi and VMT for Other Mode. Compared with the literature, our model can better capture the natural characteristics of people's travel behavior, and thus segment them into clusters with homogeneity more accurately.

4.2.5. Contributions

Based on the literature review, we summarize the contributions of this paper that distinguish from the literature. Specifically, the contributions of this paper are:

- Instead of using single VMT value as in most cases (Chatman, 2003; Hong et al., 2014; Lin and Long, 2008), we use VMT for both POV/Taxi and Other Mode to specify POV/Taxi and transit travel distance respectively. Namely, not only the travel distance made by a household as in the cases of most previous studies, we include the specific travel pattern of different modes, which demonstrates the characteristics of household. Considering both VMT for POV/Taxi and transit will help capture the different choices between POV/Taxi and transit of travel behaviors, and also avoid biased conclusion as discussed in Section 4.2.1. This is validated in Sections 4.5 where we compare the results of both excluding and including the VMT for transit. Furthermore, the considered mode split travel pattern helps us better identify people with different characteristics, and to further investigate the relationships between travel behavior, SED factors, and built environment factors.

- Considering the possibility that households living in a certain zone might not share the same characteristics of traveling, we focus on the characteristics of individual households instead of average statistics within zones. Instead of predetermining the partitioning of people by observed variables, for example Kressner and Garrow (2012), we apply a *post-hoc* method instead of *a priori* method for segmentation, which avoids making prior assumptions about the segmentation, and captures the unobserved heterogeneity in the data itself. Also, as opposed to the segmentation with basis of explanatory variables by Pinjari et al. (2008) and Lin and Long (2008), the segmentation we perform uses basis of response variables – the VMT for POV/Taxi and Other Mode – to model the travel pattern of households, in order to generate smaller clusters, within which households share similar travel behavior with cluster members. Specifically, we applied a multivariate finite mixture model, which is the first in travel behavior analysis to the best of the authors' knowledge. In this way, we put explanatory variables aside from this procedure and focus on the travel behavior of households. Then we link travel pattern to SED and built environment factor (i.e., explanatory variables), which is independent of the travel pattern segmentation .
- Instead of using regression model to capture the relationship between VMT and household characteristic factors like Bhat (1997a), we use intermediate latent variables of cluster membership probability and set up the relationship between the latent variables and household characteristic factors by correlation analysis. Besides, we profile the clusters obtained by segmentation analysis to demonstrate the characteristics of travel behavior of different clusters. Namely, we set up the

relationships between household characteristics and travel behaviors within clusters, which provides deeper insights of how household characteristics are related to people's travel behavior in different clusters.

- In addition to analyzing the factors that are related to the different clusters with various traits, we also include analysis on spatial distribution pattern of clusters and compare it with urban-form data such as accessibility of transit, which indicates the interaction between land use, public transport and people's travel behavior. This provides a deeper insight of mutual effect between travel behavior and urban form that, although urban form may not directly have a causal effect on travel behavior of people, the interaction between these two factors can still be shown.

4.3. Data

There are 2 types of data used in this study. The first type of data, which comes from travel tracker survey data, provides detailed household travel information and their SED information. The second type of data, which is collected from 2 sources, describes the urban form measure of entire Chicago area, including built environment factors of household and distribution of transit stations. The household travel information is used in the VMT segmentation. As discussed previously, following the VMT segmentation, 2 types of analysis – aggregate analysis and disaggregate analysis – will be carried out in this paper. Household SED and built environment data are used in aggregate analysis, while the distribution of trasit stations is used in disaggregate analysis.

Beside the travel tracker survey data from CMAP, in order to validate the advantages and effectiveness of our model, we also adopt similar data of other two cities – New York and Seattle, which also have developed transit systems as Chicago, and a national travel survey data. 2010/2011 Regional Household Travel Survey (RHTS) data (NYMTC, 2011) from New York Metropolitan Transportation Council (NYMTC) is used for New York metropolitan area. For Seattle, Spring 2017 Household Travel Survey data (PSRC, 2018) from Puget Sound Regional Council (PSRC), as a part of PSRC's Household Travel Survey Program, is used. The 2017 National Household Travel Survey (NHTS) data (U.S. Department of Transportation, Federal Highway Administration, 2018) conducted by Federal Highway Administration (FHWA) is used as a nationwide travel survey source. Details about the 3 datasets are introduced in F.

4.3.1. Travel tracker survey[2]

This study describes household travel behavior using data from a household travel survey conducted in 2007 and 2008, which is the latest comprehensive travel activity survey for Northeastern Illinois, performed by Chicago Metropolitan Agency for Planning (CMAP, 2008). The survey includes trip data and demographic information for each member in 10,552 households. Household members were asked to complete travel journal for 1-day or 2-day periods. About 60% of the households were asked to report trips over 1 day while 40% of households provided recorded trips for two days. This survey covered the Chicago

[2] The original travel tracker survey data can be downloaded from CMAP website (the download page address is provided in the reference). If the reader is interested in the processed data used for modeling, please reach out to the authors through the contact information provided in the paper.

Metropolitan Area, including McHenry, Lake, Kane, Cook, DuPage, Kendall, Grundy, and Will counties.

Since this study focuses on commute mode choice, we only include commuting-related trips according to the Primary Trip Purpose information available in the survey data. Here, our definition of commuting includes trips going to work for workers and trips going to school for students. Workers and students are also two types of people in households that we focus on in the analysis. Specifically, trips with 5 types of Primary Trip Purposes are selected: "Work/Job", "All other activities at work", "Attending class", "All other activities at school", and "Work/Business related". Notice that these trips include both home-to-work/school trips and work/school-to-home trips[3]. Then the travel distance information, i.e. VMT, of these commute trips needs to be obtained. Here, since the coordinates of origin and destination for each trip are included in the data, we obtain the Euclidean Distance, i.e. straight-line distance, based on the longitude and latitude information of the origin and destination pairs[4]. Also, in order to make 1-day survey and 2-day survey comparable to each other, we calculate the daily average travel distances for the 2-day surveys. Notice that we include the round trips of both going to and coming back from work/school. This is because asymmetry is found in some observations in the data. Specifically, going to and coming back from work/school might not have the same distance of number of trips. For example, people might go to groceries store or pass by restaurant after work, leading to larger number and longer distance of trips. In order to

[3] When non-commute trips are chained to commute trips, we only calculate the distance between workplace/school and home.

[4] Although actual travel distance varies because of complication of road network and traffic, straight distance can be taken as proxy of VMT. Also, it can be shown that there is linear relationship between straight-line distance and urban road distance (Boarnet and Chalermpong, 2001).

capture the actual travel pattern, both trips are included. A problem here is that, since non-commute trips chained to commute trips are excluded, some special travel pattern of non-commute and commute trips connecting with each other might not be captured by our model and analysis. This will be discussed in detail in Section 4.6.4.

As mentioned, we distinguish POV/Taxi VMT from VMT for other modes. This is based on the Mode of Trip information available in the survey data. We categorize 3 types of Mode of Trip – "Auto/Van/Truck driver", "Auto/Van/Truck passenger", and "Taxi" – as the POV/Taxi Mode, and others as the Other Mode, which includes CTA/Pace/Metra transit, school bus, walking, biking, etc. This means that some composite travel behaviors, such as walking/biking from home to transit stops or from transit stops to working place, are also included in the Other Mode. Since we count in both trips as driver and passenger, in some cases where more than one household members share the same POV, we might double-count the VMT because only one of the VMT actually occurred, which cannot be corrected because of the limited information available in the survey. This double-counting seems to be biased, however, it is meanwhile reasonable to some extent, since multiple household members sharing a POV is not exactly the same with only one member driving the vehicle after all. This will be discussed Section 4.6.4 as a limitation in detail. Among the original 10,552 households in the raw data, only 10 households show the case where two members in the household are POV driver and passenger respectively. Therefore, the impact of the double-counting is negligible.

Furthermore, we remove some households with missing variables. Households with missing travel distance information are excluded. Also, some households that have unknown income, which means they refused to reveal this information, are excluded as well.

Table 4.1. Summary of household SED information

	Min	Mean	Max	St. deviation
Income ($ in thousands)[5]	15	69.53	110	33.54
Household size	1	2.59	8	1.33
Number of workers	0	1.63	5	0.71
Number of students	0	0.72	6	1.03
Number of vehicles	0	1.83	8	0.97
Number of licenses	0	1.85	6	0.79

Table 4.2. Summary of household VMT data

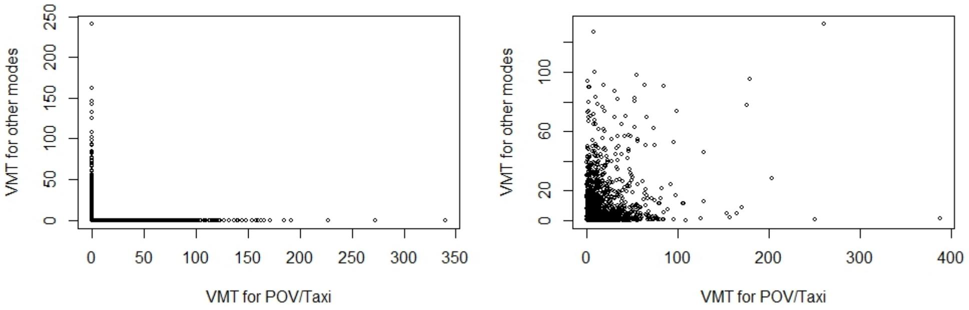
	Min	Mean	Max	Number of 0	St. deviation
VMT for POV/Taxi	0.00	16.52	387.65	1161	22.52
VMT for Other Mode	0.00	4.45	240.77	4207	12.09

We end up with a dataset of 6,886 observations, each corresponding to an individual household.

For simplicity, we only keep SED information that is relevant to our context, including household size, household income, number of workers, number of students, number of vehicles, and number of vehicle licenses in a household, along with VMT for POV/Taxi and VMT for Other Mode. In this study, we develop a segmentation model with response variable of VMT data only, instead of a functional model between VMT and explanatory variables. Therefore, here we present the summaries of VMT and SED data separately. The summary of statistics of SED data is shown in Table 4.1, and VMT data in Table 4.2. Note that in Table 4.2, the mean and standard deviation both include the 0 values.

According to statistics summarizing the household VMT data shown in Table 4.2, the VMT for POV/Taxi and Other Mode show very different distributions. In addition,

[5] The income data collected from the CMAP travel tracker survey is categorical. The 7 categories correspond to 7 income intervals. In order to conduct numeric analysis, we convert it to numeric data by replacing the categories the mean values of corresponding income interval.



(a) Household VMT with only 0 (b) Household VMT with 0 excluded

Figure 4.1. Complementary distributions of household VMT for POV/Taxi and Other Mode

Figure 4.1 shows two plots of with only observations with 0 VMT measure for either mode included and excluded respectively. Figure 4.1a shows household VMT on the two axes, i.e., with non-zero VMT for one mode and 0 VMT for the other, while Figure 4.1b shows household VMT with non-zero VMT for both modes. These two plots are mutually complement set with respect to the whole dataset (union of the two plots is the plot of the whole dataset). From Figure 4.1 we see that, there are at least 3 types of households – POV/Taxi exclusive, transit exclusive, and mixture of both modes. This suggests that different types of households may follow different distributions, which means that modeling the VMT dataset with only one single distribution is not enough to describe VMT of all the households. This motivates us to apply finite mixture models, which will be discussed in the following section.

To summarize, the travel tracker survey data in the study is in the units of households, including: VMT for POV/Taxi and VMT for Other Mode (both with 0's included), household size, household income, number of workers, number of students, number of vehicles, and number of vehicle licenses in a household.

4.3.2. Urban-form data

In order to capture the effect of urban-form variables on travel behavior, this study use some macroscopic land use measures, including public transit coverage, land use mix, distance to CBD, population density and household density. In order to match the time period in which travel tracker data is conducted, we use data from 2010, which is the closest to 2007-2008 that are available.

Those measures come from the following sources:

- Census data 2010
- Land use mix data from GEODA center
- Location of CTA train stops from City of Chicago data portal
- Location of METRA stops from City of Chicago data portal

From these data sources, for each household from the travel tracker survey part, we summarize variables: distance to CBD, land use mix measure, population density, and household density. The summary of these variables is shown in Talbe 4.3.

For distance to CBD, we choose a coordinate of (41.8781,-87.6298) (which is on Jackson station of CTA Blue Line) to represent the Chicago downtown area. The distance to that point from each household is calculated.

The land use mix measure here is an indicator of neighborhood-level diversity with a range between 0 and 1 (with no unit). Larger number means more convenient access to various jobs and services within the area. Spears et al. (2014) summarized three measures

Table 4.3. Summary of household built environment information

	Min	Mean	Max	St. deviation
Distance to CBD (miles)	0.08	18.45	70.70	15.69
Land use mix	0.00	0.72	0.98	0.14
Population density (per square mile)	7.48	3930.67	196409.20[6]	9382.67
Household density (per square mile)	0.98	2009.05	139293.90	6610.33

of land us mix of job-housing balance, land-use dissimilarity, and land-use entropy index. The last one is adopted by the data used here. Specifically, it's an entropy index of different types of land use. Equation 4.1 shows how the land use mix entropy index is calculated:

$$H = -\frac{\sum_{j=1}^{S} p_j \ln p_j}{\ln S} \tag{4.1}$$

where H is the entropy index value, S is the number of types of land use considered, and p_j is the proportion of land use type j. For a single land use, H will be 0 and for a equal distribution of land use where p_j of all types are equal, H will be 1.

Also, location data of CTA train and METRA stops is summarized to be used in the aggregate analysis. They will be shown in the results in Section 4.5.5 to be compared with segmentation result.

[6] The 196409.20 of population density and 139293.90 of household density, which seem extremely large, come from 13 households in the survey. All these 13 households are from a neighborhood near the CTA red line Argyle station. The GeoID of the neighborhood is 17031030702. Since they're concentrated in the same region, we didn't exclude them in case that bias is generated. Beside these 13 households, the max of population and hh density are 35564.24 and 29829.02 per sq mile.

4.4. Model

In this section, we introduce the basis of segmentation and present the finite mixture model for segmentation. We also present the Expectation-Maximization (EM) algorithm used to estimate associated parameters in G.

4.4.1. Segmentation basis

In this study, we focus on capturing the characteristic of travel behavior rather than directly forecasting future travel behavior by explanatory variables. Therefore, instead of modeling the functional form of relationship between response variable and explanatory variables, we developed a model that describes the distributions of VMT of different clusters in the whole population. In this way, we also avoid making some functional assumptions as made in regression analysis.

Specifically, we develop a finite mixture model for segmentation in this study. Two important contributions here are: (1) we use segmentation basis of response variable instead of explanatory variables; (2) we generate a multidimensional response variable as vector to model the joint distributions of different clusters. As discussed, both VMT for POV/Taxi and Other Mode are used as a two-dimensional vector for response variable. The response variable of VMT for POV/Taxi only as a scalar is used as benchmark for comparison. If we denote segmentation basis – the response variable vector – as $\boldsymbol{y}$, we have $\boldsymbol{y} = \{VMT_{POV/Taxi}, VMT_{Other}\}$ for our segmentation analysis and $\boldsymbol{y} = \{VMT_{POV/Taxi}\}$ as benchmark for comparison.

4.4.2. Finite mixture model

In this section, we present a Gaussian mixture model to describe the travel behavior of the hosueholds and to support segmentation based on their VMT of commute trips. To capture the information contained in the VMT vectors, we consider the problem of describing and classifying the VMT vectors. In the rest of this section, we first introduce the notations and assumptions of to formulate the Gaussian mixture model. Then we explain how the Gaussian mixture model support the segmentation analysis that we conduct. The parameter estimation of the model by Expectation-Maximization (EM) algorithm is presented in G separately.

Notations and assumptions to formulate the Gaussian mixture model are as follows:

- The response variable vector of household i is denoted by $\boldsymbol{y}_i = \{y_i^1, y_i^2, ..., y_i^K\} = \{y_i^k\}_{k=1}^K$, where y_i^k is the VMT for mode k, $i = 1, ..., I$, $k = 1, ..., K$. I represents the total number of households and K represents number of modes considered. In the specific case of this study, we consider VMT for both POV/Taxi and Other Mode, and thus $K = 2$ and $\boldsymbol{y}_i = \{y_i^P, y_i^O\}$ where P stands for POV/Taxi and O stands for Other Mode. In the benchmark case where only VMT for POV/Taxi is considered, the response variable turns into a scalar of $\boldsymbol{y}_i = \{y_i\}$.
- We assume that the households is comprised of S clusters, with proportions of $\lambda_1, \lambda_2, ..., \lambda_S$. Each household belongs to one and only one cluster. Correspondingly, $\boldsymbol{\lambda} = \{\lambda_s\}_{s=1}^S$ represents the proportion vector. This mixture proportion λ_s represents the prior probability that a randomly selected household from the population belongs to cluster s. Therefore, we have $\sum_{s=1}^S \lambda_s = 1$. Also, we define a latent variable indicating which cluster an observations belongs to, which

is denoted by z_i. Specifically, $z_i = s$ if the ith observation falls in cluster s, $s = 1, 2, ..., S$.

- Each cluster is assumed to follow a multi-variate Gaussian distribution, and is characterized by a Gaussian probability function $f_s(\boldsymbol{y}_i|\boldsymbol{\theta}_s)$, representing the probability that a household belonging to cluster s makes VMT of $\boldsymbol{y}_i$. Also we denote $\boldsymbol{\Theta} = \{\boldsymbol{\theta}_s\}_{s=1}^S$, and $\boldsymbol{\theta}_s = \{\boldsymbol{\mu}_s, \boldsymbol{\Sigma}_s\}$ represents the set of parameters that define the function $f_s(\cdot)$. In the case of VMT for both modes are considered, $\boldsymbol{\mu}_s = [\mu_s^P, \mu_s^O]$, and $\boldsymbol{\Sigma}_s = \begin{bmatrix} \sigma_s^{P2} & \rho_s\sigma_s^P\sigma_s^O \\ \rho_s\sigma_s^P\sigma_s^O & \sigma_s^{P2} \end{bmatrix}$, as P representing POV/Taxi and O representing Other Mode. Then the $f_s(\cdot)$ functions, which is also the conditional probability of a household making VMT of $\boldsymbol{y}_i$ given that it belongs to cluster s, are given as follows:

$$f_s(\boldsymbol{y}_i|\boldsymbol{\theta}_s) = \frac{1}{\sqrt{(2\pi)^K|\boldsymbol{\Sigma}_s|}}\exp(-\frac{1}{2}(\boldsymbol{y}_i - \boldsymbol{\mu}_s)^T\boldsymbol{\Sigma}_s^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu}_s)) \tag{4.2}$$

Based on the notations and assumptions above, the total probability of a household making VMT of $\boldsymbol{y}_i$ is:

$$f(\boldsymbol{y}_i|\boldsymbol{\lambda}, \boldsymbol{\Theta}) = \sum_{s=1}^{S} \lambda_s f_s(\boldsymbol{y}_i|\boldsymbol{\theta}_s) \tag{4.3}$$

This total probability is the weighted sum of the conditional probabilities of each cluster s, and Equation 4.3 is referred to as finite mixture model.

Beside describing the distribution of household's VMT vector, finite mixture model also provides the cluster membership probability which is calculated by applying Bayes Law. Given $\boldsymbol{y}_i$, the probability of the household i belonging to cluster s, p_{is}, is as follow:

$$p_{is} = P(z_i = s|\boldsymbol{y}_i; \boldsymbol{\lambda}, \boldsymbol{\Theta}) = \frac{\lambda_s f_s(\boldsymbol{y}_i|\boldsymbol{\theta}_s)}{\sum_{r=1}^{S} \lambda_r f_r(\boldsymbol{y}_i|\boldsymbol{\theta}_r)} \tag{4.4}$$

Notice that, although each household is assumed to belong to only one cluster, the cluster membership probabilities Equation 4.4 yield an overlapping segmentation, which means for each household, all the S clusters have corresponding cluster membership probability respectively. The assignment of household to clusters, on the contrary, is exclusive. Each household is eventually assigned to a cluster which has the highest corresponding membership probability.

The parameters of the mixture model can then be estimated by EM algorithm. The specific parameter estimation process is shown in G.

4.4.3. Silhouette of segmentation

Silhouette is a method of representing how well the dataset is classified into clusters. The silhouette value of a single observation ranges from -1 to 1, where higher value means that the observation is well matched to the cluster that it is classified into, and poorly matched to other clusters. In our context, higher value means better homogeneity within cluster and heterogeneity between clusters.

Silhouette is calculated based on an existing segmentation, i.e., we assume that each observation in the dataset has been assigned to one of the clusters. Then for a single observation i with the cluster that it is assigned to C_i, we first define the within cluster mean distance by:

$$a(i) = \frac{1}{|C_i| - 1} \sum_{j \in C_i, i \neq j} d(i, j) \tag{4.5}$$

where $d(i, j)$ is the distance between observations i and j. In this case, we use Euclidean distance. Here $a(i)$ measures how well observation i is assigned to its own cluster. Next we define the smallest mean distance to other clusters by:

$$b(i) = \min_{k \neq i} \frac{1}{|C_k|} \sum_{j \in C_k} d(i, j) \tag{4.6}$$

which is the mean distance to one of the other clusters that has the smallest mean distance to observation i. Then we can define the silhouette value of observation i by:

$$s(i) = \frac{b(i) - a(i)}{\max a(i), b(i)} \tag{4.7}$$

where we have $-1 \leq s(i) \leq 1$. Usually we calculate the mean silhouette value of a cluster $\tilde{s}(k)$, which is the mean of silhouette values of observations in cluster k. Based on the mean silhouette value, another statistic of silhouette coefficient is usually used to measure goodness of segmentation. The silhouette is simply the maximum value of mean silhouette values among all cluster:

$$SC = \max_{k} \tilde{s}(k) \tag{4.8}$$

The silhouette coefficient will be used in Section 4.5 to demonstrate the advantage of segmentation with both VMT for POV/Taxi and Other Mode.

4.5. Results

In this section, we firstly determine the number of clusters based on statistical criteria in order to provide appropriate segmentation for further analysis. As discussed above, we compare generalized VMT vector segmentation, i.e., VMT for both POV/Taxi and Other Mode, with conventional POV/Taxi only VMT segmentation which serves as benchmark, in order to see whether VMT for Other Mode helps model people's travel behavior. Results of both cases are included and compared with each other in the determination of number of clusters. Results show that generalized VMT vector segmentation outperforms conventional segmentation by partitioning households into more representative clusters.

As stated in Section 3.4.1, we adopt datasets of New York, Seattle, and national travel survey to validate our model. The segmentation results of the 3 datasets with both generalized VMT vector and conventional POT/Taxi only VMT are presented and discussed in F. The results also validates the effectiveness of generalized VMT vector segmentation as the CMAP's data of Chicago.

Two types of analyses – disaggregate and aggregate analysis – are conducted in this section. First, based on the segmentation results, we characterize these clusters in detail to provide comprehensive profiles of them, which also serves as conclusive summary of the results. For disaggregate analysis, we describe the different travel behaviors of clusters, and relate them to built environment and SED factors. The impacts of these factors on group-level travel behavior are discussed. After that, for aggregate analysis, we first aggregate household by their SED factors to analyze the distribution of clusters among households with different characteristics. Then we aggregate zip-code travel behavior measures and perform a GIS-based spatial analysis in order to investigate inter-dependency between household travel behavior pattern and their location of residence.

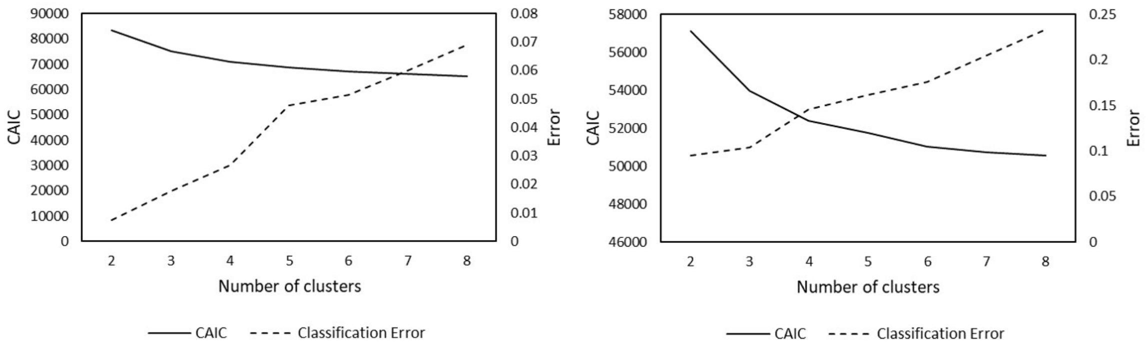
4.5.1. Number of clusters

In our analysis, we rely on the Consistent Akaike Information Criteria (CAIC) and proportion of Classification Error to assess model instances and select the best model from them. They are 2 widely used criteria to deal with the trade off between goodness of fit and the complexity (such as number of parameters) of a model.[7]

Here, we benchmark segmentation with VMT for POV/Taxi only, and compare it with segmentation with VMT for both POV/Taxi and Other Mode. The results of segmentation performance with different number of clusters of both cases of POV/Taxi + Other Mode and POV/Taxi only are shown in Figure 4.2a and Figure 4.2b. We can see that Classification Errors of the case without Other Mode are almost 3-4 times as the case

[7] The CAIC tradesoff goodness-of-fit and complexity. It is given as $CAIC \equiv -2 \cdot LL + npar(\ln(n) + 1)$, where LL is log-likelihood, n is number of observations and $npar$ is the total number of parameters. Following Vermunt and Magidson (2013), the proportion of Classification Error in overlapping cluster models is defined as $\frac{\sum_{i=1}^{I} w_i [1 - \max P(z|x_i)]}{\sum_{i=1}^{I} w_i}$, where $P(z|x_i)$ is the posterior of observation i, and w_i is sample weight for observation i. For simplification, we can assume sample weights are identical among the population, i.e., $w_i = 1$ for all observations. Intuitively, the Error is measured by weighted average of probability for an observation to be classified into the most likely cluster. For example, if the Error is 5%, we can roughly think that there is a approximately 95% probability that most of observation fall into their most likely cluster.

with Other Mode, which means we have much less confidence in the segmentation of the first case. We observe that the CAIC's decreasing rates of both cases with S larger than 4 are small. In addition, S larger than 4 also leads to a significant increases in the Classification Errors. Therefore, based on the two criteria, we choose the models with $S = 4$ for following analysis for both cases. Notice that, CAIC and Classification Error might not be proper metrics to compare across models with different structures. Thus, we only use CAIC and Classification Error for determining number of clusters, not comparing model with POV/Taxi + Other Mode and model with POV/Taxi only. Instead, silhouette coefficient, as described in Section 4.4.3, will be used below for this purpose. The estimated parameters of the two Gaussian mixture models are listed in Table 4.4. Notice that as stated in Section 4.4, the parameters of VMT for both POV/Taxi and Other Mode model in each cluster is shown in the form of "$\begin{bmatrix} \mu_s^P \\ \mu_s^O \end{bmatrix}; \begin{bmatrix} {\sigma_s^P}^2 & \rho_s \sigma_s^P \sigma_s^O \\ \rho_s \sigma_s^P \sigma_s^O & {\sigma_s^P}^2 \end{bmatrix}$", and parameters of VMT for POV/Taxi exclusively model in each cluster is shown as "$\mu_s^P; {\sigma_s^P}^2$". We see that for cluster 1 and 2, the covariances between VMT for POV/Taxi and Other Mode are really small. For cluster 3 and 4, both covariances are negative, which means that the ratio between VMT for POV/Taxi and Other Mode is fixed. Instead, for cluster 3 and 4, the homogeneity within cluster more exists in total VMT instead of travel mode pattern. The summaries of variables in each clusters are shown in Table 4.5 and Table 4.6. Based on the results, a brief profile labeling is provided in Table 4.5 for segmentation of VMT for both POV/Taxi and Other Mode.



(a) Both POV/Taxi and Other Mode

(b) POV/Taxi exclusively

Figure 4.2. CAIC and Classification Error for $S = 2, 3, ..., 8$

Table 4.4. Estimated parameters of Gaussian mixture model

	VMT for POV/Taxi and Other Mode	VMT for POV/Taxi exclusively
Cluster 1	$\begin{bmatrix} 18.77 \\ 0.01 \end{bmatrix}; \begin{bmatrix} 348.54 & -0.03 \\ -0.03 & 0.01 \end{bmatrix}$	22.02;116.92
Cluster 2	$\begin{bmatrix} 0.06 \\ 11.20 \end{bmatrix}; \begin{bmatrix} 0.16 & 0.20 \\ 0.20 & 128.35 \end{bmatrix}$	6.14;14.48
Cluster 3	$\begin{bmatrix} 13.79 \\ 4.20 \end{bmatrix}; \begin{bmatrix} 132.88 & -6.30 \\ -6.30 & 16.05 \end{bmatrix}$	0.13;0.19
Cluster 4	$\begin{bmatrix} 40.76 \\ 29.08 \end{bmatrix}; \begin{bmatrix} 2575.49 & -578.51 \\ -578.51 & 865.86 \end{bmatrix}$	61.53;1434.06

When looking at segmentation with VMT for POV/Taxi only in Table 4.6, we observe some results indicating that this segmentation doesn't partition households into representative clusters as well as segmentation with both VMT for POV/Taxi and Other Mode (as in Table 4.5). For cluster 3 in Table 4.6, which has the lowest VMT POV/Taxi, the average VMT for POV/Taxi (0.13) is much higher than cluster 2 in Table 4.5 with the lowest VMT for POV/Taxi (0.05). While both of them represent the group of households that barely use POV/Taxi, some households frequently using POV/Taxi for commute are also included in cluster 3 in Table 4.6, and different types of households still mix up

Table 4.5. Clusters profiles with both POV/Taxi and Other Mode

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Total
Cluster proportion	0.63	0.17	0.13	0.07	1.00
VMT for POV/Taxi	18.75	0.05	14.25	42.33	16.52
VMT for Other Mode	0.01	11.28	4.31	30.69	4.45
Income (in thousands $)	69.62	60.98	73.83	82.61	69.52
Household size	2.39	2.28	3.57	3.46	2.59
Number of workers	1.60	1.33	1.91	2.10	1.63
Number of students	0.51	0.62	1.55	1.30	0.72
Number of vehicles	1.98	1.05	1.94	2.25	1.83
Number of licenses	1.87	1.39	2.11	2.30	1.85
Distance to CBD	20.98	8.81	15.55	24.64	18.45
Land use mix	0.73	0.69	0.71	0.74	0.72
Population density	3120.96	6901.55	4805.36	2363.65	3930.67
Household density	1514.94	3884.06	2454.79	1064.96	2009.05

Table 4.6. Clusters profiles with POV/Taxi exclusively

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Total
Cluster proportion	0.36	0.34	0.12	0.10	1.00
VMT for POV/Taxi	24.50	6.14	0.13	74.89	16.52
VMT for Other Mode	2.213	2.85	11.27	4.05	4.45
Income (in thousands $)	74.61	67.53	61.17	78.90	69.52
Household size	2.81	2.47	2.25	3.10	2.59
Number of workers	1.81	1.54	1.33	2.07	1.63
Number of students	0.82	0.64	0.59	0.96	0.72
Number of vehicles	2.15	1.80	1.14	2.44	1.83
Number of licenses	2.05	1.79	1.43	2.29	1.85
Distance to CBD	21.95	17.60	10.78	27.51	18.45
Land use mix	0.74	0.73	0.70	0.72	0.72
Population density	2830.47	3945.87	6260.84	2544.64	3930.67
Household density	1314.73	1994.79	3489.82	1222.65	2009.05

within the cluster. The silhouette coefficients in Table 4.7, which represent the between-cluster heterogeneity and within-cluster homogeneity of segmentation, further quantifies this. According to Table 4.7, the POV/Taxi + Other Mode segmentation model is better than the POV/Taxi only segmentation model with respect to both two VMT variables. In addition, this difference not only exists in the VMT variables, but also the SED and built environment variables. The differences between clusters of number of students, household income and population/household density in Table 4.6 are not as significant as Table 4.5. This also can be seen from Table 4.7 by comparing the silhouette coefficients of these variables, and in addition to the variables mentioned above, distance to CBD also shows a better separation from the POV/Taxi + Other Mode segmentation.

To summarize, the POV/Taxi + Other Mode segmentation outperforms the POV/Taxi only segmentation in terms of the separation of both two VMT variables. Much more between-cluster heterogeneity and within-cluster homogeneity are obtained from the POV/Taxi + Other Mode segmentation. And specifically in this case, it even gives out a better separation of SED and built environment variables, although the segmentation is not conducted on these variables. Therefore, VMT for Other Mode does provide valuable information in addition to conventional VMT data. In the following sections of this paper, we will focus on the case where VMT for both POV/Taxi and Other Mode are included. In F, we present the same comparison to further demonstrate the advantage. For the New York, Seattle and National datasets, we don't include the built environment variables, so the comparison of silhouette coefficients is conducted only on VMT and SED variables.

Figure 4.3 directly presents the segmentation result of VMT for both POV/Taxi and Other Mode. Distributions of VMT data of the 4 clusters are shown in colors. It can be observed from it that cluster 1 and 2 are basically POV/Taxi and Other Mode exclusively, respectively. Cluster 3 and 4 show mixtures of both modes, and cluster 3 normally has

Table 4.7. Silhouette coefficients of POV/Taxi & Other Mode model and POV/Taxi exclusively model

	POV/Taxi and Other Mode	POV/Taxi exclusively
VMT for POV/Taxi	0.993	0.736
VMT for Other Mode	0.996	0.294
Income	0.008	-0.029
Household size	-0.065	-0.062
Number of workers	0.033	-0.052
Number of students	0.073	0.008
Number of vehicles	0.153	0.005
Number of licenses	-0.035	-0.043
Distance to CBD	0.287	0.212
Land use mix	-0.010	0.049
Population density	0.203	-0.001
Household density	0.267	0.010

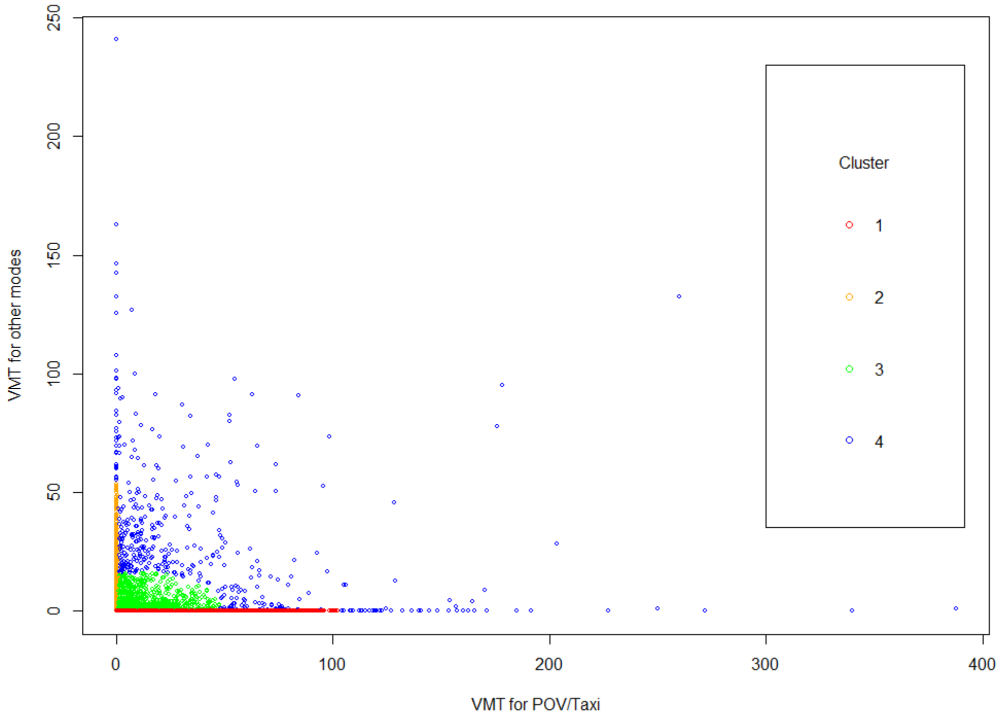


Figure 4.3. Distribution of VMT for POV/Taxi and Other Mode in clusters

less VMT than cluster 4. These observations are consistent with the profile labels in Table 4.5.

4.5.2. Cluster characterization

Based on the results above, we observe that the 4 clusters exhibit heterogeneity, i.e., the differences in either VMT for POV/Taxi, VMT for Other Mode, or both are significant. Here, we characterize the 4 clusters by their travel behavior and summarize the characteristics of their household SED and built environment factors based on the analyses below, which will be discussed in detail in following sub-sections. This characterization serves both as a summary and an introduction so the results in the rest of this section will make more sense. In particular:

- Cluster 1 (POV/Taxi dominated): Households in this cluster account for majority (63%) of the population. Residents in these households barely use travel mode other than POV/Taxi for commute, and their commute distance is relatively long if we look at the total VMT of both modes (larger than Cluster 2 and similar with Cluster 3). The average household size is relatively small (less than 3) and few of them are students (average of 0.51 per household). Even though their income, relative to other segments, is not high, the small household size means that their disposable income is high. Considering their small household, they have very high vehicle ownership (1.98) and number of license (1.87). For built environment factors, they live in areas far away from CBD with low population and household density. In conclusion, they can afford and also have the need for driving (long commute distance). On the other hand, sufficient income gives them freedom to choose their residential location and work place by using private vehicles.

- Cluster 2 (Other Mode dominated): Households in this cluster, accounting for 17% of all households, have the lowest average household income and lowest average income per household member. Moreover, they also have the lowest vehicle ownership (1.05 per household) and number of license (1.39 per household). As a result, very few of them choose driving as their travel mode for commute. For most of them, they cannot afford driving as households in cluster 1 because of income constraint. Also, the short commute distance may explain that they tend to choose workplace near their home (or the opposite, choose to live near workplace) due to the income constraint. Also, they have small household sizes (average of 2.28) with very few students (average of 0.62), so public transit is cost-effective for households in this cluster. For built environment factors, they live the closest to CBD area, which has the highest population density. This type of households mainly concentrate around downtown Chicago area. On the other hand, some households with relatively high income level also fall into this cluster. Their travel choice is more affected by the good accessibility to express transit service than income level. These households are mainly located in some suburban areas along the transit lines.
- Cluster 3 (Other Mode for student commute): Households in this cluster, accounting for 13% of all households, have large household sizes (average of 3.57) and relatively more students (average of 1.55 per household). Their average built environment factors are in between Cluster 1 and 2. A typical household in this

cluster consists of one or two school-age children and 2 adult workers as parents. Their have relatively higher total income and larger number of vehicles (average of 1.94 per household) for adult workers to drive to work and have their children taking school bus.

- Cluster 4 (Long distance with student commute): Households in this cluster, accounting for 7% of all households, have the longest commute distance for both modes (42.33 for POV/Taxi and 30.69 for Other Mode). Moreover, this cluster shows a similar household structure with cluster 3, but with higher total income and more vehicles owned (average of 2.25 per household). Also, their average distance to CBD is the largest, with the lowest population and household density. According to their long travel distance, they are likely to choose location of their house mainly based on living environment instead of the distance to work. This shows a difference with cluster 2, among which households don't have the space of choice. They are minority of the entire population.

4.5.3. Disaggregate analysis: factors on travel behavior

As previously discussed, we consider both SED and built environment factors to examine their impacts on household travel behavior. In order to evaluate these impacts, we perform correlation analysis between travel behavior and those potential factors. Moreover, in addition to analyze direct relationship between VMT and factors, we analyze the factors' impacts on cluster membership probabilities, which serve as a representation of travel behavior pattern. By using this approach, we can achieve a more accurate relationship between factors and travel behavior pattern than only the VMT measure. For

Table 4.8. Correlation coefficients between travel behavior and factors

Type	Variable	VMT		Cluster membership probability			
		POV/Taxi	Other Mode	Cluster 1	Cluster 2	Cluster 3	Cluster 4
SED	Income	0.13	0.06	0.02	-0.13	0.04	0.10
	Workers per household	0.27	0.07	-0.05	-0.19	0.15	0.19
	Students per household	0.12	0.10	-0.26	-0.05	0.32	0.17
	License per household	0.26	0.03	0.04	-0.26	0.12	0.17
	Vehicle per household	0.30	-0.03	0.20	-0.37	0.03	0.12
	Household size	0.18	0.11	-0.20	-0.11	0.29	0.19
Built environment	Distance to CBD	0.25	0.04	0.22	-0.28	-0.08	0.11
	Land use mix	0.03	0.02	0.09	-0.1	-0.05	0.03
	Population density	-0.09	-0.01	-0.12	0.14	0.04	-0.05
	Household density	-0.08	-0.01	-0.1	0.13	0.03	-0.04

example, although income's correlation coefficient with VMT for POV/Taxi is larger than correlation coefficient with VMT for Other Mode, it might not be reasonable to conclude that household with higher income tends to choose POV/Taxi instead of Other Mode. While households in cluster 1, cluster 3 and cluster 4 all have high VMT for POV/Taxi, some of them, households in cluster 4 for instance, still have demand for Other Mode commute even with high income.

Pearson correlation coefficients between VMT, cluster membership probabilities and household characteristic variables are listed in Table 4.8. SED variables including income, household size, number of workers, number of students, number of vehicles, number of licenses and built environment variables including land use mix measure, distance to CBD, household density, population density are included. To test significance of correlation, we find the two-sided 95% significance level for the correlation coefficient with sample size of 6,886. The corresponding boundaries are 0.024 and -0.024[8]. Namely, if the correlation coefficient is either greater than 0.024 or smaller than -0.024, we consider that correlation to be significant. From Table 4.8 we see that only correlations between population/household density and VMT for Other Mode are not significant. All other correlations are significant.

According to the table, overall, built environment effects are weaker than the effects of SED variables except distance to CBD. Number of workers, licenses, and vehicles in household have the strongest positive correlations with VMT for POV/Taxi (0.27, 0.26, and 0.30 respectively). As for VMT for Other Mode, number of workers and licenses have a positive but relatively weak correlation, while number of vehicles has a negative but relatively weak correlation, which is intuitively reasonable – more people means longer travel distances for both mode, while more vehicles may reduce commute of Other Mode. Distance to CBD also positively associated with VMT for POV/Taxi, which is very likely to be explained by the general lack of accessibility to transit service in areas far away from CBD, although there are some exceptions of areas far away from CBD with good accessibility to transit service. However, this strong correlation does not indicate direct relationship between VMT for POV/Taxi and distance from household location to CBD, since only a minority of employments in Chicago are located in downtown Chicago area. This correlation might be caused by the inter-dependencies between household location, transit accessibility, and distance to CBD. This can be explained by arguing that lack of transit accessibility generally means being far way from CBD, which is because the household does not depend on transit service. Therefore, they are more likely to have larger VMT for POV/Taxi.

[8] To test significance of correlation coefficient, the null hypothesis is that population correlation coefficient $\rho = 0$. Then we calculate the test statistic $t = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}}$, where r is sample correlation coefficient and N is sample size. The test statistic t follows a Student's t-distribution with degree of freedom $N - 2$. So in our case with $N = 6886$, by calculating quantiles $t_{0.025,6884}$ and $t_{0.975,6884}$, the corresponding $r = \pm 0.024$ then can be found.

More importantly, we analyze the factors on travel behavior pattern, which is described as cluster membership. Number of students per household, which seems to have relatively weak effect on VMT(0.12), actually have strong correlation with cluster membership probabilities of being in cluster 1 (-0.26) and cluster 3 (0.32). That is, a household with more students means that it's less likely to be in cluster 1 and more likely to be in cluster 3. Household size shows the similar effect (-0.20 and 0.29 respectively). Another finding is that even though vehicle ownership is positively associated with cluster 1, which is specified as group of exclusive drivers, number of licenses per household seems only have negative strong impact on cluster 2, but not on cluster 1. That is possibly due to the fact that people with license do not necessarily own vehicles.

Among built environment factors, distance to CBD has the strongest effect on cluster membership, which is positively associated with cluster membership probability of being in cluster 1 and negatively with being in cluster 2. Most of the others are significant, but much weaker compared with distance to CBD. This indicates that the built environment factors don't have as much impact on household travel pattern as SED factors. People tend to choose travel mode based more on their income level and household structure than the environment factors in their neighborhood. Also, although distance to CBD, the only one out of built environment factors, has significant effects, it is important to notice that correlations between distance to CBD and VMT for Other Mode is non-significant. Especially when its strong negative correlation with cluster membership of cluster 2 is considered, this means that although households in cluster 2 are unlikely to be far away from CBD, the distance to CBD doesn't affect the choice of taking other travel mode (transit in this case) very much. In other words, it's more likely that people in cluster 2 choose transit because of their income level and household structure. Besides, as discussed above, the the strong correlation with distance to CBD might be due to the inter-dependencies between factors, since people working in downtown Chicago area are only a minority of the population. Namely, households in cluster 2 tend to live around places with good accessibility to transit service which have the trend of being close to CBD, although their travel behavior, which is mostly determined by socioeconomic factor, is not sensitive to where they live. Therefore, the SED factors still seem to be the motivation of travel behavior in this case.

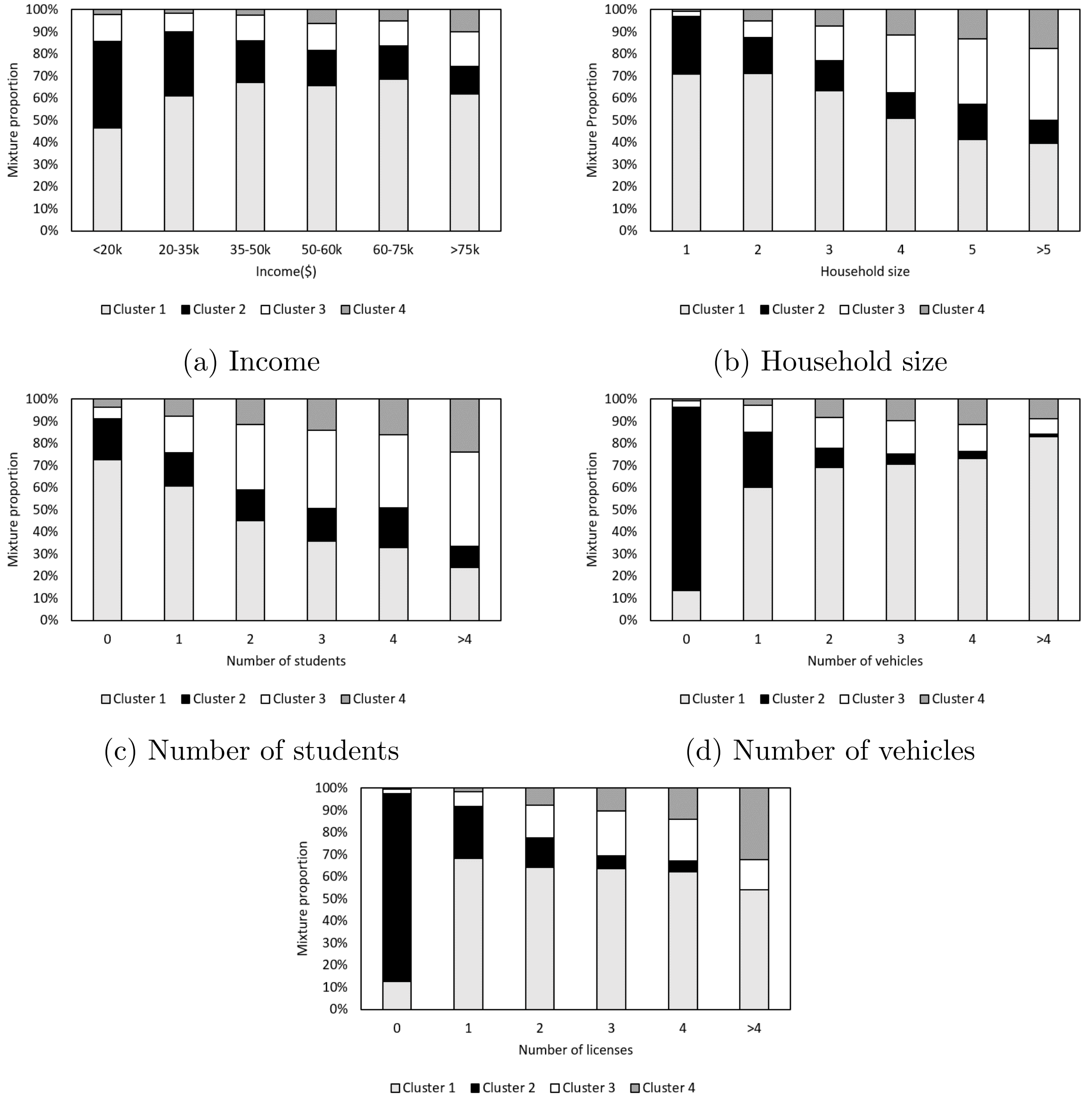
4.5.4. Aggregate analysis: mixture proportions

The above correlation analysis gives us statistical correlation between household characteristics and travel behavior. However, this approach requires linear relationship and doesn't work well with categorical data. Therefore, we also analyze the distribution of mixture proportion over some categorical factors, including income[9], number of students, household size, number of licenses and vehicle ownership[10] in order to further investigate their impacts, as shown in Figure 4.4. Namely, we aggregate the segmentation result by these categorical factors, and analyze how cluster proportions are distributed within single categories of the factors.

Figure 4.4a shows changes of mixture proportions by varying income range. We observe that proportion of cluster 2 keeps dropping, while proportion of cluster 4 increases

[9] Income data collected from survey is categorical, but we convert it to numerical data for the correlation analysis above. Here in the mixture proportion analysis, we use original categorical data.

[10] The strongest correlated factor of built environment - distance to CBD – is a continuous measure, which fits better in correlation analysis above. Moreover, we also examine its effect in the following spatial pattern analysis section. So we don't present analysis on built environment factors in this section.



(a) Income

(b) Household size

(c) Number of students

(d) Number of vehicles

(e) Number of licenses

Figure 4.4. Mixture proportions of clusters with SED factors

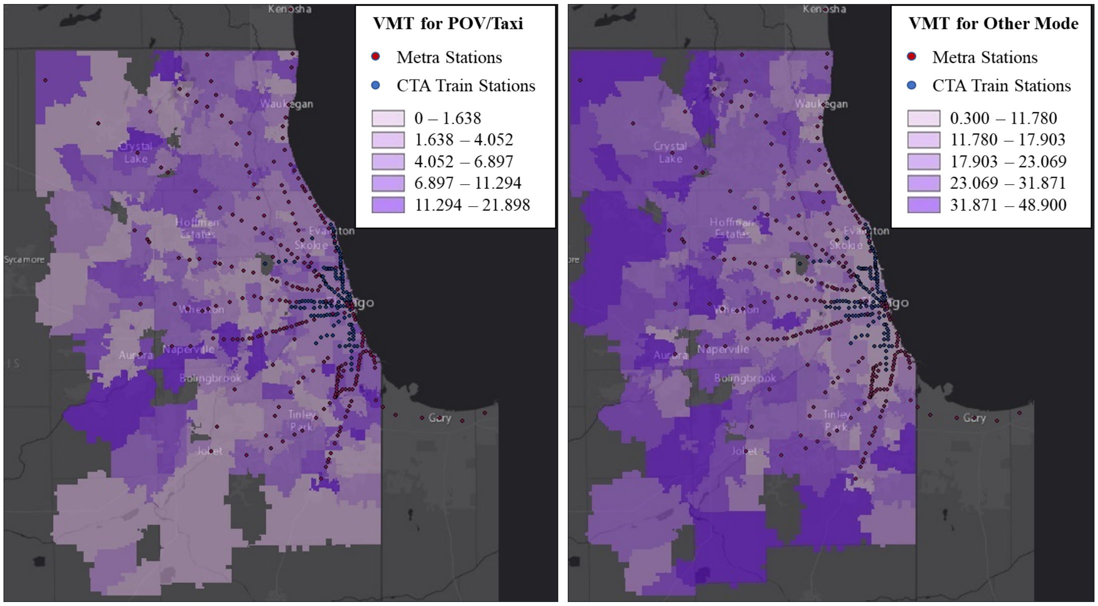
as household income increases, which is consistent with our discussion that households in cluster 2 are restricted by income while households in cluster 4 can afford driving. However, cluster 3 are not that sensitive to change of income, which remains the proportion about 15% for each income range. Cluster 1 has the same property except for income less than 20k. Mixture proportions are sometimes highly sensitive to factors other than income. For example, households in cluster 3 typically have students. No matter how much their parents earn, most of them have to take school bus.

Figure 4.4b shows the relationship between mixture proportions and household size. Proportions of cluster 3 and 4 increase dramatically when household size increases. One finding worth noticing is that proportion of cluster 1 drops while proportion of cluster 3 increases as household size goes up. However, cluster 1 has longer average total commute distance than cluster 3, which indicates that larger household size doesn't necessary mean larger commute distance. Consequently, we need to focus on the specific structure of household instead of the household size only. As discussed above, there usually is at least one student in each household in cluster 3. Students and their parents form a typical household in cluster 3 which has at least 3 people per household. We can also explain why proportion of cluster 2 is higher in small households than in large households. Households in cluster 2 use transit for their commute and have lowest vehicle ownership. Because of their relatively low income, they care about the cost of commute very much. In a large household, residents may want to buy a car for commute so that every one can use it instead of paying transit fare for everyone. Being consistent with correlation analysis above, the relationship between mixture proportion and number of students is almost the same as relationship between mixture proportion and household size, as can be observed in Figure 4.4c.

Figure 4.4d shows the relationship between mixture proportions and vehicle ownership, i.e., number of vehicles owned. Proportion of cluster 2, the Other Mode commuters, dominates households without vehicle and this proportion drops greatly when vehicle ownership is increasing. On the other hand, proportion of cluster 1 keeps increasing with vehicle ownership increasing, which is consistent with the result that households in cluster 1 almost have no share for Other Mode. Cluster 4 has the similar feature as cluster 1. Besides, cluster 3 is relatively uniformly distributed among all vehicle occupancy level, which might have the same reason as our analysis on income that proportion of cluster 3 is more likely to be affected by household structure (with students). The impact of number of licenses, as shown in Figure 4.4e, is similar with the impact of vehicle ownership. However, there are some differences in the mixture proportion of cluster 1. Mixture proportion of cluster 1 reaches its maximal when there is only 1 license in household. A potential reason is that large number of licenses indicates large household size, while cluster 1 is characterized by small household size, relatively short commute distance and heavy dependence on driving. The reason that proportion of cluster 1 still being high is that cluster 1 is the majority of the households (with overall proportion of 63%) after all.

4.5.5. Aggregate analysis: spatial pattern

In this section, we aggregate the household VMT data and segmentation result by their regional area. With the help of GIS system, we plot segmentation results and VMT data in Northeastern Illinois map. In order to analyze spatial pattern, we divide our analysis area into zip code blocks and calculate corresponding statistics per zip code block. Figure 4.5a and Figure 4.5b show the pattern of average VMT for POV/Taxi and VMT for Other Mode per block. And Figure 4.6a to Figure 4.6d show the pattern of average cluster membership probability of each cluster per block. The deeper color means the higher probability/VMT it has.

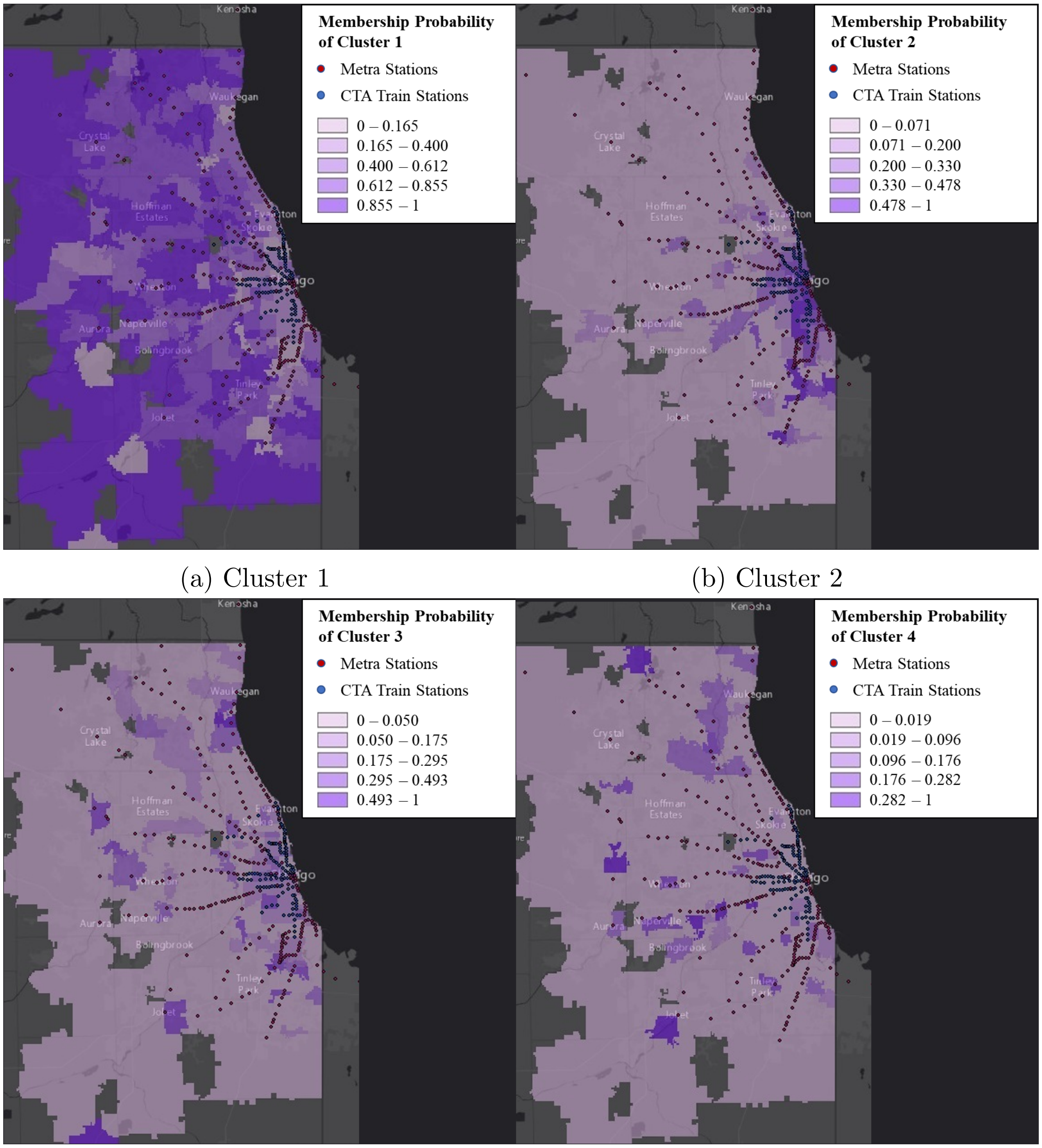


(a) VMT for POV/Taxi

(b) VMT for Other Mode

Figure 4.5. VMT distribution map

First, we noticed that map for VMT for POV/Taxi and VMT for Other Mode are complementary – most of areas have a deep color for one and light for the other, which shows that most of areas are dominated by only one travel mode. However, Chicago downtown area shows similar levels for both modes, indicating households in this area have both many POV/Taxi trips and Other Mode trips for commute. This is an issue with this type of aggregate analysis: can we conclude that, in Chicago downtown area, most households make use of both two travel modes for commute and transit-users and drivers are mixed up here? The answer is no only based on these two maps, because zonal summaries keep the heterogeneity in each geographic unit area – we don't distinguish different clusters of household within the areas. Analyses of clusters in the following figures will tell more stories.



(a) Cluster 1

(b) Cluster 2

(c) Cluster 3

(d) Cluster 4

Figure 4.6. Cluster membership distribution map

We then focus on the spatial distribution of clusters. Because cluster 1 has the most share of the entire population (63%), colors of most areas in Figure 4.6a are deep, except some isolated suburb areas near Joliet and Aurora and Chicago downtown area. Distribution of cluster 2, which represents the transit dependent commuters, is concentrated, mostly located in Chicago downtown area, south Chicago, and some suburbs with transit coverage. This is reasonable because these areas are mostly covered by CTA trains and other public transports, and households in cluster 2 highly rely on Other Mode commute. There are some households in cluster 3 located in Chicago downtown area, and the rest of them is located in surrounding cities like Elgin, Wheaton, and Joliet. Based on the fact that most households in cluster 3 have students, their parents may also take children's education into consideration for the location selection of their houses (for example, choose to live near good-quality schools, especially public schools). Distribution of cluster 4 is very isolated, and there are very few of them located in Chicago downtown area compared with other clusters. Those residents are likely to choose to live in suburbs than urban area, which is also consistent with their very long commute distance. Recall the problem above, we realize that households in Chicago downtown area have strong heterogeneity in travel behavior. Most of households in cluster 2, some of households in cluster 1, some in cluster 3 and very few in cluster 4 form the entire population of households in Chicago downtown area.

We also present the dominant cluster in each geographic units in Figure 4.7. Cluster 1 covers most part of the map, while cluster 2 covers area near CTA red, green, purple, brown, yellow lines. Cluster 3 and cluster 4 dominate some isolated areas in suburbs. This shows a more direct illustration of the same finding that cluster 2 dominates the Chicago downtown and South Chicago areas. While most units are covered by cluster 1, cluster 3 and 4 dominate some suburban areas.

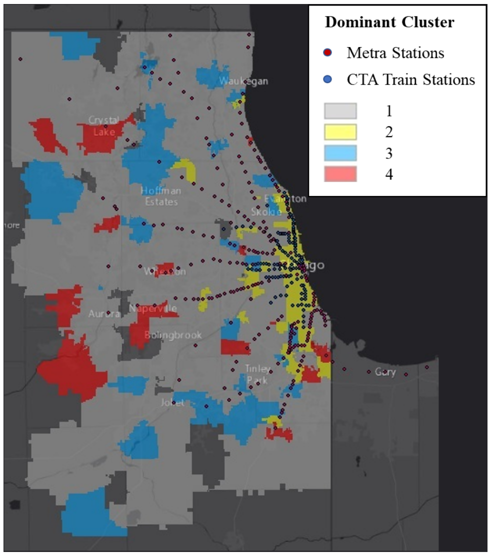


Figure 4.7. Dominant cluster map

An interesting finding from Figure 4.7 is that some areas featured by high income households, for example Evanston, are dominated by cluster 2. This is consistent with the fact that people in Evanston tend to take transit because of the convenient access to Metra and CTA express line regardless of the income level. These households consist of the part of cluster with high income level. Therefore, within cluster 2 where most of households featured by relatively low income, there are also some households choose Other Mode because of the accessibility to transit service. According to Figure 4.7, those suburban areas dominated by cluster 2 are all along the transit lines, which is consistent with our discussion.

4.6. Discussion

In this section, we provide comprehensive discussions on the respective effects of SED and built environment factors on travel behavior pattern based on the results we obtain. Based on the analysis results, we also discuss the policies that are relevant, both efficiency of current policies and possibility of future policies. In addition, some limitations of the study are also discussed.

4.6.1. Effects of SED factors on travel behavior pattern

In the previous section we have examined the heterogeneity in household's travel behavior pattern. The four clusters, which we have characterized in Section 4.5.1, represent different households with different characteristics: households with enough income for driving; with lower income and having to take transit to work; with high income but having students to take transit; with high income, having students to take transit, and living far way from workplace. The factors of income (which consequently affect vehicle ownership) and household structure (including household size and number of students) have the most impact on households' travel pattern.

Moreover, based on results of disaggregate analysis, the way that household characteristics affect travel behavior pattern is not only non-linear but also discontinuous. For vehicle ownership measure, there is a significant difference in mixture proportions of cluster 1 and cluster 2 between households without a vehicle and households with only one vehicle. However, the extra vehicles do not make significant difference in these two clusters. Besides, as discussed in the previous section, the location of schools, which is not reflected in this study, might also have a great impact on households in cluster 3 and 4. This leads to the discussion that, the location of these 2 types of households is a result of both the household structure and income level, and the location further affects their travel patterns. This makes the distribution of cluster 3 and 4 more discrete. Moreover, overall, income level has a negative correlation with mixture proportion of cluster 2, but mixture proportion of cluster 2 does not significantly drops for group of people with more than 50k income. This is consistent with the discussion that some households with considerable income level, for example, some households in Evanston, tend to take transit because of the good accessibility to express transit service. We learn that the effects of household characteristics on travel behavior pattern is too complicated to be modeled as a functional relationship.

4.6.2. Effects of urban form and built environment factors on travel behavior pattern

Based on disaggregate analysis, land use factors have non-significant effects on travel behavior pattern. The only significant one – distance to CBD – is surprising since only a minority of the whole population work in downtown Chicago area. As discussed in the previous section, this correlation might result from the inter-dependencies between household location, transit accessibility, and distance to CBD. As households in cluster 2 mainly located in Chicago downtown area, they are basically covered by CTA train, Metra and bus service. After showing the distribution of households in cluster 2 approximately matches the coverage of CTA train and Metra, we believe households in cluster 2 select their location of residence based on accessibility to transit facilities rather than distance to CBD. This turns out to be similar with the conclusion in Section 4.5.3, which is made based on disaggregate analysis. The strong correlation between travel behavior is more of a result of the accessibility of transit service, and areas covered by transit service generally are relatively close to CBD area. Therefore, instead of that distance to CBD determines people's travel mode choice, their mode choice is affected by their SED factors, and the mode choice leads to their choice of living location. Furthermore, the spatial clustering phenomenon of people, as shown in Figure 4.6a to Figure 4.6d, might also affect the built environment and urban form in turn. However, living in areas with good transit accessibility does not necessarily mean that the household belongs to cluster 2 or is transit-dependent. As discussed above, many other factors may lead to the result of living close to transit and CBD area.

The influence of urban form on people's travel behavior is complicated. The land use, home value and built environment of each neighborhood are predetermined by urban planning. Then heavy transit, as a part of built environment, is built based on and affected by local residents. After that, people with different SED characteristics would chose their living places or relocate according to built environment (or macroscopically speaking, urban form), which also has an impact on public transport planning in turn. Afterwards, land use, urban form, public transport and people's travel behavior keep on interacting with each other. Just like Handy et al. (2005) argues, we can hardly determine if built environment and land use have a causal effect on travel behavior. Consequently, we really need to have a second thought on a popular belief that change of urban form will efficiently change people's travel behavior. We believe that household's SED factors have more causal effects on travel behavior than simply urban form structure. It is possible that people from different clusters with different travel behaviors and household's SED live in the same area. Some of them show strong location-dependent feature, while some of them have much more freedom to choose where they live. Similarly, people within a zone may or may not have the same travel pattern, but they do show the travel behavior of a mixture of different types. That is, to some extent, a good example of the quote in McFadden (2000)'s review - “Zones don't travel; people travel!”.

4.6.3. Relevant policy

Based on the analysis of travel behavior segmentation and spatial pattern, some policy insights can be informed by the results regarding both the efficiency of current policies and possibility of future policies. For current policy, we focus on whether the location of new station or renewal of old station is reasonable and efficient based on our analysis. For future policy, we focus on how subsidy should be distributed among people in different clusters.

The spatial pattern analysis shows the dominating clusters and cluster probabilities of the geographic units. Therefore, whether a transit line or a transit station will be fully utilized by passengers can be inferred from the spatial pattern analysis. Ideally, construction of new station or renewal of old station should focus on areas where heavy demand of transit service is observed, i.e., areas dominated by cluster 2-4, especially cluster 2. For example, in southwest Chicago, the new Joliet Transportation Center was completed and brought into service in early 2018 after the 6-year construction. The new station serves both Amtrak and Metra trains. According to Figure 4.6 and 4.7, the location of the new Joliet Transportation Center is the a large area of cluster 3, with a considerable group of households in cluster 4. Therefore, there's a great probability that the new station will cover a large population who actually need the better transit service, which indicates that the our analysis result is consistent with the transit planning. Another recent proposal of extending CTA Red Line further to the south is also supported by our analysis result. From Figure 4.7, we can observe that several cluster 2 dominated areas exist at the further south of the current destination of Red Line. Although these areas are covered by Metra ME Line, the extension still might greatly improve the transit service in those areas, especially considering that CTA provides a much lower fare than Metra.

Another important relevant policy is subsidy. In order to encourage people to shift to transit from POV/Taxi, subsidizing people is a commonly used and efficient method. However, the subsidy should be targeted on certain group of people to maximize the efficiency. Based on our analysis, households in cluster 2 are already highly dependent on transit. Although cluster 3 and cluster 4 have some need for transit service, which makes them seem easy to be encouraged, the long distance of commute and poor transit accessibility of many households in these two clusters make it hard for them to shift to transit. For cluster 1, the spatial pattern shows that many areas well covered by transit service of both Metra and CTA are still dominated by cluster 1. Also, most households in cluster 1 have small households sizes, which makes it inefficient to drive. Therefore, targeting households in cluster 1 with small household size and living near transit stations, based on our analysis, would be the best subsidy policy.

4.6.4. Limitations

As stated in Section 4.3.1, non-commute trips chained to commute trips are excluded. Although this exclusion makes the study focus on commute trips, it might also have limited capability of capturing some more complicated travel behaviors than simply going to and coming back from work/school. For example, if a person regularly drop by a grocery store on the way home (for example, daily or weekly on Friday), intuitively in the survey both the two sections – from work to grocery store and from grocery store to home – should be classified as commute trip. However, our model and analysis would fail to reflect this special travel pattern, which actually seems to be a mixture of commute and non-commute trip. Namely, this study focuses on the simple travel pattern of going to and coming back from work/school. Travel behaviors that are more complicated cannot be precisely reflected. Also, in this example, the total distance of the two sections of trip might be larger than the actual distance between working place and home, which may cause bias to some extent (though the bias is unlikely to be large, since taking a large detour to drop by grocery store is uncommon).

The other limitation mentioned in Section 4.3.1 is that we double-count the VMT in some cases where multiple household members share the same POV for commute. The realistic issue is that due to the limited information available in the survey, we cannot tell if two members in a household were in the same vehicle for all cases, and thus cannot avoid the double-counting. However, this double-counting can still be reasonable, since although the VMT that actual happened should be only counted once, there is a difference between household members sharing a POV and only the driver driving the vehicle. For example, a couple in a household shares a car for commute with a driving distance of 10 miles, and another single person drives to work alone with a same driving distance of 10 miles. If we don't double-count the VMT, then for the two households, the household VMT will be the same, which obviously ignore the difference between the travel behaviors of the two households. Besides, another argument is that, some consequences of VMT that would be directly affected by the double-counting of VMT are not directly considered in this study.

For studies evaluating the energy consumption or emission caused by VMT, the double-counting will definitely affect the precision of quantifying those measures. However in this paper, the double-counting only captures the travel behavior, but does not affect those consequences of VMT because they are not considered here. But the limitation would be that we cannot directly analogize the results of VMT analysis in this study to energy consumption and emission field, since the VMT we use is not precise enough for energy consumption and emission evaluation.

4.7. Conclusion

We perform a multi-dimensional finite mixture model for joint segmentation analysis of the VMT for both POV/Taxi and Other Mode to capture the travel behaviors of households with different characteristics. The segmentation result shows that heterogeneity does exist in travel behavior between different clusters. Statistics of important SED and built environment factors are summarized, and correlation analysis between VMT, cluster membership probabilities and the factors. Based on the statistics and correlation analysis, and by comparing SED factors in different clusters, we come up with qualitative effects of those variables on generalized VMT information and cluster membership. The effects of factors indicate that, in this case, household's travel behavior is more affected by SED factors than built environment factors. Specifically, vehicle ownership, number of students in household and household size are the most significant factors. Within built environment factors, only distance to CBD is shown to be significant. However, based on our discussion, the significance of distance to CBD is due to the inter-dependencies between factors, but itself does not have direct impact on household's travel behavior.

A spatial pattern analysis of households in different clusters is also provided. And we find, for some clusters such as cluster 2, the distribution of households has a strong dependence on spatial location. But for other clusters such as cluster 1, the distribution of households only follows the distribution of total population. But with the result obtained by statistics of SED variables, we still explored possible explanations for these phenomenon. Moreover, we have the reason to believe, in some case, people with same driving behavior don't necessarily live in the same area, as opposed to "neighborhood" clustering methods defined in the literature. After the analyses, we characterize the 4 clusters by summarizing their household and travel behavior characteristics.

Policy insights are also informed by the factor impact and spatial pattern analysis. Ideal location for new transit station or renewal of old station can be inferred from the distribution of cluster 2, 3, and 4. Two examples of completed and proposed project, the recent renewal of Joliet Transportation Center and proposed extension plan of CTA Red Line, are found to be supported by our result. The two project both focus on areas with high demand for transit service, which validates the efficiency of the policies. Also, potential target population of subsidy policy is discussed. Within the whole population, targeting households in cluster 1 with small household size and living near transit stations might be the most efficient way to encourage more people to shift to transit.

In addition, segmentation results also help us understand the complication of interdependency among SED, urban form, transit service and people's travel behavior pattern. Unlike conventional analysis which focus on correlation between factors and VMT or mode choice measure, we focus on impacts on travel behavior pattern which are results of cluster membership probability. That is, we reveal the essential latent variable describing people's travel behavior and then link it to factors. The result suggests that changes in urban form variables, such as mixed land use policy, may not directly take effect in a short period of time. People's commute mode choice and household structure may play a bigger role. We can hardly make people switch from driving commuters to transit commuters without improving public transit accessibility and change people's attitude towards transit.

References

Aaker, J., Vohs, K. D., and Mogilner, C. (2010). Nonprofits are seen as warm and for-profits as competent: Firm stereotypes matter. *Journal of Consumer Research*, 37(2):224–237.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, 43(4):561–573.

Aron, A., Aron, E. N., and Smollan, D. (1992). Inclusion of other in the self scale and the structure of interpersonal closeness. *Journal of Personality and Social Psychology*, 63(4):596.

Bekkers, R. and Wiepking, P. (2007). Understanding philanthropy: A review of 50 years of theories and research. In *35th annual conference of the Association for Research on Nonprofit and Voluntary Action, Chicago*.

Bekkers, R. and Wiepking, P. (2011). A literature review of empirical studies of philanthropy: Eight mechanisms that drive charitable giving. *Nonprofit and Voluntary Sector Quarterly*, 40(5):924–973.

Bennett, R. (2006). Predicting the lifetime durations of donors to charities. *Journal of Nonprofit & Public Sector Marketing*, 15(1-2):45–67.

Bhat, C. R. (1997a). An endogenous segmentation mode choice model with an application to intercity travel. *Transportation science*, 31(1):34–48.

Bhat, C. R. (1997b). Work travel mode choice and number of non-work commute stops. *Transportation Research Part B: Methodological*, 31(1):41–54.

Bhat, C. R. (1998). Accommodating variations in responsiveness to level-of-service measures in travel mode choice modeling. *Transportation Research Part A: Policy and Practice*, 32(7):495–507.

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

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

Boarnet, M. and Crane, R. (2001). The influence of land use on travel behavior: specification and estimation strategies. *Transportation Research Part A: Policy and Practice*, 35(9):823–845.

Boarnet, M. G. and Chalermpong, S. (2001). New highways, house prices, and urban development: A case study of toll roads in orange county, ca. *Housing Policy Debate*, 12(3):575–605.

Boarnet, M. G. and Sarmiento, S. (1998). Can land-use policy really affect travel behaviour? a study of the link between non-work travel and land-use characteristics. *Urban Studies*, 35(7):1155–1169.

Boatwright, P., Borle, S., and Kadane, J. B. (2003). A model of the joint distribution of purchase quantity and timing. *Journal of the American Statistical Association*, 98(463):564–572.

Buuren, S. v. and Groothuis-Oudshoorn, K. (2010). mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, pages 1–68.

Chatman, D. (2003). How density and mixed uses at the workplace affect personal commercial travel and commute mode choice. *Transportation Research Record: Journal of the Transportation Research Board*, (1831):193–201.

Clotfelter, C. T. (2003). Alumni giving to elite private colleges and universities. *Economics of Education Review*, 22(2):109–120.

CMAP (2008). Traveler tracker survey 2007-2008. `https://www.cmap.illinois.gov/data/transportation/travel-survey`. [Online].

Council for Aid to Education, New York, N. (2020). *Voluntary support of education 2019*. Council for Aid to Education.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22.

Domencich, T. A. and McFadden, D. (1975). Urban travel demand-a behavioral analysis. Technical report.

Dunn, P. K. (2004). Occurrence and quantity of precipitation can be modelled simultaneously. *International Journal of Climatology*, 24(10):1231–1239.

Durango-Cohen, E. J., Torres, R. L., and Durango-Cohen, P. L. (2013a). Donor segmentation: When summary statistics don't tell the whole story. *Journal of Interactive Marketing*, 27(3):172–184.

Durango-Cohen, P. L., Durango-Cohen, E. J., and Torres, R. L. (2013b). A bernoulli–gaussian mixture model of donation likelihood and monetary value: An application to alumni segmentation in a university setting. *Computers & Industrial Engineering*, 66(4):1085–1095.

Dvorak, T. and Toubman, S. R. (2013). Are women more generous than men? evidence from alumni donations. *Eastern Economic Journal*, 39(1):121–131.

Ein-Gar, D. and Levontin, L. (2013). Giving from a distance: Putting the charitable organization at the center of the donation appeal. *Journal of Consumer Psychology*, 23(2):197–211.

Elldér, E. (2014). Residential location and daily travel distances: the influence of trip purpose. *Journal of Transport Geography*, 34:121–130.

Ewing, R. and Cervero, R. (2010). Travel and the built environment: a meta-analysis. *Journal of the American Planning Association*, 76(3):265–294.

Fischer, R. (2004). Standardization to account for cross-cultural response bias: A classification of score adjustment procedures and review of research in jccp. *Journal of Cross-Cultural Psychology*, 35(3):263–282.

Friedman, H. H. and Amoo, T. (1999). Multiple biases in rating scale construction. *Friedman, HH and Amoo*, pages 115–126.

Gaba, A. and Winkler, R. L. (1992). Implications of errors in survey data: a bayesian model. *Management Science*, 38(7):913–925.

Gaier, S. (2005). Alumni satisfaction with their undergraduate academic experience and the impact on alumni giving and participation. *International Journal of Educational Advancement*, 5(4):279–288.

Green, P. E. (1977). A new approach to market segmentation. *Business Horizons*, 20(1):61–73.

Greene, W. H., Hensher, D. A., and Rose, J. (2006). Accounting for heterogeneity in the variance of unobserved effects in mixed logit models. *Transportation Research Part B: Methodological*, 40(1):75–92.

Greenleaf, E. A. (1992). Improving rating scale measures by detecting and correcting bias components in some response styles. *Journal of Marketing Research*, 29(2):176–188.

Grün, B. and Dolnicar, S. (2016). Response style corrected market segmentation for ordinal data. *Marketing Letters*, 27(4):729–741.

Grün, B. and Leisch, F. (2008). FlexMix version 2: finite mixtures with concomitant variables and varying and constant parameters. *Journal of Statistical Software*, 28(4):1–35.

Gunsalus, R. (2005). The relationship of institutional characteristics and giving participation rates of alumni. *International Journal of Educational Advancement*, 5(2):162–170.

Handy, S. (1996). Methodologies for exploring the link between urban form and travel behavior. *Transportation Research Part D: Transport and Environment*, 1(2):151–165.

Handy, S., Cao, X., and Mokhtarian, P. (2005). Correlation or causality between the built environment and travel behavior? evidence from northern california. *Transportation Research Part D: Transport and Environment*, 10(6):427–444.

Hoeffding, W. (1948). A non-parametric test of independence. *The Annals of Mathematical Statistics*, pages 546–557.

Holmes, J. (2009). Prestige, charitable deductions and other determinants of alumni giving: Evidence from a highly selective liberal arts college. *Economics of Education Review*, 28(1):18–28.

Holmes, J. A., Meditz, J. A., and Sommers, P. M. (2008). Athletics and alumni giving: Evidence from a highly selective liberal arts college. *Journal of Sports Economics*, 9(5):538–552.

Hong, J., Shen, Q., and Zhang, L. (2014). How do built-environment factors affect travel behavior? a spatial analysis at different geographic scales. *Transportation*, 41(3):419–440.

Huang, G. and Sudhir, K. (2021). The causal effect of service satisfaction on customer loyalty. *Management Science*, 67(1):317–341.

Ilieva, J., Baron, S., and Healey, N. M. (2002). Online surveys in marketing research. *International Journal of Market Research*, 44(3):1–14.

Javaras, K. N. and Ripley, B. D. (2007). An "unfolding" latent variable model for likert attitude data: Drawing inferences adjusted for response style. *Journal of the American Statistical Association*, 102(478):454–463.

Jen, L., Chou, C.-H., and Allenby, G. M. (2009). The importance of modeling temporal dependence of timing and quantity in direct marketing. *Journal of Marketing Research*, 46(4):482–493.

Jørgensen, B. and Paes De Souza, M. C. (1994). Fitting tweedie's compound poisson model to insurance claims data. *Scandinavian Actuarial Journal*, 1994(1):69–93.

Jung, H. S. and Yoon, H. H. (2013). Do employees' satisfied customers respond with an satisfactory relationship? the effects of employees' satisfaction on customers' satisfaction and loyalty in a family restaurant. *International Journal of Hospitality Management*, 34:1–8.

Katz, R. W. (2002). Stochastic modeling of hurricane damage. *Journal of Applied Meteorology and Climatology*, 41(7):754–762.

Koppelman, F. S. and Sethi, V. (2005). Incorporating variance and covariance heterogeneity in the generalized nested logit model: an application to modeling long distance travel choice behavior. *Transportation Research Part B: Methodological*, 39(9):825–853.

Kramer, T. and Yoon, S.-O. (2007). Approach-avoidance motivation and the use of affect as information. *Journal of Consumer Psychology*, 17(2):128–138.

Kressner, J. and Garrow, L. (2012). Lifestyle segmentation variables as predictors of home-based trips for atlanta, georgia, airport. *Transportation Research Record: Journal of the Transportation Research Board*, (2266):20–30.

Le Blanc, L. A. and Rucks, C. T. (2009). Data mining of university philanthropic giving: Cluster-discriminant analysis and pareto effects. *International Journal of Educational Advancement*, 9(2):64–82.

Leslie, L. L. and Ramey, G. (1988). Donor behavior and voluntary support for higher education institutions. *The Journal of Higher Education*, 59(2):115–132.

Levontin, L., Ein-Gar, D., and Lee, A. (2015). Acts of emptying promote self-focus: A perceived resource deficiency perspective. *Journal of Consumer Psychology*, 25(2):257–267.

Lin, J. and Long, L. (2008). What neighborhood are you in? empirical findings of relationships between household travel and neighborhood characteristics. *Transportation*, 35(6):739–758.

Lindahl, W. E. and Conley, A. T. (2002). Literature review: Philanthropic fundraising. *Nonprofit Management and Leadership*, 13(1):91–112.

Lindahl, W. E. and Winship, C. (1992). Predictive models for annual fundraising and major gift fundraising. *Nonprofit Management and Leadership*, 3(1):43–64.

Lindsey, M., Schofer, J. L., Durango-Cohen, P., and Gray, K. A. (2011). The effect of residential location on vehicle miles of travel, energy consumption and greenhouse gas emissions: Chicago case study. *Transportation Research Part D: Transport and Environment*, 16(1):1–9.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Routledge.

Ma, J. and Goulias, K. (1996). Multivariate marginal frequency analysis of activity and travel patterns in first four waves of puget sound transportation panel. *Transportation Research Record: Journal of the Transportation Research Board*, (1556):67–76.

Mair, P. and Hatzinger, R. (2007). Extended rasch modeling: The erm package for the application of irt models in r. *Journal of Statistical Software*, 20(9):1–20.

Masters, G. N. (1982). A rasch model for partial credit scoring. *Psychometrika*, 47(2):149–174.

McAlexander, J. and Koenig, H. (2001). University experiences, the student-college relationship, and alumni support. *Journal of Marketing for Higher Education*, 10(3):21–44.

McDearmon, J. T. (2013). Hail to thee, our alma mater: Alumni role identity and the relationship to institutional support behaviors. *Research in Higher Education*, 54(3):283–302.

McFadden, D. (2000). Disaggregate behavioral travel demand's rum side. *Travel Behaviour Research*, pages 17–63.

McFadden, D. and Reid, F. (1975). *Aggregate travel demand forecasting from disaggregated behavioral models*. Institute of Transportation and Traffic Engineering, University of California.

McLachlan, G. and Peel, D. (2004). *Finite mixture models*. John Wiley & Sons.

Meer, J. and Rosen, H. S. (2009). The impact of athletic performance on alumni giving: An analysis of microdata. *Economics of Education Review*, 28(3):287–294.

Merchant, A., Ford, J. B., and Sargeant, A. (2010). 'don't forget to say thank you': The effect of an acknowledgement on donor relationships. *Journal of Marketing Management*, 26(7-8):593–611.

Monks, J. (2003). Patterns of giving to one's alma mater among young graduates from selective institutions. *Economics of Education Review*, 22(2):121–130.

Moon, S. and Azizi, K. (2013). Finding donors by relationship fundraising. *Journal of Interactive Marketing*, 27(2):112–129.

Naess, P. (2011). 'New urbanism' or metropolitan-level centralization? a comparison of the influences of metropolitan-level and neighborhood-level urban form characteristics on travel behavior. *Journal of Transport and Land Use*, 4(1):25–44.

NYMTC (2011). 2010/2011 Regional Household Travel Survey (RHTS). `https://www.nymtc.org/DATA-AND-MODELING/Travel-Surveys/2010-11-Travel-Survey`. [Online].

Okunade, A. A. (1996). Graduate school alumni donations to academic funds: micro-data evidence. *American Journal of Economics and Sociology*, 55(2):213–229.

Öztürk, A. (1981). On the study of a probability distribution for precipitation totals. *Journal of Applied Meteorology and Climatology*, 20(12):1499–1505.

Paulssen, M., Temme, D., Vij, A., and Walker, J. L. (2014). Values, attitudes and travel behavior: a hierarchical latent variable mixed logit model of travel mode choice. *Transportation*, 41(4):873–888.

Pedro, I. M., Pereira, L. N., and Carrasqueira, H. B. (2018). Determinants for the commitment relationship maintenance between the alumni and the alma mater. *Journal of Marketing for Higher Education*, 28(1):128–152.

Ping Jr, R. A. (2004). On assuring valid measures for theoretical models using survey data. *Journal of Business Research*, 57(2):125–141.

Pinjari, A., Eluru, N., Bhat, C., Pendyala, R., and Spissu, E. (2008). Joint model of choice of residential neighborhood and bicycle ownership: accounting for self-selection and unobserved heterogeneity. *Transportation Research Record: Journal of the Transportation Research Board*, (2082):17–26.

Piquero, A. R., MacIntosh, R., and Hickman, M. (2000). Does self-control affect survey response? applying exploratory, confirmatory, and item response theory analysis to grasmick et al.'s self-control scale. *Criminology*, 38(3):897–930.

PSRC (2018). Spring 2017 Household Travel Survey. `https://www.psrc.org/household-travel-survey-program`. [Online].

Quigley Jr., J., Bingham Jr., F., and Murray, K. (2002). An analysis of the impact of acknowledgement programs on alumni giving. *Journal of Marketing Theory and Practice*, 10(3):75–86.

Rasch, G. (1961). On general laws and the meaning of measurement in psychology. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, volume 4, pages 321–333.

Rasch, G. (1993). *Probabilistic models for some intelligence and attainment tests.* ERIC.

Reibstein, D. (2002). What attracts customers to online stores, and what keeps them coming back? *Journal of the Academy of Marketing Science*, 30(4):465–473.

Sargeant, A. (2013). *Donor Retention: What Do We Know & What Can We Do about It?* Nonprofit Quarterly.

Schröder, N., Falke, A., Hruschka, H., and Reutterer, T. (2019). Analyzing the browsing basket: A latent interests-based segmentation tool. *Journal of Interactive Marketing*, 47:181–197.

Schwab, D. P., Heneman III, H., and DeCotiis, T. A. (1975). Behaviorally anchored rating scales: A review of the literature. In *Academy of Management Proceedings*, volume 1975, pages 222–224. Academy of Management Briarcliff Manor, NY 10510.

Simonson, I. (1990). The effect of purchase quantity and timing on variety-seeking behavior. *Journal of Marketing Research*, 27(2):150–162.

Smith, A. B., Rush, R., Fallowfield, L. J., Velikova, G., and Sharpe, M. (2008). Rasch fit statistics and sample size considerations for polytomous data. *BMC Medical Research Methodology*, 8(1):33.

Smyth, G. K. and Jørgensen, B. (2002). Fitting tweedie's compound poisson model to insurance claims data: dispersion modelling. *ASTIN Bulletin*, 32(1):143–157.

Spears, S., Boarnet, M. G., Handy, S., and Rodier, C. (2014). Impacts of land-use mix on passenger vehicle use and greenhouse gas emissions. *Policy*, 9:30.

Sun, X., Hoffman, S. C., and Grady, M. L. (2007). A multivariate causal model of alumni giving: Implications for alumni fundraisers. *International Journal of Educational Advancement*, 7(4):307–332.

Teichert, T., Shehu, E., and von Wartburg, I. (2008). Customer segmentation revisited: The case of the airline industry. *Transportation Research Part A: Policy and Practice*, 42(1):227–242.

Terry, N. and Macy, A. (2007). Determinants of alumni giving rates. *Journal of Economics and Economic Education Research*, 8(3):3–17.

Turner, S. E., Meserve, L. A., and Bowen, W. G. (2001). Winning and giving: Football results and alumni giving at selective private colleges and universities. *Social Science Quarterly*, 82(4):812–826.

U.S. Department of Transportation, Federal Highway Administration (2018). 2017 National Household Travel Survey. `http://nhts.ornl.gov`. [Online].

Vafainia, S., Breugelmans, E., and Bijmolt, T. (2019). Calling customers to take action: The impact of incentive and customer characteristics on direct mailing effectiveness. *Journal of Interactive Marketing*, 45:62–80.

Van Rosmalen, J., Van Herk, H., and Groenen, P. J. (2010). Identifying response styles: A latent-class bilinear multinomial logit model. *Journal of Marketing Research*, 47(1):157–172.

Vermunt, J. K. and Magidson, J. (2013). Technical guide for latent gold 5.0: Basic, advanced, and syntax. *Statistical Innovations Inc., Belmont, MA*.

Washington, S. P., Karlaftis, M. G., and Mannering, F. L. (2010). *Statistical and econometric methods for transportation data analysis*. CRC press.

Wedel, M. and Kamakura, W. (2000). *Market segmentation: Conceptual and methodological foundations, Second Edition*. Kluwer.

Wedel, M. and Kamakura, W. A. (2012). *Market segmentation: Conceptual and methodological foundations*, volume 8. Springer Science & Business Media.

Weerts, D. J. and Ronca, J. M. (2007). Profiles of supportive alumni: Donors, volunteers, and those who "do it all". *International Journal of Educational Advancement*, 7(1):20–34.

Weerts, D. J. and Ronca, J. M. (2009). Using classification trees to predict alumni giving for higher education. *Education Economics*, 17(1):95–122.

Wright, B. D. and Masters, G. N. (1982). *Rating scale analysis.* MESA Press.

Wunnava, P. V. and Lauze, M. A. (2001). Alumni giving at a small liberal arts college: Evidence from consistent and occasional donors. *Economics of Education Review*, 20(6):533–543.

Yang, S., Zhao, Y., and Dhar, R. (2010). Modeling the underreporting bias in panel survey data. *Marketing Science*, 29(3):525–539.

APPENDIX A

Data Imputation

Two commonly used data imputation methods are Single Imputation and Multivariate Imputation. The main difference is that, Single Imputation generates one complete dataset with filled missing values, while Multivariate Imputation generates multiple compete datasets, analyzes each dataset, and combine the analysis results together. Here we simply use a Single Imputation.

First we define the notations. Let $Y_j, j = 1, 2, ..., p$ be p variables in our dataset, with some missing values in it, and let Y_j^{obs} be the observed values in Y_j, and Y_j^{mis} be the missing values. Then Y_{-j} stands for the data of $(Y_1, ..., Y_{j-1}, Y_{j+1}, ..., Y_p)$, i.e. the whole data except variable Y_j.

The nature of a Single Imputation is a Gibbs Sampler. The general framework is, from the first to the p-th (last) variable, to do the sampling as below:

$$\theta_j^* \sim P(\theta_j | Y_j^{obs}, Y_{-j}) \tag{A.1}$$

$$Y_j^* \sim P(Y_j | Y_j^{obs}, Y_{-j}) \tag{A.2}$$

To further explain the sampling above, first we estimate the distribution of parameters θ_j (Equation A.1) which describes relationship between Y_j and Y_{-j}, and then sample a value of θ_j^* from the distribution. Then with the sampled θ_j^*, for the missing Y_j values, we predict the distribution of the corresponding missing Y_j value, and sample a Y_j^* (Equation A.2) as the filling value. After we do this for each Y_j variable, the data is then complete.

The sampling techniques that we use is Predictive Mean Matching (PMM). Below we describe the specific sampling process and explain how it corresponds to the general framework above. The specific sampling process for each variable Y_j is as below:

- Fit a linear regression of Y_j on Y_{-j} (i.e., Y_j as the response and Y_{-j} as the predictors). From the linear regression model we obtain the regression coefficients $\boldsymbol{\beta}$ and the posterior predictive distribution of $\boldsymbol{\beta}$, typically a multivariate normal distribution (corresponds to obtaining $P(\theta_j|Y_j^{obs}, Y_{-j})$).
- With the posterior predictive distribution, sample θ_j^* from the distribution (corresponds to sampling in Equation A.1).
- Use the sampled regression coefficients θ_j^* and Y_{-j} to calculate predicted values for all Y_j, including the observed and missing ones.
- For each missing value in Y_j, identify the set of observed Y_j whose *predicted* values are close to the *predicted* value of the missing ones (the set of observed Y_j corresponds to obtaining $P(Y_j|Y_j^{obs}, Y_{-j})$). Here we use a set size of 5.
- For each missing value in Y_j, within its set of close samples, randomly choose one and assign the *observed* Y_j to the missing Y_j (corresponds to Equation A.2). Then the imputation of Y_j is completed.

APPENDIX B

Parameter Estimation of Mixture Regression Model by EM Algorithm

As discussed in Section 2.4, the regression coefficients, segment proportions and segment memberships of individuals are estimated by EM algorithm. Based on Equations 2.6, we already have the probability distribution of an individual observation, and thus we can calculate the log-likelihood of the data and use Maximum Likelihood Estimation to estimate parameters. However, in mixture models, we need to incorporate the segment information. Specifically, we calculate the Complete Likelihood of:

(B.1)

$$L_c(\boldsymbol{\pi}, \boldsymbol{\Theta}; \boldsymbol{y}, \boldsymbol{x}, \boldsymbol{z}) = \prod_{i=1}^{N} \pi_{z_i} f_{z_i}(\boldsymbol{y}_i | \boldsymbol{x}_i, \boldsymbol{\theta}_{z_i}) = \prod_{i=1}^{N} \pi_{z_i} f_{freq,z_i}(y_{f,i} | \boldsymbol{x}_{f,i}, \boldsymbol{\beta}_{z_i}) f_{amt,z_i}(y^l_{a,i} | \boldsymbol{x}_{a,i}, \boldsymbol{\gamma}_{z_i})$$

Then we take log for both sides, and we get the Complete log-Likelihood:

(B.2)

$$\ln L_c(\boldsymbol{\pi}, \boldsymbol{\Theta}; \boldsymbol{y}, \boldsymbol{x}, \boldsymbol{z}) = \sum_{i=1}^{N} [\ln(\pi_{z_i}) + \ln(f_{freq,z_i}(y_{f,i} | \boldsymbol{x}_{f,i}, \boldsymbol{\beta}_{z_i})) + \ln(f_{amt,z_i}(y^l_{a,i} | \boldsymbol{x}_{a,i}, \boldsymbol{\gamma}_{z_i}))]$$

Then we apply EM algorithm to maximize the Complete log-Likelihood. The EM algorithm consists of two steps: E-step where we calculate expectation, and M-step where we do maximization. In E-step, we keep the parameters fixed, and calculate the expectation of latent variable $\boldsymbol{z}$, which is given by:

$$E_{z_i|\boldsymbol{\Theta}}[z_i] = P(z_i = k|\boldsymbol{y}_i, \boldsymbol{x}_i, \boldsymbol{\pi}, \boldsymbol{\Theta}) = p_{ik} \tag{B.3}$$

which is the posterior probability given by Equation 2.8. With the posterior posterior probabilities updated, we can then update the segment proportions.

In M-step, we calculate the Q-function in EM algorithm, which is the expectation of Complete log-Likelihood function:

$$Q(\boldsymbol{\pi}, \boldsymbol{\Theta}) = E_{\boldsymbol{z}|\boldsymbol{\Theta}}[\ln L_c(\boldsymbol{\pi}, \boldsymbol{\Theta}; \boldsymbol{y}, \boldsymbol{x}, \boldsymbol{z})] \tag{B.4}$$

Then we keep the posterior probabilities fixed and update the parameters $\boldsymbol{\Theta} = \{\beta_1, ..., \beta_K; \gamma_1, ..., \gamma_K\}$ by maximizing the Q-function. The algorithm is shown below in Algorithm 1:

Once the stopping criterion is met (either maximum number of iterations reached or increase of Q-function is smaller than threshold), the last $\hat{\boldsymbol{\pi}}^t$ and $\hat{\boldsymbol{\Theta}}^t$ are the estimated parameters. Then by applying Equation 2.8, we can calculate the posterior probabilities p_{ik} of each individual i belonging to segment k. The final segment assignment can be done by assigning individual i to segment k with largest p_{ik}.

Algorithm 1 EM Algorithm for Mixture Regression Model

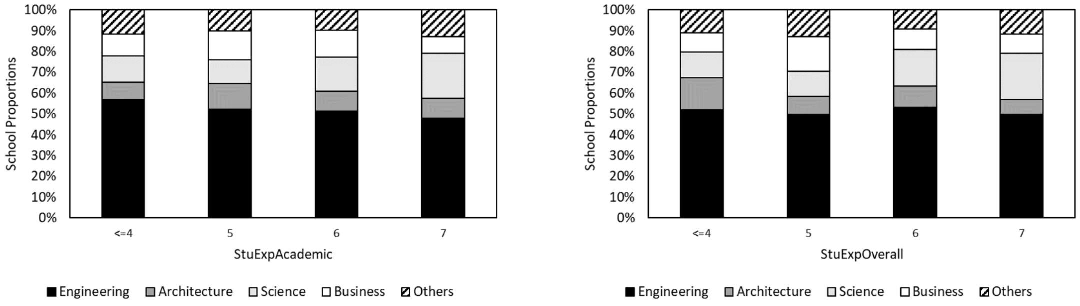
```
Initialize
$t \leftarrow 0, \hat{\boldsymbol{\pi}}^t, \hat{\boldsymbol{\Theta}}^t$
while $t < T$ and $Q(\hat{\boldsymbol{\pi}}^t, \hat{\boldsymbol{\Theta}}^t) - Q(\hat{\boldsymbol{\pi}}^{t-1}, \hat{\boldsymbol{\Theta}}^{t-1}) > \epsilon$ do
    E-Step:
    for $k = 1, ..., K$ do
        $\hat{p}_{ik} = P(z_i = k|\boldsymbol{y}_i, \boldsymbol{x}_i, \hat{\boldsymbol{\pi}}^t, \hat{\boldsymbol{\Theta}}^t) = \frac{\hat{\pi}_k^t f_k(\boldsymbol{y}_i|\boldsymbol{x}_i, \hat{\boldsymbol{\theta}}_k^t)}{\sum_{j=1}^K \hat{\pi}_j^t f_j(\boldsymbol{y}_i|\boldsymbol{x}_i, \hat{\boldsymbol{\theta}}_j^t)}$
        $\hat{\pi}_k^{t+1} = \frac{1}{N} \sum_{i=1}^N \hat{p}_{ik}$
    end for
    M-Step:
    $\hat{\boldsymbol{\Theta}}^{t+1} \leftarrow \arg\max_{\boldsymbol{\Theta}} Q(\hat{\boldsymbol{\pi}}^{t+1}, \boldsymbol{\Theta})$

    $t \leftarrow t + 1$
end while
```

APPENDIX C

Effect of student experience variables

From Table 2.8 we see that the StuExpOverall (overall student experience) is significantly positive in Frequency submodel, and StuExpAcademic (overall academic experience) has significantly negative coefficients. We also note that StuExpLife has a significant negative effect on average donation amounts. The signs of these effects are not consistent with intuition that better experiences as a student lead to positive outcomes. Correlation between the student experience variables with Field of Study may provide a possible explanation. Figure C.1 shows the distributions of Field of Study under different Student Experience question scores, i.e., among the alumni that gave a certain answer score to the question, how many of them are from each of the schools. Both StuExpAcademic and StuExpOverall are included, and since not many people gave a score under 4 for both questions, we aggregate score of 1-4 in to the "≤4" category.



(a) StuExpAcademic

(b) StuExpOverall

Figure C.1. School Type Distribution under Different Student Experience Question Scores

According to Figure C.1a, there is a trend in StuExpAcademic that as the score gets higher, the proportion of School of Engineering alumni decreases from $\approx 57\%$ to $\approx 48\%$. Taking the overall effect that Engineering alumni donate more frequently than other alumni from other schools into consideration, we infer that Engineering alumni tend to give low scores to the student academic experience question. This results in the marginal effect that worse student academic experience leads to more donations. On the other hand, from Figure C.1b, the Field of Study distribution under different overall student experience answer scores turns out to be more stable. The proportion of School of Engineering ranges from $\approx 50\%$ to $\approx 53\%$. This explains why the overall student experience question (StuExpOverall) still shows a positive marginal effect on donation frequency.

APPENDIX D

Segmentation of behavior assignment by Markov Chain Mixture model

We applied the Markov chain mixture model for donor segmentation developed by Durango-Cohen et al. (2013a) on our data to obtain the segmentation of longitudinal alumni donation behavior. To introduce the methodology, we first define the longitudinal data sequence of alumnus m as $\mathbf{y} = \{\mathbf{y}_m\}$, $m = 1, ..., M$. For each sequence $\mathbf{y}_m = \{y_m^1, y_m^2, ..., y_m^T\}$, where y_m^t is the annual donation amount of alumnus m in year t. M and T represent the total number of individuals in the population and total number of years considered. In the specific case of this study, $T = 17$ since we consider alumni donation from 2000 to 2016. We firstly present the general framework of mixture model and Expectation-Maximization (EM) algorithm to estimate the mixture model parameters. Then we present the specific Markov chain mixture model for the segmentation of sequence alumni donation data.

D.1. Segmentation by Mixture model

In mixture model framework, the sequences $\mathbf{y}$ comes from a population being a mixture of S segments with segment proportions $\lambda = \{\lambda_s\}$, $s = 1, ..., S$, and constraints $\sum_s = 1^S \lambda_s = 1$ and $\lambda_s \geq 0$. Then for a certain sequence $\mathbf{y}_m$, the probability density is given by:

$$f(\mathbf{y}_m|\mathbf{\Theta}) = \sum_{s=1}^{S} \lambda_s f_s(\mathbf{y}_m|\boldsymbol{\theta}_s) \tag{D.1}$$

where $\mathbf{\Theta} = \{\boldsymbol{\theta}_s\}$ are the sets of parameters that define the function $f_s(\cdot)$. Namely, the total probability is a weighted sum of the probabilities associated with each segment. For each given $\mathbf{y}_m$, Bayes law can be applied to calculate probability of individual m belonging to segment s by:

$$p_{ms} \equiv P(z_{ms} = 1|\mathbf{y}_m) = \frac{\lambda_s f_s(\mathbf{y}_m|\boldsymbol{\theta}_s)}{\sum_{r=1}^{S} \lambda_r f_r(\mathbf{y}_m|\boldsymbol{\theta}_r)} \tag{D.2}$$

where z_{ms} is the hard assignment of $z_{ms} = 1$ if individual m belongs to segment s and $z_{ms} = 0$ otherwise. Given the probability distribution of an individual and incorporate the segment proportion, the Complete Likelihood of mixture model can be written as:

$$L_c(\mathbf{y}, \mathbf{z}; \boldsymbol{\lambda}, \mathbf{\Theta}) = \prod_{m=1}^{M} \prod_{s=1}^{S} [f_s(\mathbf{y}_m|\boldsymbol{\theta}_s) P(z_{ms} = 1|\mathbf{y}_m, \boldsymbol{\lambda}, \mathbf{\Theta})]^{p_{ms}} \tag{D.3}$$

and the log of Complete Likelihood can be obtained as:

$$\ln L_c(\mathbf{y}, \mathbf{z}; \boldsymbol{\lambda}, \mathbf{\Theta}) = \sum_{m=1}^{M} \sum_{s=1}^{S} [z_{ms} \ln f_s(\mathbf{y}_m|\boldsymbol{\theta}_s) + z_{ms} \ln p_{ms}] \tag{D.4}$$

As discussed by McLachlan and Peel (2004), the issue to estimate the parameters $\mathbf{\Theta}$ and λ is that the individual membership is unknown, and thus we cannot use simple

Maximum Likelihood Estimation to estimate the parameters. Therefore, the mixture model is considered as an incomplete data problem, and EM algorithm (Dempster et al., 1977) is used to iteratively update parameters and reach convergence. The EM algorithm consists of two steps – E-step and M-step. In E-step, we evaluate the expectation of $\ln L_c(\cdot)$ described by Equation D.4 and estimates $\boldsymbol{\lambda}$ and p_{ms}; in M-step, we maximize the expectation of $\ln L_c(\cdot)$ to update $\boldsymbol{\Theta}$. The E-step and M-step are repeated until convergence criterion is met. The general EM algorithm is shown below in Algorithm 2:

Algorithm 2 A General EM Algorithm Framework

```
**Initialize**
$t \leftarrow 0, \hat{\boldsymbol{\lambda}}^t, \hat{\boldsymbol{\Theta}}^t$
**while** $t < T$ and $Q(\hat{\boldsymbol{\lambda}}^t, \hat{\boldsymbol{\Theta}}^t) - Q(\hat{\boldsymbol{\lambda}}^{t-1}, \hat{\boldsymbol{\Theta}}^{t-1}) > \epsilon$ **do**
    E-Step:
    **for** $s = 1, ..., S$ **do**
        $\hat{p}_{ms} = P(z_{ms} = 1|\boldsymbol{y}_m, \hat{\boldsymbol{\lambda}}^t, \hat{\boldsymbol{\Theta}}^t) = \frac{\hat{\lambda}_s^t f_s(\boldsymbol{y}_m|\hat{\boldsymbol{\theta}}_s^t)}{\sum_{r=1}^S \hat{\lambda}_r^t f_r(\boldsymbol{y}_m|\hat{\boldsymbol{\theta}}_r^t)}$
        $\hat{\lambda}_s^{t+1} = \frac{1}{M}\sum_{m=1}^M \hat{p}_{ms}$
    **end for**
    M-Step:
    $\hat{\boldsymbol{\Theta}}^{t+1} \leftarrow \arg\max_{\boldsymbol{\Theta}} Q(\hat{\boldsymbol{\lambda}}^{t+1}, \boldsymbol{\Theta})$
    $t \leftarrow t + 1$
**end while**
```

D.2. Markov Chain Mixture model for alumni donation

Here in this section, we first present the specific Markov chain model for alumni donation sequence. We model each sequence as a finite set of states, and each segment ...

The notations and assumptions are as below:

- We transform the value of annual donation of an individual to a manifestation of a set of L discrete states. Let x_m^t be the state/level representation of the individual annual donation amount y_m^t. Such relationship between x_m^t and y_m^t is defined by

$x_m^t = l \iff b_{l-1} < y_m^t \leq b_l$, $l = 1, ..., L$, where the points $b_0, ..., b_L$ define a partition of the state/level space of y_m^t. In this study, $L = 6$ is taken, and specific values of $b_0, ..., b_L$ used in this study are listed in Table D.1. For the state/level variable, we have $\mathbf{x}_m \equiv \{x_m^t\}$, $t = 1, ..., T$, and $\mathbf{x} \equiv \{\mathbf{x}_m\}$, $m = 1, ..., M$.

- The probability density function $f_s(\cdot)$ is given by:

$$f_s(\mathbf{y}_m|\boldsymbol{\theta}_s) = P(\mathbf{x}_m|\boldsymbol{\theta}_s) = \prod_{t=1}^{T} \pi^s_{x_m^{t-1}, x_m^t} = \prod_{i=1}^{L}\prod_{j=1}^{L} \left[\pi_{ij}^s\right]^{\nu_{ij}^m} \tag{D.5}$$

where π_{ij}^s is the probability that annual donation amount of an individual from segment s transitions from state i to state j in two consecutive years. The parameters $\boldsymbol{\theta}_s \equiv \pi_{ij}^s$, $i, j = 1, ..., L$ of segment s correspond to the set of transition probabilities that define a time-homogeneous Markov chain. ν_{ij}^m is the total number of "from i to j" transitions in sequence $\mathbf{x}_m$, which is another form of representation of $\mathbf{x}_m$.

- Based on the probability density function $f_s(\cdot)$, the log Complete Likelihood can be obtained as:

$$\ln L_c(\mathbf{y}, \mathbf{z}; \boldsymbol{\lambda}, \boldsymbol{\Theta}) = \sum_{m=1}^{M}\sum_{s=1}^{S} z_{ms}\left[\sum_{i=1}^{L}\sum_{j=1}^{L} \nu_{ij}^m \ln\left(\pi_{ij}^s\right)\right] + z_{ms}\ln p_{ms} \tag{D.6}$$

Then the log Complete Likelihood can be used in Appendix 2 to estimate the parameters.

As mentioned in Section 3.4.1, we obtain a 3-segment MCMM for donation behavior assignment. The transition probability matrix $\Pi^s = \{\pi_{ij}^s\}$ of each segment is presented below in Table D.2, where rows correspond to state i and columns correspond to state j.

Table D.1. State Definitions for Markov Chain Mixture Model

State Definition	0	1	2	3
Annual Donation Amount	Inactive	(\$0,\$25]	(\$25,\$50]	(\$50,\$100]
State Definition	4	5	6	7
Annual Donation Amount	(\$100,\$275]	(\$275,\$600]	(\$600,\$1500)	(\$1500,∞)

The probability transition matrices demonstrate characteristics of the 3 segments, that Low Variance segment has high probabilities of keeping current state along the diagonal line, and Transient segment has high probabilities going into inactive (state 0) from any other states.

Table D.2. Transition probability matrices of 3-segment MCMM

State	0	1	2	3	4	5	6	7
High Variance								
0	44.4%	11.1%	19.6%	3.7%	10.8%	5.5%	3.3%	1.6%
1	32.5%	38.3%	21.2%	2.7%	3.2%	0.9%	0.2%	0.0%
2	29.9%	11.3%	39.3%	5.8%	10.1%	2.0%	0.5%	0.1%
3	29.2%	8.1%	19.8%	21.2%	14.5%	5.7%	1.1%	0.3%
4	30.7%	2.8%	21.3%	5.4%	27.8%	9.5%	2.1%	0.4%
5	26.3%	1.3%	8.5%	3.8%	18.9%	28.8%	10.2%	2.2%
6	24.2%	0.3%	2.2%	0.0%	2.5%	11.2%	45.4%	14.1%
7	19.6%	0.3%	0.5%	0.3%	0.4%	3.4%	27.4%	48.2%
Low Variance								
0	74.3%	9.0%	7.3%	2.2%	3.3%	1.7%	1.3%	0.8%
1	18.3%	66.6%	12.6%	1.9%	0.6%	0.0%	0.0%	0.0%
2	10.6%	7.2%	67.5%	5.6%	7.9%	0.9%	0.3%	0.0%
3	10.7%	6.0%	16.6%	43.3%	17.4%	5.8%	0.3%	0.0%
4	8.2%	0.3%	13.3%	4.5%	56.6%	15.6%	1.2%	0.3%
5	5.8%	0.2%	2.2%	2.6%	11.6%	67.2%	9.9%	0.5%
6	6.9%	0.0%	0.2%	0.4%	1.0%	8.8%	79.5%	3.3%
7	8.6%	0.0%	0.5%	0.4%	1.0%	3.0%	8.5%	78.0%
Transient								
0	94.3%	2.3%	1.9%	0.3%	0.6%	0.3%	0.2%	0.1%
1	76.4%	14.2%	6.9%	0.7%	1.0%	0.5%	0.3%	0.0%
2	72.1%	7.5%	12.1%	2.4%	3.4%	1.7%	0.6%	0.2%
3	71.9%	5.6%	11.8%	4.8%	3.3%	1.9%	0.6%	0.2%
4	62.1%	2.6%	13.2%	1.5%	12.8%	5.0%	1.6%	1.1%
5	59.8%	1.3%	6.0%	4.1%	7.4%	15.3%	4.8%	1.5%
6	68.3%	0.0%	2.7%	0.9%	5.9%	5.4%	11.2%	5.6%
7	62.9%	0.0%	1.5%	0.0%	3.4%	1.2%	2.9%	28.1%

APPENDIX E

Survey question details

- Connected: How connected do you feel to the Illinois Institute of Technology? (From “1 – Not very connected” to “7 – Very connected”)
- FeelingsUniv: How would you describe your feelings toward the university today? (From “1 – Very cold” to “7 – Very warm”)
- FeelingsAA: How would you describe your feelings toward the university Alumni Association? (From “1 – Very cold” to “7 – Very warm”)
- Competent: How competent do you perceive the university to be? (From “1 – Very incompetent” to “7 – Very competent”)
- Similarity: Please rate the similarities between you and the university. (From “1 – Not very similar” to “7 – Very similar”)
- UseWordUs: To what extent do you use the word “us” to describe you and the university community? (From “1 – Never” to “7 – Always”)
- DonationImpact[1]:

  For donors: How impactful do you feel your donation(s) has/have been? (From “1 – Not impactful at all” to “7 – Very impactful”)

  For non-donors: What do you think the impact of a donation from you would be? (From “1 – Not impactful at all” to “7 – Very impactful”)

[1] This variable is obtained by combining the responses of two questions below, for donors and non-donors respectively. Namely, we take responses of first question for donors, and second question for non-donors.

- StuExpAcademic: How would you assess your experience as a student? Overall academic experience. (From “1 – Completely dissatisfied” to “7 – Completely satisfied”)
- StuExpAcademicDept: How would you assess your experience as a student? Overall academic experience in your department or program. (From “1 – Completely dissatisfied” to “7 – Completely satisfied”)
- StuExpLife: How would you assess your experience as a student? Non-academic or student life experience. (From “1 – Completely dissatisfied” to “7 – Completely satisfied”)
- StuExpOverall: How would you assess your experience as a student? Overall experience at the university. (From “1 – Completely dissatisfied” to “7 – Completely satisfied”)
- UnivOweSuccess: Please rate your level of agreement with the statement – I owe a portion of my career success to the university. (From “1 – Strongly disagree” to “7 – Strongly agree”)
- UnivRecommend: Please rate your level of agreement with the statement – I would recommend the university to friends or family. (From “1 – Strongly disagree” to “7 – Strongly agree”)
- ContactSatisfied: How satisfied are you with the amount of contact from the university? (From “-3 – Far too little” to “3 – Far too much”)

APPENDIX F

Data and segmentation results of other cities

In this appendix, we present the details of the 2010/2011 RHTS data of New York, Spring 2017 Household Travel Survey data of Seattle, and 2017 NHTS national data, and also the segmentation results of the 3 datasets with both generalized VMT vector and conventional POV/Taxi only VMT. The results validates the effectiveness of generalized VMT vector segmentation, which is the model that we propose in this paper.

F.1. 2010/2011 RHTS data of New York

The 2010/2011 RHTS data includes totally 18,965 households within 28 counties in of the New York-New Jersey-Connecticut metropolitan area (NYMTC, 2011). In order to do similar segmentation as CMAP's data, we take SED variables of household size, vehicle ownership, number of students, number of workers, number of licenses, and household income into consideration. Also, VMT for POV/Taxi and Other Mode are used for segmentation.

For this dataset, we follow the same procedure to process the raw data as the CMAP's data. Only commuting-related trips are included according to the trip purpose information. Both trip origin purpose and destination purpose are available in the dataset. Here we include a trip if either the origin or destination purpose is going to work, university, or school (i.e. trips going to or coming back from work/university/school are both considered). The trip distance information is already available in the raw data, which is also

Table F.1. Summary of SED information of New York data

	Min	Mean	Max	St. deviation
Income ($ in thousands)	15.00	97.60	200.00	56.07
Household size	1	2.61	10	1.28
Number of workers	0	1.55	7	0.73
Number of students	0	0.75	7	1.00
Number of vehicles	0	1.72	8	1.17
Number of licenses	0	1.78	7	0.83

Table F.2. Summary of VMT of New York data

	Min	Mean	Max	Number of 0	St. deviation
VMT for POV/Taxi	0.00	22.28	390.41	3232	30.66
VMT for Other Mode	0.00	3.67	2585.37	5764	54.14

measured by Euclidean Distance in miles. For trip mode, we categorize 3 types of trip mode – “Auto (Car or Small Truck) Driver”, “Auto (Car or Small Truck) Passenger”, and “Taxi (Yellow, Medallion Cab)” – as the POV/Taxi Mode, and others as the Other Mode, which includes walking, biking, local bus, subway, school bus, etc. Besides, households with unknown income information are removed. Households with 0 VMT for both POV/Taxi and Other Mode are also removed. The categorical income interval variable is transformed into numeric variable by taking the mean value of the interval. This also eliminates the impact of outliers. The final processed data includes 12,278 households. The summary of SED information and VMT data of the final processed data are shown in Table F.1 and Table F.2.

To determine the number of clusters for segmentation, similarly we use CAIC and Classification Error to asses the models. In this case, number of clusters of $S = 4$ still works best for the 2010/2011 RHTS data. Based on this parameter, segmentation models

Table F.3. Clusters profiles with both POV/Taxi and Other Mode of New York data

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Total
Cluster proportion	0.48	0.03	0.24	0.26	1.00
VMT for POV/Taxi	31.22	62.65	0.00	22.12	22.28
VMT for Other Mode	0.00	92.26	1.11	3.51	3.67
Income (in thousands $)	98.53	115.29	80.72	109.62	97.60
Household size	2.33	3.68	2.16	3.45	2.61
Number of workers	1.57	1.92	1.26	1.75	1.55
Number of students	0.44	1.61	0.53	1.43	0.75
Number of vehicles	2.13	2.15	0.69	1.88	1.72
Number of licenses	1.92	2.14	1.22	1.98	1.78

Table F.4. Clusters profiles with POV/Taxi exclusively of New York data

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Total
Cluster proportion	0.26	0.23	0.48	0.03	1.00
VMT for POV/Taxi	0.00	52.54	12.65	134.87	22.28
VMT for Other Mode	4.91	2.14	3.82	1.81	3.67
Income (in thousands $)	81.82	113.69	96.87	123.46	97.60
Household size	2.25	2.87	2.65	3.19	2.61
Number of workers	1.26	1.85	1.53	2.15	1.55
Number of students	0.60	0.83	0.77	0.98	0.75
Number of vehicles	0.73	2.37	1.90	2.74	1.72
Number of licenses	1.24	2.13	1.86	2.49	1.78

with VMT for both POV/Taxi and Other Mode, and with VMT for POV/Taxi only are developed. The cluster profiles for both cases are shown in Table F.3 and Table F.4.

According to Table F.3, the households are segmented into 4 clusters that are similar with the CMAP's data of Chicago. Cluster 1 represents POV/Taxi dominated households, while Cluster 2 represents Other Mode dominated households. Cluster 2 and 4 show mixture use of both modes, where Cluster 2 shows longer travel distance. Similar with the result of CMAP's data, Cluster 3 has lowest average household income, while Cluster

Table F.5. Silhouette coefficients of POV/Taxi & Other Mode model and POV/Taxi exclusively model of New York data

	POV/Taxi and Other Mode	POV/Taxi exclusively
VMT for POV/Taxi	1.000	1.000
VMT for Other Mode	0.997	0.126
Income	-0.015	-0.047
Household size	-0.065	0.040
Number of workers	0.109	0.056
Number of students	0.072	0.123
Number of vehicles	0.352	0.297
Number of licenses	0.042	0.004

2 and 4 is characterized by their large household size and especially large number of students in household.

However, when looking at Table F.4, the conventional segmentation doesn't partition households into representative clusters as the generalized vector segmentation does. Table F.5 shows the silhouette coefficients from the two segmentation models. We see that beside household size and number of students, the generalized VMT vector segmentation has larger silhouette coefficients than the conventional segmentation, especially on VMT for Other Mode, number of workers and number of licenses. Namely, the conventional segmentation, as discussed in Section 4.5, doesn't reflect the effect of household structure on travel behavior. Therefore, the 2010/2011 RHTS data of New York well validates the effectiveness and advantage of the generalized VMT vector segmentation.

F.2. Spring 2017 Household Travel Survey data of Seattle

The Spring 2017 Household Travel Survey data of Seattle includes totally 3,285 households in the entire PSRC four-county region, which includes King, Kitsap, Pierce, and Snohomish counties (PSRC, 2018). A difference here in the data is that information of number of students and number of licenses in household is not available. We use number of children in household as a substitute of number of students, which reflect similar information. Therefore, the SED variables in the Spring 2017 Household Travel Survey data are household size, vehicle ownership, number of children, number of workers, and household income. The VMT information is the same with Chicago and New York data, consisting of VMT for POV/Taxi and Other Mode.

Similar data processing procedures are done for this dataset. As the 2010/2011 RHTS data, origin and destination trip purpose information is available in the Spring 2017 Household Travel Survey data. We include a trip if either the origin or destination purpose is “School/daycare”, “Primary workplace”, “Work-related place”, or “Other work-related activity”. Thus, trips going to or coming back from work/school are both considered. The trip distance information, measured by Euclidean Distance in miles, is available in the raw data. For trip mode, we categorize 3 types of trip mode — “Household vehicle”, “Taxi (e.g., Yellow Cab”, and “Other hired service (e.g., Lyft, Uber)” – as the POV/Taxi Mode, and others as the Other Mode, which includes walking, biking, public transit, school bus, urban rail, etc. Households with unknown income information and households with 0 VMT for both POV/Taxi and Other Mode are also removed. The categorical income interval variable is also transformed into interval mean numeric variable. The final processed data includes 2,255 households. The summary of SED information and VMT data of the final processed data are shown in Table F.6 and Table F.7.

Based on CAIC and Classification Error, we develop segmentation models with number of clusters $S = 4$, which is the best parameter for Spring 2017 Household Travel

Table F.6. Summary of SED information of Seattle data

	Min	Mean	Max	St. deviation
Income ($ in thousands)	10.00	105.47	250.00	64.93
Household size	1	2.00	9	1.02
Number of workers	0	1.42	5	0.62
Number of children	0	0.31	5	0.68
Number of vehicles	0	1.33	8	0.88

Table F.7. Summary of VMT of Seattle data

	Min	Mean	Max	Number of 0	St. deviation
VMT for POV/Taxi	0.00	29.15	521.12	697	50.23
VMT for Other Mode	0.00	37.48	7217.85	783	261.00

Table F.8. Clusters profiles with both POV/Taxi and Other Mode of Seattle data

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Total
Cluster proportion	0.35	0.27	0.08	0.30	1.00
VMT for POV/Taxi	34.85	27.76	124.16	0.00	29.15
VMT for Other Mode	0.00	13.54	361.68	19.84	37.48
Income (in thousands $)	104.84	116.14	123.24	92.19	105.47
Household size	1.97	2.41	2.20	1.62	2.00
Number of workers	1.37	1.61	1.71	1.24	1.42
Number of children	0.28	0.54	0.35	0.12	0.31
Number of vehicles	1.65	1.49	1.69	0.73	1.33

Survey data. The cluster profiles for both conventional POV/Taxi only segmentation and generalized VMT vector segmentation are shown in Table F.8 and Table F.9.

In the Spring 2017 Household Travel Survey data, the 4 clusters have similar characterization as the other datasets. Cluster 1 and 4 represent POV/Taxi dominated and Other Mode dominated, respectively. Cluster 2 and 3 represents mixture use of both mode with shorter and longer distance, respectively. Similarly, Cluster 2 and 3 have larger household

Table F.9. Clusters profiles with POV/Taxi exclusively of Seattle data

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Total
Cluster proportion	0.31	0.19	0.45	0.05	1.00
VMT for POV/Taxi	0.00	65.24	15.77	206.55	29.15
VMT for Other Mode	22.54	50.34	30.17	156.60	37.48
Income (in thousands $)	92.68	112.47	110.11	116.60	105.47
Household size	1.63	2.32	2.11	2.10	2.00
Number of workers	1.24	1.63	1.43	1.65	1.42
Number of children	0.12	0.46	0.38	0.29	0.31
Number of vehicles	0.74	1.81	1.48	1.84	1.33

Table F.10. Silhouette coefficients of POV/Taxi & Other Mode model and POV/Taxi exclusively model of Seattle data

	POV/Taxi and Other Mode	POV/Taxi exclusively
VMT for POV/Taxi	1.000	1.000
VMT for Other Mode	1.000	0.428
Income	-0.011	0.015
Household size	0.103	0.135
Number of workers	-0.010	0.026
Number of children	0.508	0.514
Number of vehicles	0.217	0.173

size and number of children, also with larger household income. Cluster 4 has the lowest income and household size, which is consistent with the other datasets.

The conventional POV/Taxi only segmentation shows same problem of not partitioning households into representative clusters. According to Table F.10, although SED variables show similar level of between cluster heterogeneity and within cluster homogeneity, the silhouette coefficients of VMT for Other Mode still shows large difference. The conventional segmentation still fails to partition VMT for Other Mode into representative clusters. Thus, the generalized VMT vector segmentation is also validated by the Spring 2017 Household Travel Survey data.

F.3. 2017 NHTS national data

The 2017 NHTS includes totally 129,696 households, which is “the only source of national data that allows one to analyze trends in personal and household travel” (U.S. Department of Transportation, Federal Highway Administration, 2018). In the 2017 NHTS data, information of number of students and number of licenses in household are not available. We use information of number of drivers in household (which can be considered the same with number of licenses) as substitute of number of licenses. With information of number of adults (at least 18 years old) in household available, we create a variable of number of children (less than 18 years old) in household by taking the difference between household size and number of adults in household. This variable serves as substitute of number of students in household as in Chicago data. The final SED variables in 2017 NHTS data are household size, vehicle ownership, number of children, number of workers, number of drivers, and household income. The VMT information is the same with other data, consisting of VMT for POV/Taxi and Other Mode.

Similar procedures are taken for data processing. Both origin and destination trip purpose are available for each trip. We include a trip if either origin or destination purpose is “Work”, “Work-related meeting trip”, or “Attend school as student”. Both trips going to or coming back from work/school are considered. The trips distance information is available in the raw data. For trip mode, 6 types of trip mode – “Car”, “SUV”, “Van”, “Motorcycle/Moped”, “RV (motor home, ATV, snowmobile)”, and “Taxi/limo (including Uber/Lyft)” – are categorized as POV/Taxi Mode, and others as the Other Mode, which includes walking, biking, school bus, subway, etc. Also, trips with unknown trip distance are removed. Households with unknown income information and households with 0 VMT

Table F.11. Summary of SED information of NHTS data

	Min	Mean	Max	St. deviation
Income ($ in thousands)	10.00	86.83	200.00	54.80
Household size	1	2.52	13	1.29
Number of workers	0	1.54	7	0.73
Number of children	0	0.56	8	0.97
Number of vehicles	0	2.23	12	1.20
Number of drivers	0	1.90	9	0.79

Table F.12. Summary of VMT of NHTS data

	Min	Mean	Max	Number of 0	St. deviation
VMT for POV/Taxi	0.00	31.80	5901.01	9531	70.21
VMT for Other Mode	0.00	17.40	7685.53	34856	126.45

for both POV/Taxi and Other Mode, as in other data, are removed. The income variable is also transformed from categorical interval to interval mean. The final processed data includes 59,085 households. The summary of SED information and VMT data of the final process ed data are shown in Table F.11 and Table F.12.

Based on CAIC and Classification Error, we develop segmentation models with number of clusters $S = 3$, which is different with the other datasets. The cluster profiles for both conventional POV/Taxi only segmentation and generalized VMT vector segmentation are shown in Table F.13 and Table F.14.

Since the NHTS data is from a national survey, most of the households included might not have good access to transit system as households in Chicago, New York, and Seattle do. Thus, according to the results in Table F.13, there's no cluster that represents Other Mode dominated households. Cluster 1, 2, and 3 represents POV/Taxi dominated, short distance mixture, and long distance mixture, respectively. Although without the Other

Table F.13. Clusters profiles with both POV/Taxi and Other Mode of NHTS data

	Cluster 1	Cluster 2	Cluster 3	Total
Cluster proportion	0.59	0.37	0.04	1.00
VMT for POV/Taxi	34.90	16.33	123.81	31.80
VMT for Other Mode	0.00	18.97	247.21	17.40
Income (in thousands $)	85.10	87.37	106.38	86.83
Household size	2.31	2.78	3.12	2.52
Number of workers	1.48	1.60	1.95	1.54
Number of children	0.42	0.76	0.80	0.56
Number of vehicles	2.17	2.24	2.88	2.23
Number of drivers	1.85	1.93	2.33	1.90

Table F.14. Clusters profiles with POV/Taxi exclusively of NHTS data

	Cluster 1	Cluster 2	Cluster 3	Total
Cluster proportion	0.67	0.02	0.31	1.00
VMT for POV/Taxi	10.98	296.50	63.75	31.80
VMT for Other Mode	18.26	19.60	15.37	17.40
Income (in thousands $)	80.93	106.43	98.90	86.83
Household size	2.37	3.00	2.81	2.52
Number of workers	1.42	1.98	1.79	1.54
Number of children	0.51	0.67	0.67	0.56
Number of vehicles	2.08	2.92	2.52	2.23
Number of drivers	1.78	2.36	2.13	1.90

Mode dominated cluster, the trends are still similar with other data. Cluster 2 and 3 have larger household size and number of children in household, and also have higher income than Cluster 1. Same with other data, Cluster 2 shows lower income per person than Cluster 1, though it has higher total income than Cluster 1.

The advantage of generalized VMT vector segmentation is less significant in 2017 NHTS data than other data. As discussed above, nationwide households might have less access to transit, and thus the generalized VMT vector might be less meaningful than in other city's data. However, we can still observe more variation in terms of household size and number of children, which are two important factors of structure of household, in generalized VMT vector segmentation than conventional POV/Taxi only segmentation. For NHTS data, we didn't calculate the silhouette coefficients due to computational capability issue, but we believe the results presented do show the difference. Thus, in the 2017 NHTS data, generalized VMT vector segmentation still shows better capability of partitioning households into representative clusters.

APPENDIX G

Parameter estimation of Gaussian mixture model

As discussed in Section 4.4, the distribution parameters, cluster proportions, and cluster memberships of observations are obtained by parameter estimation process. As commonly done, we can find the expression of (log)likelihood with respect to each combination of parameter and then maximize it. Given data set $\boldsymbol{Y} = \{\boldsymbol{y}_i\}_{i=1}^{I}$, the likelihood for $\boldsymbol{\Theta}$ and $\boldsymbol{\lambda}$ is:

$$L(\boldsymbol{\lambda}, \boldsymbol{\Theta}; \boldsymbol{y}) = \prod_{i=1}^{I} f(\boldsymbol{y}_i | \boldsymbol{\lambda}, \boldsymbol{\Theta}) \tag{G.1}$$

In fact, we can directly use (3) to conduct likelihood maximization as what is commonly done for regular MLE calculation. However, that procedure ignores latent variable $\boldsymbol{z}$ and the correlation between $\boldsymbol{z}$ and $\boldsymbol{\lambda}$ that $\boldsymbol{z}$ follows prior distribution characterized by $\boldsymbol{\lambda}$. In order to incorporate membership information, we can use (2) express the complete data likelihood equation:

$$L_c(\boldsymbol{\lambda}, \boldsymbol{\Theta}; \boldsymbol{y}, \boldsymbol{z}) = \prod_{i=1}^{I} f(\boldsymbol{y}_i, z_i; \boldsymbol{\lambda}, \boldsymbol{\Theta}) = \prod_{i=1}^{I} \lambda_{z_i} f_{z_i}(\boldsymbol{y}_i | \boldsymbol{\theta}_{z_i}) \tag{G.2}$$

If we take log for both sides:

$$\ln L_c(\boldsymbol{\lambda}, \boldsymbol{\Theta}; \boldsymbol{y}, \boldsymbol{z}) = \sum_{i=1}^{I} (\ln(\lambda_{z_i}) + \ln(f_{z_i}(\boldsymbol{y}_i | \boldsymbol{\theta}_{z_i}))) \tag{G.3}$$

With consideration on both computational efficiency and accuracy, we applied EM algorithm to maximize (log)likelihood. The EM algorithm consists of two steps: An expectation step, E-Step, and a maximization step, M-Step. In E-Step, we keep the parameters $\boldsymbol{\Theta}$ fixed, and update the posterior probabilities $P(z_i = s|\boldsymbol{y}_i; \boldsymbol{\Theta}, \boldsymbol{\lambda})$ and the prior probabilities $\boldsymbol{\lambda}$. We also calculate the expectation of log likelihood by:

$$Q(\boldsymbol{\lambda}, \boldsymbol{\Theta}) = \mathbb{E}[\ln L_c(\boldsymbol{\lambda}, \boldsymbol{\Theta}; \boldsymbol{y}, \boldsymbol{z})] = \mathbb{E}_{z_1,\ldots,z_I}\left[\sum_{i=1}^{I}[\ln(\lambda_{z_i}) + \ln(f_{z_i}(\boldsymbol{y}_i|\boldsymbol{\theta}_{z_i}))\right] \tag{G.4}$$

In the M-Step, we keep the posterior probabilities fixed and update the parameters $\boldsymbol{\Theta} = \{\boldsymbol{\theta}_s\} = \{\{\boldsymbol{\mu}_s, \boldsymbol{\Sigma}_s\}\}$ by maximizing the expectation of log likelihood. The algorithm is as below in Algorithm 3.

Algorithm 3 EM algorithm for Gaussian mixture model

```
**Initialize**
$k \leftarrow 0, \hat{\boldsymbol{\lambda}}^k, \hat{\boldsymbol{\Theta}}^k$
**while** $k < K$ and $Q(\hat{\boldsymbol{\lambda}}^k, \hat{\boldsymbol{\Theta}}^k) - Q(\hat{\boldsymbol{\lambda}}^{k-1}, \hat{\boldsymbol{\Theta}}^{k-1}) > \epsilon$ **do**
    E-Step:
    **for** $s = 1, ..., S$ **do**
        $\hat{p}_{is} = P(z_i = s|\boldsymbol{y}_i; \hat{\boldsymbol{\lambda}}^k, \hat{\boldsymbol{\Theta}}^k) = \frac{\hat{\lambda}_s^k f_s(\boldsymbol{y}_i|\hat{\boldsymbol{\theta}}_s^k)}{\sum_{r=1}^{S} \hat{\lambda}_r^k f_r(\boldsymbol{y}_i|\hat{\boldsymbol{\theta}}_r^k)}$
        $\hat{\lambda}_s^{k+1} = \frac{1}{I}\sum_{i=1}^{I} \hat{p}_{is}$
    **end for**
    M-Step:
    $\hat{\boldsymbol{\Theta}}^{k+1} \leftarrow \arg\max_{\boldsymbol{\Theta}} Q(\hat{\boldsymbol{\lambda}}^{k+1}, \boldsymbol{\Theta})$

    $k \leftarrow k + 1$
**end while**
```

Once the stop condition is met in the algorithm (number of iterations or predetermined gap reached), we obtain the result of $\hat{\boldsymbol{\Theta}}$ and $\hat{\boldsymbol{\lambda}}$. Then we can use Equation 4.4 to compute the cluster membership probability p_{is} of each cluster for each observation. Based on the cluster membership probabilities, we can categorize each observation into corresponding cluster.

Vita

Jingyuan Bao was born and raise in Beijing, China, where he spent the first 22 years of his life. After he earned a Bachelor's degree in Civil Engineering at Tsinghua University, he arrived at the United States, and started his study in the Department of Civil and Environmental Engineering at Northwestern University under the supervision of Prof. Pablo L. Durango-Cohen. His research focuses on statistical modeling of fundraising and travel behavior.