The Dissertation Committee for Nazneen Ferdous certifies that this is the approved version of the following dissertation:

A New Estimation Approach for Modeling Activity-Travel Behavior: Applications of the Composite Marginal Likelihood Approach in Modeling Multidimensional Choices

Committee:

Chandra R. Bhat, Supervisor

Randy B. Machemehl

Jason Abrevaya

Steven T. Waller

Chandler Stolp

A New Estimation Approach for Modeling Activity-Travel Behavior:

Applications of the Composite Marginal Likelihood Approach in Modeling Multidimensional Choices

by

Nazneen Ferdous, B.Sc.; M.Sc.

Dissertation

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

The University of Texas at Austin

August, 2011

Dedication

To my mother, Ferdous Ara, who has always been there for me.

To my father, Abul Kashem, whose love and encouragement shaped who I am today.

Acknowledgments

First, I would like to express my sincere gratitude to my advisor, Dr. Chandra Bhat. Dr. Bhat belongs to a group of scholars who, in my mind, come under the class of “true academics, both in spirit and soul”. So, I am most appreciative of the fact that I got to work with Dr. Bhat for 4 years. I would also like to thank Dr. Randy Machemehl, Dr. Jason Abrevaya, Dr. S. Travis Waller, and Dr. Chandler Stolp for serving on my dissertation committee and for providing helpful comments/suggestions throughout the course of my Ph.D. research work. A big thank you to Lisa Macias for her friendship and kind assistance with numerous issues.

Thanks to many staff members of the department and my fellow grad students at The University of Texas at Austin. Special thanks to Abani and Namita for their unyielding friendship.

Lastly, but most importantly, I would like to warmly acknowledge and extend my sincere gratitude to my mother, my father, and my sisters for their love, encouragement, and steadfast support. As with all my past accomplishments, the credit for this dissertation should also go to my family.

A New Estimation Approach for Modeling Activity-Travel Behavior:

Applications of the Composite Marginal Likelihood Approach in Modeling Multidimensional Choices

Publication No. ____________

Nazneen Ferdous, Ph.D.

The University of Texas at Austin, 2011

Supervisor: Chandra R. Bhat

The research in the field of travel demand modeling is driven by the need to understand individuals' behavior in the context of travel-related decisions as accurately as possible. In this regard, the activity-based approach to modeling travel demand has received substantial attention in the past decade, both in the research arena as well as in practice. At the same time, recent efforts have been focused on more fully realizing the potential of activity-based models by explicitly recognizing the multi-dimensional nature of activity-travel decisions. However, as more behavioral elements/dimensions are added, the dimensionality of the model systems tends to explode, making the estimation of such models all but infeasible using traditional inference methods. As a result, analysts and practitioners often trade-off between recognizing attributes that will make a model behaviorally more representative (from a theoretical viewpoint) and being able to estimate/implement a model (from a practical viewpoint).

An alternative approach to deal with the estimation complications arising from multi-dimensional choice situations is the technique of composite marginal likelihood

(CML). This is an estimation technique that is gaining substantial attention in the statistics field, though there has been relatively little coverage of this method in transportation and other fields. The CML approach is a conceptually and pedagogically simpler simulation-free procedure (relative to traditional approaches that employ simulation techniques), and has the advantage of reproducibility of the results. Under the usual regularity assumptions, the CML estimator is consistent, unbiased, and asymptotically normally distributed.

The discussion above indicates that the CML approach has the potential to contribute in the area of travel demand modeling in a significant way. For example, the approach can be used to develop conceptually and behaviorally more appealing models to examine individuals' travel decisions in a joint framework. The overarching goal of the current research work is to demonstrate the applicability of the CML approach in the area of activity-travel demand modeling and to highlight the enhanced features of the choice models estimated using the CML approach. The goal of the dissertation is achieved in three steps as follows: (1) by evaluating the performance of the CML approach in multivariate situations, (2) by developing multidimensional choice models using the CML approach, and (3) by demonstrating applications of the multidimensional choice models developed in the current dissertation.

Table of Contents

List of Tables

List of Figures

Chapter 1
Introduction

1.1 Background: The Role of Travel Demand Models

Travel demand models (TDMs) are used to predict individuals' travel behaviors over a period of time, typically a weekday but sometimes also weekend days. Specifically, TDMs are designed to predict different dimensions of an individual's/agent's activity and travel behavior, including number of activity episodes, accompaniment arrangements, travel modes, destinations, activity durations, and other time-use behavior. Such models can be used to provide information on existing travel patterns as well as to forecast the future activity-travel patterns of individuals. For example, the outputs from a TDM can be used to analyze directional traffic flow on a roadway, obtain information on modal share, calculate travel time and delay, evaluate monetary and non-monetary benefits of building a new infrastructure, and quantify vehicular emission, to list just a few applications. In addition, a TDM may also be used as a tool to design and develop strategies that will proactively influence individuals' travel behaviors. For instance, assume that a toll is introduced on a segment of a congested road as part of a traffic management strategy. A commuter, who usually uses this road, may decide not to pay the toll and instead use an alternative (even if longer) route to get to work. Thus, introduction of a toll road has an immediate or short-term effect on this individual's travel pattern. In the long-term, the individual may decide to move to a new residential location to avoid longer commute times. A TDM can quantify these changes and aid decision-makers in developing strategies designed to affect individual's short-term and long-term travel behavior. In addition to route choice, other short-term travel-related choices of an individual that may be influenced include mode choice, drop-off/pick-up responsibility, number of non-mandatory activity episode participations, activity duration, time-of-day, and trip-chaining propensity. Longer term behavioral shifts may include individual's work location choice and household-level choices such as residential location, car ownership, and vehicular fleet composition.

It is clear from the above discussion that a TDM can be a powerful tool to manage, influence, and control individual's travel behavior at a disaggregate level and the overall demand for travel at an aggregate level. Of course, the effectiveness of a TDM depends on the level of accuracy of its prediction. In general, the more accurately a model can capture behavior and the responsiveness of an agent to observed (and unobserved) factors/stimuli, the better is its prediction capability and the overall performance level. In this regard, there are three general types of models that are commonly used to predict demand for travel. These are the trip-based model, the tour-based model, and the activity-based model. Each of these models is discussed in the subsequent sections.

1.1.1 The Trip-Based Class of Travel Demand Models

The trip-based class of models uses trips made by each agent in the study area as the unit of analysis. A trip-based model typically comprises four sequential steps: trip generation, trip distribution, mode choice, and traffic assignment. The trip generation step involves the estimation of the number of home-based and non home-based person trips produced from, and attracted to, each traffic analysis zone (TAZ) in the study area. The home-based trips are often divided into two categories: home-based work trips and home-based other (or non-work) trips. The second step, trip distribution, determines the number of trips from each zone to each other zone in the study area (that is, this step produces origin-destination (O-D) matrices of trip by purpose). The third step, mode choice, determines the mode of travel for each person trip. The travel mode choice usually includes at least the personal vehicle mode and the transit mode. This step converts person trips to vehicular trips. The fourth and the final step, traffic assignment, assigns the vehicle trips to the road network to obtain link-level vehicle volumes and travel times. In addition to the outputs from the mode choice step, external trip matrices containing information on truck flows may also be assigned to the road network in this step.

The trip-based class of models is the most widely used framework for modeling travel demand. Though this class of models has the virtue of simplicity, one of the major drawbacks is that it considers trips to be independent of one another. That is, the trip-based class of models assumes that there is no spatial and temporal linkage between the successive trips of the same agent. For illustration, consider Figure 1.1, which depicts the travel pattern of a fictitious worker on a weekday. In the figure, the individual undertakes four trips: (1) a trip from home to the coffee shop, (considered a home-based non-work trip), (2) a trip from the coffee shop to the individual's workplace (considered a non home-based work trip), (3) a trip from the workplace to the restaurant in the afternoon (considered a non home-based non-work trip), and (4) a trip from the restaurant to home (considered a home-based non-work trip). The problem with the trip-based approach is that it does not consider the linkages between the four trips just listed. That is, it is likely that an individual will use the same travel mode for all the four trips, and that the locations of the coffee shop and the restaurant will be determined, at least in part, by the location of the home and workplaces. But the trip-based class of models characterizes the travel behavior of the individual as comprising two home-based non-work trips, one non home-based work trip, and one non home-based non-work trip. There is no relationship retained between these trips, because of the individual trip unit of analysis. Consequently, the trip-based model does not preserve the integrity of mode choices, location choices, and time-of-day of participation choices among the different activity episodes.

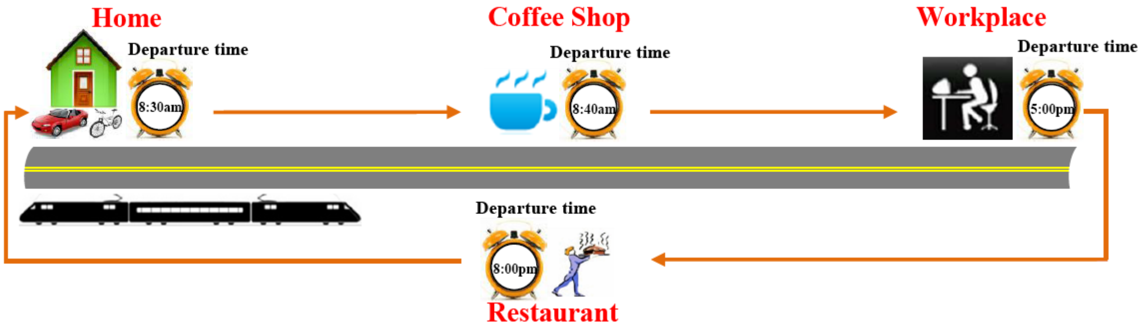


Figure 1.1 Travel Pattern of an Individual

1.1.2 The Tour-Based Class of Travel Demand Models

The tour-based class of models uses tours as the basic unit of analysis. A tour may be defined as a closed chain of trips beginning and ending at home (defined as a home tour), or beginning and ending at work (defined as a work tour, and applicable only for employed individual). Within a tour, the individual makes one or more stops and the trips in the tour are the result of the stops being at locations dispersed in space (and at a different location than the origin point of the tour). In Figure 1.1, the individual participates in a home-based tour with three stops – the coffee shop, the workplace, and the restaurant, all of which are dispersed in space. Thus, the tour includes four trips: (1) home to coffee shop, (2) coffee shop to workplace, (3) workplace to restaurant, and (4) restaurant to home. The tour-based model structure ensures that the integrity of the sequence of the trips in a tour, the destination choice, the mode choice, and the time-of-day of the trips in the tour are all preserved. For example, the tour-based model will, in general, predict that the coffee shop is located between the individual's home and workplace, rather than at a location that is on the other side of the work place from the person's home. Also, if the individual drives a car to the coffee place, the tour-based model will assign a very high probability that the person will also use the car for other trips in the tour.

1.1.3 The Activity-Based Class of Travel Demand Models

The activity-based class of models also uses tours as the basic unit of analysis. However, the tour-based approach and the activity-based approach view travel quite differently. Specifically, the activity-based approach regards travel as a demand derived from the need to pursue activities (Jones, 1979, Jones *et al.* 1990, Bhat and Koppelman, 1999, and Pendyala and Goulias, 2002). That is, an activity-based travel demand model assumes that individuals usually travel to participate in activities, and considers the activity episode as the unit of analysis by analyzing such activity episode dimensions as the number of activity episodes by purpose, activity episode companion choice, activity episode location, activity episode duration, and activity episode participation time (as opposed to focusing on the characteristics of the trips comprising a tour). For example, in Figure 1.1, the individual travels to his/her workplace to participate in a work activity episode. Similarly, in the afternoon, the individual travels to the restaurant to participate in an eat-out activity episode. In such a framework, the focus, for instance, is on the duration in continuous time (in contrast to in 30-minute or 1-hour "chunks") of the coffee stop rather than, as does the tour-based approach, on the end-time of the trip terminating at the coffee stop location and the start time of the trip immediately after the coffee stop. By using activity episodes as the building blocks and using continuous time, the activity-based class of models ties directly to a time-use decision framework in which time is treated as an all-encompassing continuous entity within which individuals make activity/travel participation decisions. This approach is also able to represent spatio-temporal interactions within and between individuals in a straightforward manner because of the consideration of continuous time. Further, the consideration of time as a continuous entity enables the analyst to maintain integrity in time and space of joint activities across household members, and enables the consideration of time-varying and dynamic pricing policies in an effective and rigorous manner.

1.2 Problem Statement

Over the past three decades, the field of travel demand modeling has experienced a shift from the traditional four-step trip-based approach to travel demand modeling toward a more behaviorally-oriented activity-based approach to travel demand modeling, prompted by the limitations of the trip-based approach and an increasing recognition of the need to understand individuals' behavioral responses to travel management measures. As just discussed, while an individual's activity participation behaviors and time use patterns are represented more accurately in an activity-based approach, the approach also leads to the econometric challenge of modeling multi-dimensional choice situations because traditional classical and Bayesian simulation techniques become extremely cumbersome and often impractical in these situations. Also, the accuracy of simulation techniques is known to degrade rapidly at medium-to-high dimensions, and the simulation noise increases substantially as well. This leads to convergence problems during estimation. This difficulty with model estimation often leads to the use of simplistic models with aggregated alternatives, or uni-dimensional models for each dimension, or a pre-specified hierarchical system of the dimensions (more on this later). But these "quick-fixes" also undo the richness of the activity-based approach. For demonstration, consider a case where an analyst wants to examine the weekday activity episode participation patterns of adult individuals, the choice of companions for each episode, and the travel mode used, all within a unified framework. Such a model will provide useful insights into inter-individual interactions and how such interactions may affect mode choice decision. For this exercise, assume that an individual can participate only in the following out-of-home activities: work, maintenance activity, and discretionary activity. The travel modes available to the individual may include drive alone (DA), shared ride (SR), public transportation (PT), and non-motorized modes (NM). Also, an individual can participate in the activities either alone, with only family member(s), or with "other" member(s) ("other" members include a combination of family and non-family members). An econometric model for this situation with correlation between all the alternatives (due to unobserved factors) will involve evaluation of a 28-dimensional integral.[1] To avoid simulation-related difficulties associated with the evaluation of a high-dimensional integral, there are, traditionally, three ways to model this situation:

(1) Develop an aggregate model with fewer alternatives. For example, reclassify the activity types into two categories: mandatory activity (includes work) and non-mandatory

[1] The feasible combinations of activity, companion type, and mode choice are as follows:

Activity Type	Activity Companion Choice	Travel Mode
Work	Alone	DA, SR, PT, or NM
Maintenance activity	Alone	DA, SR, PT, or NM
Maintenance activity	With only family member(s)	DA, SR, PT, or NM
Maintenance activity	With "other" member(s)	DA, SR, PT, or NM
Discretionary activity	Alone	DA, SR, PT, or NM
Discretionary activity	With only family member(s)	DA, SR, PT, or NM
Discretionary activity	With "other" member(s)	DA, SR, PT, or NM

Total number of alternatives = (4×7) = 28.

activity (includes maintenance and discretionary activities). Similarly, consider only two mode choices (motorized and non-motorized) and two companion types (alone and not alone). This will reduce the dimensionally of the integral from 28 to 6. Then, the model system may be estimated using maximum simulated likelihood approach or the Bayesian approach without encountering any significant difficulty.

(2) Develop a disaggregate model with no dependence or partial dependence between the alternatives due to unobserved factors, or

(3) Develop a sequential modeling framework. For example, one may develop a model for the activity participation frequency first, followed by a joint model of companion type and mode choice.

Of these three options, whichever modeling approach is chosen by the analyst, the resulting model will be less sensitive in terms of capturing the effects of observed and unobserved variables (such as intra-household interactions, peer-influence, built-environment related factors) on individuals' activity-travel behaviors. This, in turn, can translate to less accurate assessment of the effects of travel demand management strategies on individuals' travel choices.

1.3 Objectives of the Dissertation

The research undertaken in the current dissertation is motivated by the discussion above. Specifically, we propose the use of an alternative approach, the composite marginal likelihood (CML) approach, which allows estimation of multidimensional models and deals with the estimation complications discussed in the previous section. The CML is an estimation technique that is gaining substantial attention in the statistics field, though there has been relatively little coverage of this method in transportation and other fields. The CML method is based on forming a surrogate likelihood function that compounds much easier-to-compute, lower-dimensional, marginal likelihoods. Very simply stated, the CML approach is based on developing the marginal log-likelihood of the joint distribution of a lower dimensional number of alternatives at one time (such as two alternatives at one time), while ignoring all other alternatives. Then, by developing and maximizing a surrogate log-likelihood function that is the sum of the log-likelihood of each possible combination of the lower dimensional marginal distribution, one obtains a consistent, unbiased, and asymptotically normally distributed estimator of all the relevant parameters characterizing the original high dimensional distribution. Thus, the CML approach represents a conceptually and pedagogically simpler simulation-free procedure relative to simulation techniques, and has the advantage of reproducibility of the results (see Bhat *et al.*, 2010a). Also, as indicated by Varin and Vidoni (2009), it is possible that the "maximum CML estimator can be consistent when the ordinary full likelihood estimator is not". This is because the CML procedures are typically more robust and can represent the underlying low-dimensional process of interest more accurately than the low dimensional process implied by an assumed (and imperfect) high-dimensional multivariate model. Finally, the CML approach can be easily implemented using simple optimization software for likelihood estimation.

The discussion above indicates that the CML approach has the potential to contribute in the area of travel demand modeling in a significant way. For example, the approach can be used to develop conceptually and behaviorally more appealing models to examine individuals' short-term travel decisions in a joint framework. Within the context of the activity-travel behavior modeling approach, application of the CML approach can be further extended to encompass the area of land use modeling, a research area that is of considerable interest to the travel demand analysts due to its direct impact on individuals long-term travel behavior. The overarching goal of the current research work is to demonstrate the applicability of the CML approach in the area of activity-travel demand modeling and to highlight the enhanced features of the choice models estimated using the CML approach. The goal of the research is realized by considering the following objectives.

The first objective is to assess the performance of the CML approach relative to the "benchmark" maximum-simulated likelihood (MSL) approach. This is because the CML estimator (theoretically speaking) loses some efficiency relative to traditional maximum likelihood estimation, though a limited investigation has shown efficiency loss to be negligible (Zhao and Joe, 2005, Lele, 2006, Joe and Lee, 2009). In the current research work, this issue is investigated further. Specifically, the performance of the CML approach is compared with the maximum-simulated likelihood (MSL) approach in multivariate situations. The ability of the two approaches to recover model parameters in simulated data sets is examined. In addition, the efficiencies of estimated parameters and the computational costs of both approaches are also compared.

The second objective is to evaluate the ability of the CML approach to recover model parameters in a multi-dimensional context in both a cross-sectional setting as well as a panel setting. Also, the potential impact of different correlation structures on the performance of the CML approach is studied.

The remaining objectives demonstrate the use of the CML technique to estimate rich model structures for activity-travel demand modeling. Specifically, the third objective is to develop a behaviorally rich model structure to analyze inter-individual interactions in activity episode generation. Specifically, a multivariate (30-variate) modeling framework is developed to examine the interactions in non-work activity episode decisions across household and non-household members at the level of activity generation. Such a model structure accommodates complementarity and substitution effects in individuals' activity participation behaviors.

The fourth objective is to formulate a joint model of walking and bicycling activity duration (also referred to as non-motorized transport modes) using a hazard based specification that recognizes the presence of unobserved heterogeneity in the activity participation behaviors of individuals. In particular, the model accounts for unobserved factors specific to individuals, family/household-level interactions, social group or peer influences, and spatial clustering effects that contribute to the heterogeneity in non-motorized transport mode use behavior.

The fifth objective is to propose and estimate a spatial panel ordered-response model with temporal autoregressive error terms to analyze changes in urban land development intensity level over time. Such a model structure maintains a close linkage between the land owner's decision and the land development intensity level. Also, the model structure recognizes that spatial dependence is a substantive issue in the current empirical context, and is caused by didactic interactions between the land owners. In addition, the model structure incorporates spatial heterogeneity, spatial heteroscedasticity, and temporal dependence. The model can be used to examine and understand the behavior of land owners, who ultimately make land use decisions.

The sixth and the final objective is to demonstrate the application of the models estimated in objectives 3, 4, and 5. In addition to illustrating the application of the models, the exercises within this sixth objective also highlight the improved performance of the models developed in this dissertation relative to their naïve counterparts.

1.4 Model System Used in the Current Dissertation

All the models developed in the current dissertation are based on an ordered-response model structure. Ordered-response model structures are used when analyzing ordinal discrete outcome data that may be considered as manifestations of an underlying scale that is endowed with a natural ordering. Examples include ratings data (of consumer products, bonds, credit evaluation, movies, *etc.*), or likert-scale type attitudinal/opinion data (of air pollution levels, traffic congestion levels, school academic curriculum satisfaction levels, teacher evaluations, *etc.*), or grouped data (such as bracketed income data in surveys or discretized rainfall data), or count data (such as the number of trips made by a household, the number of episodes of physical activity pursued by an individual, and the number of cars owned by a household). In all of these situations, the observed outcome data may be considered as censored (or coarse) measurements of an underlying latent continuous random variable. The censoring mechanism is usually characterized as a partitioning or thresholding of the latent continuous variable into mutually exclusive (non-overlapping) intervals. The reader is referred to McKelvey and Zavoina (1971) and Winship and Mare (1984) for some early expositions of the ordered-response model formulation, and Liu and Agresti (2005) for a survey of recent developments. The reader is also referred to a recent book by Greene and Hensher (2010) for a comprehensive history and treatment of the ordered-response model structure. These recent reviews indicate the abundance of applications of the ordered-response model in the sociological, biological, marketing, and transportation sciences, and the list of applications only continues to grow rapidly.

While the applications of the ordered-response model are quite widespread, much of these are confined to the analysis of a single outcome, with a sprinkling of applications associated with two and three correlated ordered-response outcomes. Some very recent studies of two correlated ordered-response outcomes include Scotti (2006), Mitchell and Weale (2007), Scott and Axhausen (2006), and LaMondia and Bhat (2011).[2] The study by Scott and Kanaroglou (2002) represents an example of three correlated ordered-response outcomes. But the examination of more than two to three correlated outcomes is rare, mainly because the extension to an arbitrary number of correlated ordered-response outcomes entails, in the usual likelihood function approach, integration of dimensionality equal to the number of outcomes. On the other hand, there are many instances when interest may be centered around analyzing several ordered-response outcomes simultaneously, such as in the case of the number of episodes of each of several activities, or satisfaction levels associated with a related set of products/services, or multiple ratings measures regarding the state of health of an individual/organization (we will refer to such outcomes as cross-sectional multivariate ordered-response outcomes). There are also instances when the analyst may want to analyze time-series or panel data of ordered-response outcomes over time, and allow flexible forms of error correlations over these outcomes. For example, the focus of analysis may be to examine rainfall levels (measured in grouped categories) over time in each of several spatial regions, or individual stop-making behavior over multiple days in a week, or individual headache severity levels at different points in time (we will refer to such outcomes as panel multivariate ordered-response outcomes).

[2] The first three of these studies use the bivariate ordered-response probit (BORP) model in which the stochastic elements in the two ordered-response equations take a bivariate normal distribution, while the last study develops a more general and flexible copula-based bivariate ordered-response model that subsumes the BORP as but one special case.

In the analysis of cross-sectional and panel ordered-response systems with more than three outcomes, the norm until very recently has been to apply numerical simulation techniques based on a maximum simulated likelihood (MSL) approach or a Bayesian inference approach. However, such simulation-based approaches become impractical in terms of computational time, or even infeasible, as the number of ordered-response outcomes increases. Even if feasible, the numerical simulation methods do get imprecise as the number of outcomes increase, leading to convergence problems during estimation.

The discussion above highlights the applications of the ordered-response model in a wide variety of fields, including the field of travel demand, and the estimation problem associated with modeling correlated multiple outcomes. Thus, the ordered-response model provides an ideal framework to undertake the current research.

1.5 Structure of the Dissertation

The six research objectives identified in Section 1.3 may be grouped into three categories, based on the nature of their contributions: (1) Group A includes objectives 1 and 2, and contributes toward an evaluate of the performance of the CML approach, (2) Group B, which includes objectives 3, 4, and 5, contributes toward the formulation and estimation of behaviorally more representative, but also analytically tractable, travel choice and land use models, and (3) Group C includes objective 6, and contributes toward reducing the widening gap between travel demand modeling research and practice by highlighting the practical applications of the models developed in the dissertation. Attainment of each group of objectives is presented sequentially (from Chapter 2 to Chapter 7) in Part I, Part II, and Part III of this dissertation. The last and the final chapter (Chapter 8) concludes the dissertation by summarizing the findings in the previous chapters (Chapter 3 to Chapter 7), discussing some limitations of the current work, and suggesting directions for future research. A schematic representation of the dissertation structure is presented in Figure 1.2. A schematic description of each part of the dissertation and the final chapter is provided below.

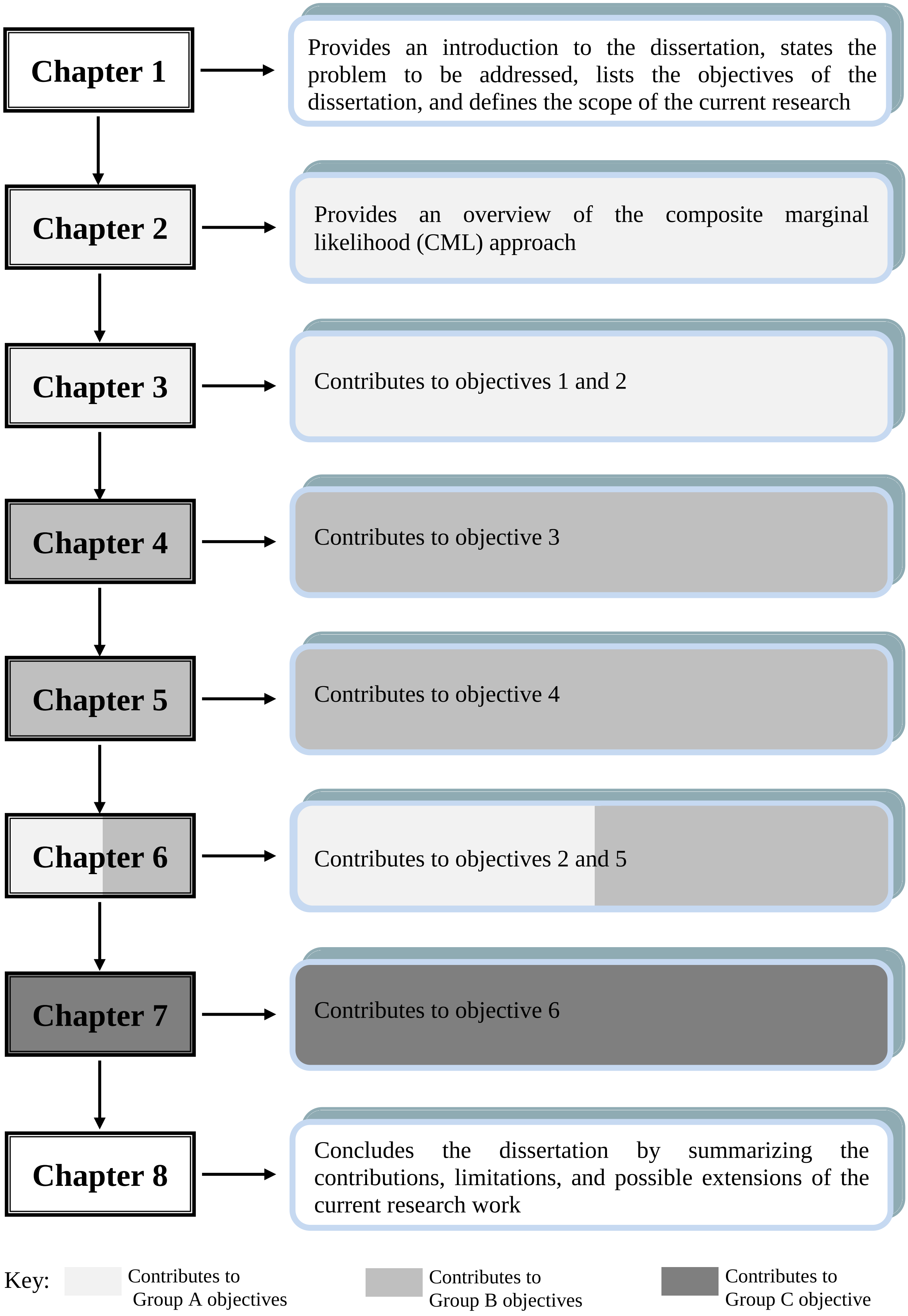

Chapter 1
Provides an introduction to the dissertation, states the problem to be addressed, lists the objectives of the dissertation, and defines the scope of the current research
Chapter 2
Provides an overview of the composite marginal likelihood (CML) approach
Chapter 3
Contributes to objectives 1 and 2
Chapter 4
Contributes to objective 3
Chapter 5
Contributes to objective 4
Chapter 6
Contributes to objectives 2 and 5
Chapter 7
Contributes to objective 6
Chapter 8
Concludes the dissertation by summarizing the contributions, limitations, and possible extensions of the current research work
Key:
Contributes to Group A objectives
Contributes to Group B objectives
Contributes to Group C objective

Figure 1.2 Dissertation Structure

- **Part I:** This part provides an overview of the CML approach, compares the performance of the CML approach with the MSL approach, and assesses the ability of the CML approach to recover model parameters in cross-sectional and panel data context. This part consists of two chapters as follows:

  Chapter 2, while not contributing to any specific research objective, provides a backdrop for the subsequent chapters (and the dissertation objectives) by presenting an overview of the CML approach. Specifically, in this chapter, the composite likelihood function approach is discussed in general, and the composite marginal likelihood (CML) approach is discussed in particular, including the properties of the CML estimator, standard error estimation technique, and hypothesis testing.

  Chapter 3 contributes to objectives 1 and 2. The first objective is achieved by comparing the performance of the CML approach with the MSL approach when the MSL approach is feasible. For this, a 5-dimensional ordered-response model is estimated using a number of simulated cross-sectional data sets corresponding to different levels of correlation. For the MSL approach, the Geweke-Hajivassiliou-Keane (GHK) Probability Simulator is used, while for the CML approach, the pairwise marginal likelihood approach is used. The performance of the two approaches is compared using three measures: the absolute percentage bias, the finite sample standard error, and the asymptotic standard error. An assessment of the performances of the two (CML and MSL) approaches due to different correlation structures is also undertaken. In addition, this third chapter also contributes partly toward the second objective by evaluating the ability of the CML approach to recover model parameters in a cross-sectional data setting.
- **Part II:** This part presents a series of econometric models that are behaviorally appealing, but are generally considered impractical to be estimated by traditional estimation approaches. Part II comprises three chapters as follows:

  Chapter 4 contributes to the third objective by developing a multivariate ordered-response model system with flexible error structure to model non-work activity episode decisions and activity companion choices in a joint framework. Such a model structure recognizes that activity participation decisions of an individual are influenced by other household and non-household members. Another salient feature of this model is that it has a flexible structure that accommodates complementarity and substitution effects in activity participation and accompaniment arrangement decisions.

Chapter 5 presents a framework that models the walking and bicycling activity durations of individuals simultaneously using a multilevel cross-cluster hazard-based model system that accounts for a range of interactions and spatial effects. Specifically, in addition to the usual individual-specific factors, family (*i.e.*, household-specific) interactions, social group (peer) influences, and spatial clustering effects are also considered as potential factors that contribute to heterogeneity in non-motorized transport mode use behavior. The proposed model system is capable of accommodating grouped duration responses often encountered in activity-travel surveys. This chapter contributes to objective 4.

Chapter 6 proposes and estimates an econometric model with spatial and temporal dependence to analyze changes in urban land development intensity level over time. The model framework developed here has several salient features. First, the model recognizes that it is important to maintain the link between the decision making agent (*i.e.*, the land owner) and the observed land development intensity level (undeveloped land, land less-intensely developed for residential use, *etc.*). Second, spatial dependence is introduced not only through explanatory variables and error terms, but also through time-invariant effects of random coefficients. Third, the model structure accommodates spatial heterogeneity and spatial heteroscedasticity. Finally, temporal dependence effects are introduced at two levels: time-invariant temporal effects and time-varying temporal effects. In addition to the empirical analysis, a simulation exercise is undertaken to assess the ability of the CML approach to recover parameter in panel data context. The performance of the CML approach was assessed using four different panel data settings: panel data with low spatial and temporal dependence, panel data with low spatial but high temporal dependence, panel data with high spatial but low temporal dependence, and panel data with high spatial and temporal dependence. The simulation exercise also highlights the consequence of ignoring spatial dependence and spatial heterogeneity when both are actually present. This chapter contributes partly to objective 2 and completely to object 6.

- **Part III:** This part includes Chapter 7, which contributes to objective 6 of the dissertation. In this chapter, the econometric models developed in Part II are applied to various empirical contexts to demonstrate their applications. The results presented here quantify the effects of employing a behaviorally-rich model and underline the advantages of incorporating such behavioral features within the modeling framework.

Finally, Chapter 8 concludes the current research work by summarizing the contributions of the research, highlighting the key empirical findings, discussing some limitations of the current dissertation, and sharing thoughts on future research in the area.

.

Part I

Chapter 2
The Composite Marginal Likelihood Approach: An Overview

2.1 Introduction

The composite marginal likelihood (CML) estimation approach is a relatively simple approach that can be used when the full likelihood function is near impossible or plain infeasible to evaluate due to the underlying complex dependencies. For instance, in a recent application, Varin and Czado (2010) examined the headache pain intensity of patients over several consecutive days. In this study, a full information likelihood estimator would have entailed as many as 815 dimensions of integration to obtain individual-specific likelihood contributions, an infeasible proposition using the computer-intensive simulation techniques. In this case and other similar cases, the CML approach provides an alternative estimation technique. The CML approach belongs to the more general class of composite likelihood (CL) function approaches. In the next section, we discuss the composite likelihood function approaches.

2.2 The Composite Likelihood Function (CLF) Approach

The composite likelihood approach is based on forming a surrogate likelihood function that compounds much easier-to-compute lower-dimensional likelihoods.[3] For illustration, let Y be a Q-dimensional random variable with density function $f(y;\boldsymbol{\theta})$ $(Y \subseteq \Re^Q, Q \geq 1)$, where $\boldsymbol{\theta}$ is a $(D\times 1)$ parameter vector ($\boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq \Re^D, D \geq 1$). Also, let { $E_1, E_2, ..., E_M$ } be a set of events with the corresponding likelihood functions $(L_1(\boldsymbol{\theta}; y), L_2(\boldsymbol{\theta}; y), ..., L_M(\boldsymbol{\theta}; y))$. Then, the composite likelihood function may be written as follows:

[3] The composite likelihood was first proposed by Besag (1974) under the name pseudolikelihood. Lindsay in 1988 first used the term composite likelihood to describe the class of likelihood functions discussed in this chapter.

$$L_{CLF}(\boldsymbol{\theta}) = \prod_{m=1}^{M} (L_m(\boldsymbol{\theta}; y))^{w_m} \tag{2.1}$$

where w_m is a non-negative weight. In the above definition of the composite likelihood function, the set of events could be either conditional or marginal. If the likelihood function associated with an event is the product of conditional densities, then the resulting likelihood function is called the Composite Conditional Likelihood (see Stein *et al.*, 2004 and Wang and Williamson, 2005 for applications of the composite conditional likelihood estimation technique). On the other hand, if the likelihood function $(L_m(\boldsymbol{\theta}; y))$ in Equation (2.1) is a marginal likelihood, then $L_{CLF}(\boldsymbol{\theta})$ may be called the composite marginal likelihood. In this research work, we focus on the composite marginal likelihood (CML) approach. In the next section the CML approach is discussed in more detail. Then, in Sections 2.4 through 2.6, the properties of the CML estimator, standard error estimation technique, and hypothesis testing procedures are presented. The final section concludes the chapter by providing a brief summary.

2.3 The Composite Marginal Likelihood (CML) Approach

The simplest CML may be formed by assuming independence across the variables. In this case, the likelihood function is the product of univariate probabilities for each variable. However, this approach does not provide estimates of correlation that are of interest in a multivariate context. Another approach is the pairwise likelihood function formed by the product of likelihood contributions of all or a selected subset of couplets (*i.e.*, pairs of variables or pairs of observations). The pairwise likelihood estimator is typically robust to misspecification (see Varin and Vidoni, 2009, and Varin, 2008). The approach is very simple computationally with literally no convergence-related issues. It can also be very easily coded in software packages that allow the computation of a bivariate normal cumulative distribution function and have an optimization procedure for maximizing a function with respect to embedded parameters. Almost all earlier research efforts employing the CML technique have used the pairwise approach, including Bellio and Varin (2005), de Leon (2005), Varin *et al.* (2005), Engle *et al.* (2007), Apanasovich *et al.*

(2008), Varin and Vidoni (2009), and Bhat *et al.* (2010a). In the current research, all estimation efforts are also undertaken using the pairwise marginal likelihood approach. Specifically, we employ a pairwise marginal likelihood estimation approach that corresponds to a composite marginal approach based on bivariate normal distribution.

In addition to the independence and the pairwise likelihood, the analyst can also consider larger subsets of observations, such as triplets or quadruplets or even higher dimensional subsets (see Engler *et al.*, 2006, and Caragea and Smith, 2007). In general, the issue of whether to use pairwise likelihoods or higher-dimensional likelihoods remains an open, and under-researched, area of research. However, it is generally agreed that the pairwise approach is a good balance between statistical and computation efficiency. The reader is referred to Varin (2008) and Varin *et al.* (2011) for a comprehensive overview of applications of the CML technique in a wide variety of fields.

2.4 Properties of the CML Estimator

The properties of the CML estimator may be derived using the theory of estimating equations (see Lindsay, 1988, Cox and Reid, 2004, and Molenberghs and Verbeke, 2005 for details). For convenience, a number of key properties of the CML estimator are summarized here. Under the usual regularity assumptions:

1) *The CML estimator is consistent*. In the context of the pairwise CML approach used in the current research, the surrogate likelihood function represented by the CML function is the product of the marginal likelihood functions formed by each pair of variables/observations. In general, maximization of the original likelihood function will result in parameters that tend to maximize each pairwise likelihood function. Since the CML is the product of pairwise likelihood contributions, it will therefore provide consistent estimates. [4]

[4] Another equivalent way to see this is to assume we are discarding all but two randomly selected variables in the original likelihood function. Of course, we will not be able to estimate all the model parameters from two random variables, but if we could, the resulting parameters would be consistent because information

2) *The CML estimator is unbiased*. This follows from the unbiasedness of the CML score function, which is a linear combination of proper score functions associated with the marginal event probabilities forming the composite likelihood.

3) *The CML estimator is asymptotically normally distributed*. Let, $\hat{\boldsymbol{\theta}}_{CML}$ be a CML estimate of the parameter vector $\boldsymbol{\theta}$. Then, $\hat{\boldsymbol{\theta}}_{CML}$ is asymptotically normal distributed as follows:

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_{CML} - \boldsymbol{\theta}) \xrightarrow{d} MVN_D[\mathbf{0}, \mathbf{V}_{CML}(\boldsymbol{\theta})]$$

where n is the sample size, MVN_D is a multivariate normal distribution of size D, and $\mathbf{V}_{CML}(\boldsymbol{\theta})$ is the inverse of Godambe's (1960) sandwich information matrix $(\mathbf{G}(\boldsymbol{\theta}))$.

2.5 Standard Error Estimation

The variance-covariance matrix above $(\mathbf{V}_{CML}(\boldsymbol{\theta}))$ may be given as follows (see Zhao and Joe, 2005):

$\mathbf{V}_{CML}(\boldsymbol{\theta}) = [\mathbf{G}(\boldsymbol{\theta})]^{-1} = [\mathbf{H}(\boldsymbol{\theta})]^{-1}[\mathbf{J}(\boldsymbol{\theta})][\mathbf{H}(\boldsymbol{\theta})]^{-1}$, where

$$\mathbf{H}(\boldsymbol{\theta}) = E\left[-\frac{\partial^2 \log L_{CML}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right] \text{ and } \tag{2.2}$$

$$\mathbf{J}(\boldsymbol{\theta}) = E\left[\left(\frac{\partial \log L_{CML}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right)\left(\frac{\partial \log L_{CML}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}'}\right)\right]$$

where $\log L_{CML}(.)$ is the logarithm of the composite marginal likelihood. In Equation (2.2), the $\mathbf{H}(\boldsymbol{\theta})$ matrix can be estimated in a straightforward manner using the Hessian of the negative of $\log L_{CML}(\boldsymbol{\theta})$, evaluated at $\hat{\boldsymbol{\theta}}_{CML}$. This is because, in the context of pairwise likelihood estimator, the information identity remains valid for each pairwise term

(captured by other variables) is being discarded in a purely random fashion. The CML estimation procedure works similarly, but combines all variables observed two at a time, while ignoring the full joint distribution of the variables.

forming the composite marginal likelihood. However, depending on the dependence structure of the model, estimation of the $\mathbf{J}(\boldsymbol{\theta})$ matrix may be more difficult. In the current research, we discuss and demonstrate estimation techniques of the $\mathbf{J}(\boldsymbol{\theta})$ matrix in three situations: (1) simple case with no underlying dependence across the observations (demonstrated in Chapter 4), (2) clustering effects creating multi-level dependence across the observations (demonstrated in Chapter 5), and (3) spatial “spillover” effects (demonstrated in Chapter 6).

2.6 Hypothesis Testing

Hypothesis testing and model selection procedures similar to those available with the full maximum likelihood approach are also available with the CML approach (see Varin and Vidoni, 2009, Pace *et al.*, 2011, and Varin and Czado, 2010; Bhat, 2011 provides a concise summary). The common statistical tests are summarized here:

1) For a single parameter, the statistical test may be pursued using the usual t-statistic.

2) When the statistical test involves multiple parameters between two nested models, the composite likelihood ratio test (CLRT) statistic, which is similar to the likelihood ratio test in full maximum likelihood estimation, may be employed. For this, consider the null hypothesis $H_0 : \boldsymbol{\tau} = \boldsymbol{\tau}_{\mathbf{0}}$ against the alternative hypothesis $H_1 : \boldsymbol{\tau} \neq \boldsymbol{\tau}_{\mathbf{0}}$, where $\boldsymbol{\tau}$ is a subvector of $\boldsymbol{\theta}$ of dimension *d*. Let $\hat{\boldsymbol{\theta}}$ be the CML estimator of the unrestricted model (without the restriction imposed by the null hypothesis), and let $\hat{\boldsymbol{\theta}}_{\mathbf{0}}$ be the CML estimator under the null hypothesis. Then, the CLRT statistic may be calculated as follows:

$$\text{CLRT } = 2[\log L_{CML}(\hat{\boldsymbol{\theta}}) - \log L_{CML}(\hat{\boldsymbol{\theta}}_{\mathbf{0}})], \qquad (2.3)$$

However, the above CLRT statistic does not have the standard chi-squared asymptotic distribution under the null hypothesis as in the case of the maximum likelihood inference procedure. One alternative is to use bootstrapping to obtain the exact distribution of the CLRT statistic. The procedure is as follows (Varin and Czado, 2008):

a. Let the estimation sample be denoted as y_{obs}, and the observed CLRT value as $\mathrm{CLRT}(y_{obs})$.

b. Generate B sample data sets $y_1, y_2, y_3, \ldots, y_B$ using the CML convergent values under the null hypothesis.

c. Compute the CLRT statistic for each generated data set, and label it as $\mathrm{CLRT}(y_b)$.

d. Calculate the p-value of the test using the following expression:

$$p = \frac{1 + \sum_{b=1}^{B} I\left\{\mathrm{CLRT}(y_b) \geq \mathrm{CLRT}(y_{obs})\right\}}{B+1}, \text{ where } I\{A\} = 1 \text{ if } A \text{ is true.}$$

Another alternative is to adjust the CLRT statistic to obtain an adjusted composite likelihood ratio test (ADCLRT) statistic (see Varin and Vidoni, 2009, Pace *et al.*, 2011 and Bhat, 2011). For this, define $[\mathbf{H}_{\boldsymbol{\tau}}(\boldsymbol{\theta})]^{-1}$ and $[\mathbf{G}_{\boldsymbol{\tau}}(\boldsymbol{\theta})]^{-1}$ as the $(d \times d)$ submatrices of $[\mathbf{H}(\boldsymbol{\theta})]^{-1}$ and $[\mathbf{G}(\boldsymbol{\theta})]^{-1}$, respectively, which correspond to the vector $\boldsymbol{\tau}$. The following ADCLRT statistic may be considered to be asymptotically chi-squared distributed with *d* degrees of freedom:

$$\mathrm{ADCLRT} = \frac{[\mathbf{S}_{\boldsymbol{\tau}}(\boldsymbol{\theta})]'[\mathbf{H}_{\boldsymbol{\tau}}(\boldsymbol{\theta})]^{-1}[\mathbf{G}_{\boldsymbol{\tau}}(\boldsymbol{\theta})][\mathbf{H}_{\boldsymbol{\tau}}(\boldsymbol{\theta})]^{-1}[\mathbf{S}_{\boldsymbol{\tau}}(\boldsymbol{\theta})]}{[\mathbf{S}_{\boldsymbol{\tau}}(\boldsymbol{\theta})]'[\mathbf{H}_{\boldsymbol{\tau}}(\boldsymbol{\theta})]^{-1}[\mathbf{S}_{\boldsymbol{\tau}}(\boldsymbol{\theta})]} \times \mathrm{CLRT}$$

where $\mathbf{S}_{\boldsymbol{\tau}}(\boldsymbol{\theta})$ is the $(d \times 1)$ submatrix of $\mathbf{S}(\boldsymbol{\theta}) = \left(\dfrac{\partial \log L_{CML}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)$ corresponding to the vector $\boldsymbol{\tau}$, and all the matrices above are computed at $\hat{\boldsymbol{\theta}}_{\mathbf{0}}$.

3) When the null hypothesis entails model selection between two competing non-nested models, the composite likelihood information criterion (CLIC) introduced by Varin and Vidoni (2005) may be used. The CLIC takes the following form:

$$\text{CLIC} = \log L_{CML}(\hat{\boldsymbol{\theta}}) - tr\left[\hat{\mathbf{J}}(\hat{\boldsymbol{\theta}})[\hat{\mathbf{H}}(\hat{\boldsymbol{\theta}})]^{-1}\right]$$

The model that provides a higher value of CLIC is preferred over the other models.

2.7 Summary

This chapter provided a description of the composite marginal likelihood (CML) approach. In addition, properties of the CML estimator, standard error estimation technique, and hypothesis testing procedures were also presented.

Chapter 3

A Comparison of the Composite Marginal Likelihood Estimation Approach with the Maximum Simulated Likelihood Approach in the Context of the Multivariate Ordered-Response Model System

3.1 Motivation

The CML approach is based on the assumption that the lower-dimensional marginal likelihoods forming the surrogate likelihood function are independent of each other. Though this allows us to specify and estimate models with complex dependence structure, a weakness of this assumption is that the second Bartlett identity $(\mathbf{H}(\boldsymbol{\theta}) = \mathbf{J}(\boldsymbol{\theta}))$ is no longer valid.[5] From this theoretical perspective, the maximum CML estimator should lose some efficiency relative to a full likelihood estimator. However, this efficiency loss appears to be empirically minimal (see Zhao and Joe, 2005, Lele, 2006, Joe and Lee, 2009).[6] On the other hand, for models with complex dependence structure such as multivariate ordered-response model system of dimensionality more than 3, the full likelihood estimator has to be approximated using simulation techniques. Application of simulation techniques such as the maximum simulated likelihood (MSL) approach also leads to a loss in estimator efficiency (see McFadden and Train, 2000). Thus, it is of interest to compare the CML and MSL estimators in terms of asymptotic efficiency.

Earlier applications of the CML approach (and specifically the pairwise likelihood approach) to multivariate ordered-response systems include de Leon (2005) and Ferdous *et al.* (2010) in the context of cross-sectional multivariate ordered-response probit systems, and Varin and Vidoni (2006) and Varin and Czado (2010) in the context of panel multivariate ordered-response probit systems. Bhat *et al.* (2010b) also use a CML approach to estimate their multivariate ordered-response probit system in the context of a

[5] For definition of matrix **H**, matrix **J**, and vector $\boldsymbol{\theta}$, see Section 2.5.

[6] A handful of studies (see Hjort and Varin, 2008, Mardia *et al.*, 2009, Cox and Reid, 2004) have also theoretically examined the limiting normality properties of the CML approach, and compared the asymptotic variance matrices from this approach with the maximum likelihood approach. However, such a precise theoretical analysis is possible only for very simple models, and becomes much harder for models such as a multivariate ordered-response system.

spatially dependent ordered-response outcome variable. In this study, we do not use the high multivariate dimensionality of most of these earlier studies. Rather, we consider relatively lower multivariate dimensionality simulation situations, so that we are able to estimate the models using MSL techniques too. Specifically, we compare the performance of the composite marginal likelihood (CML) approach with the maximum-simulated likelihood (MSL) approach in 5-variate ordered-response situations. We use simulated data sets with known underlying model parameters to evaluate the two estimation approaches. The ability of the two approaches to recover model parameters is examined, as is the sampling variance and the simulation variance of parameters in the MSL approach relative to the sampling variance in the CML approach. The computational costs of the two approaches are also presented.

The rest of this chapter is structured as follows. In the next section, we present the structure of the cross-sectional multivariate ordered-response system. Section 3.3 discusses the simulation estimation methods (with an emphasis on the MSL approach). Section 3.4 presents the experimental design for the simulation experiments, while Section 3.5 discusses the results. Section 3.6 concludes the chapter by highlighting the important findings.

3.2 The Cross-Sectional Multivariate Ordered-Response Probit (CMOP) Formulation

Let q be an index for individuals ($q = 1, 2, \ldots, Q$, where Q denotes the total number of individuals in the data set), and let i be an index for the ordered-response variable ($i = 1, 2, \ldots, I$, where I denotes the total number of ordered-response variables for each individual). Let the observed discrete (ordinal) level for individual q and variable i be m_{qi} (m_{qi} may take one of K_i values; *i.e.*, $m_{qi} \in \{1, 2, \ldots, K_i\}$ for variable i). In the usual ordered-response framework notation, we write the latent propensity (y^*_{qi}) for each ordered-response variable as a function of relevant covariates and relate this latent propensity to the observed discrete level m_{qi} through threshold bounds (see McKelvey and Zavoina, 1975):

$$y_{qi}^{*} = \boldsymbol{\beta}_{\mathbf{i}}^{'}\mathbf{x}_{\mathbf{qi}} + \varepsilon_{qi}, y_{qi} = m_{qi} \text{ if } \theta_i^{m_{qi}-1} < y_{qi}^{*} < \theta_i^{m_{qi}}, \tag{3.1}$$

where $\mathbf{x}_{\mathbf{qi}}$ is a $(L \times 1)$ vector of exogenous variables (not including a constant), $\boldsymbol{\beta}_{\mathbf{i}}$ is a corresponding $(L \times 1)$ vector of coefficients to be estimated, ε_{qi} is a standard normal error term, and $\theta_i^{m_{qi}}$ is the upper bound threshold for discrete level m_{qi} of variable i ($\theta_i^0 < \theta_i^1 < \theta_i^2 ... < \theta_i^{K_i-1} < \theta_i^{K_i}$; $\theta_i^0 = -\infty$, $\theta_i^{K_i} = +\infty$ for each variable i). The ε_{qi} terms are assumed independent and identical across individuals (for each and all i). For identification reasons, the variance of each ε_{qi} term is normalized to 1. However, we allow correlation in the ε_{qi} terms across variables i for each individual q. Specifically, we define $\boldsymbol{\varepsilon}_{\mathbf{q}} = (\varepsilon_{q1}, \varepsilon_{q2}, \varepsilon_{q3}, \ldots, \varepsilon_{qI})'$. Then, $\boldsymbol{\varepsilon}_{\mathbf{q}}$ is multivariate normal distributed with a mean vector of zeros and a correlation matrix as follows:

$$\boldsymbol{\varepsilon}_{\mathbf{q}} \sim N\left[\begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho_{12} & \rho_{13} & \cdots & \rho_{1I} \\ \rho_{21} & 1 & \rho_{23} & \cdots & \rho_{2I} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ \rho_{I1} & \rho_{I2} & \rho_{I3} & \cdots & 1 \end{pmatrix}\right], \text{ or} \tag{3.2}$$

$$\boldsymbol{\varepsilon}_{\mathbf{q}} \sim N[\mathbf{0}, \boldsymbol{\Sigma}]$$

The off-diagonal terms of $\boldsymbol{\Sigma}$ capture the error covariance across the underlying latent continuous variables; that is, they capture the effects of common unobserved factors influencing the underlying latent propensities. These are the so-called polychoric correlations between pairs of observed ordered-response variables. Of course, if all the correlation parameters (*i.e.*, off-diagonal elements of $\boldsymbol{\Sigma}$), which we will stack into a vertical vector $\boldsymbol{\Omega}$, are identically zero, the model system in Equation (3.1) collapses to independent ordered-response probit models for each variable. Note that the diagonal elements of $\boldsymbol{\Sigma}$ are normalized to one for identification purposes.

The parameter vector (to be estimated) of the cross-sectional multivariate probit model is $\boldsymbol{\delta} = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2', ..., \boldsymbol{\beta}_I';\ \boldsymbol{\theta}_1', \boldsymbol{\theta}_2', ..., \boldsymbol{\theta}_I';\ \boldsymbol{\Omega}')'$, where $\boldsymbol{\theta}_\mathbf{i} = (\theta_i^1, \theta_i^2, ..., \theta_i^{K_i-1})'$ for $i = 1, 2, ..., I$. The likelihood function for individual q may be written as follows:

$$L_q(\boldsymbol{\delta}) = \Pr(y_{q1} = m_{q1}, y_{q2} = m_{q2}, ..., y_{qI} = m_{qI}) \qquad (3.3)$$

$$L_q(\boldsymbol{\delta}) = \int_{v_1=\theta_1^{m_{q1}-1}-\beta_1' x_{q1}}^{\theta_1^{m_{q1}}-\beta_1' x_{q1}} \int_{v_2=\theta_2^{m_{q2}-1}-\beta_2' x_{q2}}^{\theta_2^{m_{q2}}-\beta_2' x_{q2}} \cdots \int_{v_I=\theta_I^{m_{qI}-1}-\beta_I' x_{qI}}^{\theta_I^{m_{qI}}-\beta_I' x_{qI}} \phi_I(v_1, v_2, ..., v_I \mid \boldsymbol{\Omega}) dv_1 dv_2 ... dv_I,$$

where ϕ_I is the standard multivariate normal density function of dimension *I*. The likelihood function above involves an *I*-dimensional integral for each individual *q*.

3.3 Overview of Simulation Approaches

As indicated in Section 1.4 and Section 3.1, models that require integration of more than three dimensions in a multivariate ordered-response model are typically estimated using simulation approaches. Two broad simulation approaches may be identified in the literature for multivariate ordered-response modeling. One is based on a frequentist approach, while the other is based on a Bayesian approach. We provide an overview of these two approaches in the next two sections (Section 3.3.1 and Section 3.3.2).

3.3.1 The Frequentist Approach

In the context of a frequentist approach, Bhat and Srinivasan (2005) suggested a maximum simulated likelihood (MSL) method for evaluating the multi-dimensional integral in a cross-sectional multivariate ordered-response model system, using quasi-Monte Carlo simulation methods proposed by Bhat (2001, 2003). In their approach, Bhat and Srinivasan (BS) partition the overall error term into one component that is independent across dimensions and another mixing component that generates the correlation across dimensions. The estimation proceeds by conditioning on the error components that cause correlation effects, writing the resulting conditional joint probability of the observed ordinal levels across the many dimensions for each individual, and then integrating out the mixing correlated error components. An important issue is to ensure that the covariance matrix of the mixing error terms remains in a correlation form (for identification reasons) and is positive definite, which BS maintain by writing the likelihood function in terms of the elements of the Cholesky decomposed-matrix of the correlation matrix of the mixing normally distributed elements and parameterizing the diagonal elements of the Cholesky matrix to guarantee unit values along the diagonal. Another alternative and related MSL method would be to consider the correlation across error terms directly without partitioning the error terms into two components. This corresponds to the formulation in Equations (1) and (2) of the current study. Balia and Jones (2008) adopt such a formulation in their eight-dimensional multivariate probit model of lifestyles, morbidity, and mortality. They estimate their model using a Geweke-Hajivassiliou-Keane (GHK) simulator (the GHK simulator is discussed in more detail later in this study). However, it is not clear how they accommodated the identification sufficiency condition that the covariance matrix be a correlation matrix and be positive definite. But one can use the GHK simulator combined with BS's approach to ensure unit elements along the diagonal of the covariance matrix. Another MSL method that can be used to approximate the multivariate rectangular (*i.e.*, truncated) normal probabilities in the likelihood functions is based on the Genz-Bretz (GB) algorithm (Genz and Bretz, 1999). In concept, all these MSL methods can be extended to any number of correlated ordered-response outcomes, but numerical stability, convergence, and precision problems start surfacing as the number of dimensions increase.

3.3.2 The Bayesian Approach

Chen and Dey (2000), Herriges *et al.* (2008), Jeliazkov *et al.* (2008), and Hasegawa (2010) have considered an alternate estimation approach for the multivariate ordered-response system based on the posterior mode in an objective Bayesian approach. As in the frequentist case, a particular challenge in the Bayesian approach is to ensure that the covariance matrix of the parameters is in a correlation form, which is a sufficient condition for identification. Chen and Dey proposed a reparametization technique that involves a rescaling of the latent variables for each ordered-response variable by the reciprocal of the largest unknown threshold. Such an approach leads to an unrestricted covariance matrix of the re-scaled latent variables, allowing for the use of standard Markov Chain Monte Carlo (MCMC) techniques for estimation. In particular, the Bayesian approach is based on assuming prior distributions on the non-threshold parameters, reparameterizing the threshold parameters, imposing a standard conjugate prior on the reparameterized version of the error covariance matrix and a flat prior on the transformed threshold, obtaining an augmented posterior density using Bayes' Theorem for the reparameterized model, and fitting the model using a Markov Chain Monte Carlo (MCMC) method. Unfortunately, the method remains cumbersome, requires extensive simulation, and is time-consuming. Further, convergence assessment becomes difficult as the number of dimensions increase. For example, Muller and Czado (2005) used a Bayesian approach for their panel multivariate ordered-response model, and found that the standard MCMC method exhibits bad convergence properties. They proposed a more sophisticated group move multigrid MCMC technique, but this only adds to the already cumbersome nature of the simulation approach. In this regard, both the MSL and the Bayesian approach are "brute force" simulation techniques that are not very straightforward to implement and can create convergence assessment problems.

3.4 Estimation Methods Used in the Current Research

In the current study, we use the frequentist approach to compare the composite marginal likelihood (CML) approach with the simulation approaches. Frequentist approaches are widely used in the literature, and are included in several software programs that are readily available. Within the frequentist approach, we consider the Geweke-Hajivassiliou-Keane (GHK) simulator. We select the GHK simulator because it is among the most effective simulators for evaluating multivariate normal probabilities. Within the CML approach, we consider the pairwise marginal likelihood approach, because a significant volume of earlier applications of the CML approach as well as all the applications of the CML approach in the current dissertation are undertaken using the pairwise marginal likelihood approach.

3.4.1 Geweke-Hajivassiliou-Keane (GHK) Probability Simulator

The GHK is perhaps the most widely used probability simulator for integration of the multivariate normal density function, and is particularly well known in the context of the estimation of the multivariate unordered probit model. It is named after Geweke (1991), Hajivassiliou (Hajivassiliou and McFadden, 1998), and Keane (1990, 1994). Train (2003) provides an excellent and concise description of the GHK simulator in the context of the multivariate unordered probit model. In the current study, we adapt the GHK simulator to the case of the multivariate ordered-response probit model.

The GHK simulator is based on directly approximating the probability of a multivariate rectangular region of the multivariate normal density distribution. To apply the simulator, we first write the likelihood function in Equation (3.3) as follows:

$$L_q(\boldsymbol{\delta}) = \Pr(y_{q1} = m_{q1})\Pr(y_{q2} = m_{q2} \mid y_{q1} = m_{q1})\Pr(y_{q3} = m_{q3} \mid y_{q1} = m_{q1}, y_{q2} = m_{q2})\ldots$$

$$\ldots\Pr(y_{qI} = m_{qI} \mid y_{q1} = m_{q1}, y_{q2} = m_{q2}, \ldots, y_{qI-1} = m_{qI-1}) \qquad (3.4)$$

Also, write the error terms in Equation (3.2) as:

$$\begin{bmatrix} \varepsilon_{q1} \\ \varepsilon_{q2} \\ \vdots \\ \varepsilon_{qI} \end{bmatrix} = \begin{bmatrix} l_{11} & 0 & 0 & \cdots & 0 \\ l_{21} & l_{22} & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ l_{I1} & l_{I2} & l_{I3} & \cdots & l_{II} \end{bmatrix} \begin{bmatrix} v_{q1} \\ v_{q2} \\ \vdots \\ v_{qI} \end{bmatrix} \qquad (3.5)$$

$$\boldsymbol{\varepsilon}_{\mathbf{q}} = \mathbf{L}\mathbf{v}_{\mathbf{q}}$$

where **L** is the lower triangular Cholesky decomposition of the correlation matrix **Σ**, and $\mathbf{v_q}$ terms are independent and identically distributed standard normal deviates (*i.e.,* $\mathbf{v_q} \sim N[\mathbf{0},\mathbf{I_I}]$). Each (unconditional/conditional) probability term in Equation (3.4) can be written as follows:

$$\Pr(y_{q1}=m_{q1})=\Pr\left(\frac{\theta_1^{\mathrm{m_{q1}}-1}-\beta_1^{'}x_{q1}}{l_{11}}<v_{q1}<\frac{\theta_1^{\mathrm{m_{q1}}}-\beta_1^{'}x_{q1}}{l_{11}}\right)$$

$$\Pr(y_{q2}=m_{q2}\mid y_{q1}=m_{q1})=\Pr\left(\frac{\theta_2^{\mathrm{m_{q2}}-1}-\beta_2^{'}x_{q2}-l_{21}v_{q1}}{l_{22}}<v_{q2}<\frac{\theta_2^{\mathrm{m_{q2}}}-\beta_2^{'}x_{q2}-l_{21}v_{q1}}{l_{22}}\middle|\frac{\theta_1^{\mathrm{m_{q1}}-1}-\beta_1^{'}x_{q1}}{l_{11}}<v_{q1}<\frac{\theta_1^{\mathrm{m_{q1}}}-\beta_1^{'}x_{q1}}{l_{11}}\right)$$

$$\Pr(y_{q3}=m_{q3}\mid y_{q1}=m_{q1},y_{q2}=m_{q2})=\Pr\left(\begin{array}{l}\dfrac{\theta_3^{\mathrm{m_{q3}}-1}-\beta_3^{'}x_{q3}-l_{31}v_{q1}-l_{32}v_{q2}}{l_{33}}<v_{q3}<\dfrac{\theta_3^{\mathrm{m_{q3}}}-\beta_3^{'}x_{q3}-l_{31}v_{q1}-l_{32}v_{q2}}{l_{33}}\\ \left|\dfrac{\theta_1^{\mathrm{m_{q1}}-1}-\beta_1^{'}x_{q1}}{l_{11}}<v_{q1}<\dfrac{\theta_1^{\mathrm{m_{q1}}}-\beta_1^{'}x_{q1}}{l_{11}},\dfrac{\theta_2^{\mathrm{m_{q2}}-1}-\beta_2^{'}x_{q2}-l_{21}v_{q1}}{l_{22}}<v_{q2}<\dfrac{\theta_2^{\mathrm{m_{q2}}}-\beta_2^{'}x_{q2}-l_{21}v_{q1}}{l_{22}}\right.\end{array}\right)$$

$$\vdots \qquad\qquad (3.6)$$

$$\begin{array}{l}\Pr(y_{qI}=m_{qI}\mid y_{q1}=m_{q1},y_{q2}=m_{q2},\dots,y_{qI-1}=m_{qI-1})=\\ \Pr\left(\begin{array}{l}\dfrac{\theta_{\mathrm{I}}^{\mathrm{m_{qI}}-1}-\beta_I^{'}x_{qI}-l_{I1}v_{q1}-l_{I2}v_{q2}-\cdots-l_{I(I-1)}v_{q(I-1)}}{l_{II}}<v_{qI}<\dfrac{\theta_{\mathrm{I}}^{\mathrm{m_{qI}}}-\beta_I^{'}x_{qI}-l_{I1}v_{q1}-l_{I2}v_{q2}-\cdots-l_{I(I-1)}v_{q(I-1)}}{l_{II}}\\ \left|\dfrac{\theta_1^{\mathrm{m_{q1}}-1}-\beta_1^{'}x_{q1}}{l_{11}}<v_{q1}<\dfrac{\theta_1^{\mathrm{m_{q1}}}-\beta_1^{'}x_{q1}}{l_{11}},\dfrac{\theta_2^{\mathrm{m_{q2}}-1}-\beta_2^{'}x_{q2}-l_{21}v_{q1}}{l_{22}}<v_{q2}<\dfrac{\theta_2^{\mathrm{m_{q2}}}-\beta_2^{'}x_{q2}-l_{21}v_{q1}}{l_{22}},\cdots,\right.\\ \dfrac{\theta_{\mathrm{I\text{-}1}}^{\mathrm{m_{q(I\text{-}1)}}-1}-\beta_{I-1}^{'}x_{qI-1}-l_{(I-1)1}v_{q1}-l_{(I-1)2}v_{q2}-\cdots-l_{(I-1)(I-2)}v_{q(I-2)}}{l_{(I-1)(I-1)}}<v_{q(I-1)}<\dfrac{\theta_{\mathrm{I\text{-}1}}^{\mathrm{m_{q(I\text{-}1)}}}-\beta_{I-1}^{'}x_{qI-1}-l_{(I-1)1}v_{q1}-l_{(I-1)2}v_{q2}-\cdots-l_{(I-1)(I-2)}v_{q(I-2)}}{l_{(I-1)(I-1)}}\end{array}\right)\end{array}$$

The error terms v_{qi} are drawn *d* times (*d* = 1, 2, …, *D*) from the univariate standard normal distribution with the lower and upper bounds as above. To be precise, we use a randomized Halton draw procedure to generate the *d* realizations of v_{qi}, where we first generate standard Halton draw sequences of size $D\times 1$ for each individual for each dimension *i* (*i* = 1, 2,…, *I*), and then randomly shift the $D\times 1$ integration nodes using a random draw from the uniform distribution (see Bhat, 2001 and 2003 for a detailed discussion of the use of Halton sequences for discrete choice models). These random shifts are employed because we generate 10 different randomized Halton sequences of size $D\times 1$ to compute simulation error. Gauss code implementing the Halton draw procedure is available for download from the home page of Chandra Bhat at http://www.caee.utexas.edu/prof/bhat/halton.html. For each randomized Halton sequence, the uniform deviates are translated to truncated draws from the normal distribution for v_{qi} that respect the lower and upper truncation points (see, for example, Train, 2003; page 210). An unbiased estimator of the likelihood function for individual *q* is obtained as:

$$L_{GHK,q}(\boldsymbol{\delta}) = \frac{1}{D}\sum_{d=1}^{D} L_q^d(\boldsymbol{\delta}) \tag{3.7}$$

where $L_q^d(\boldsymbol{\delta})$ is an estimate of Equation (3.4) for simulation draw *d*. A consistent and asymptotically normal distributed GHK estimator $\hat{\boldsymbol{\delta}}_{GHK}$ is obtained by maximizing the logarithm of the simulated likelihood function $L_{GHK}(\boldsymbol{\delta}) = \prod_q L_{GHK,q}(\boldsymbol{\delta})$. The covariance matrix of parameters is estimated using the inverse of the sandwich information matrix (*i.e.*, using the robust asymptotic covariance matrix estimator associated with quasi-maximum likelihood; see McFadden and Train, 2000).

The likelihood function (and hence, the log-likelihood function) mentioned above is parameterized with respect to the parameters of the Cholesky decomposition matrix **L** rather than the parameters of the original covariance parameter **Σ**. This ensures the positive definiteness of **Σ**, but also raises two new issues: (1) the parameters of the Cholesky matrix **L** should be such that **Σ** should be a correlation matrix, and (2) the estimated parameter values (and asymptotic covariance matrix) do not correspond to **Σ**, but to **L**. The first issue is overcome by parameterizing the diagonal terms of **L** as shown below (see Bhat and Srinivasan, 2005):

$$\mathbf{L} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ l_{21} & \sqrt{1-l_{21}^2} & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ l_{I1} & l_{I2} & l_{I3} & \cdots & \sqrt{1-l_{I1}^2-l_{I2}^2\cdots-l_{I(I-1)}^2} \end{bmatrix} \tag{3.8}$$

The second issue is easily resolved by estimating **Σ** from the convergent values of the Cholesky decomposition parameters $(\mathbf{\Sigma} = \mathbf{LL}')$, and then running the parameter estimation procedure one more time with the likelihood function parameterized with the terms of **Σ**.

3.4.2 Pairwise Likelihood Approach

The pairwise marginal likelihood function for individual *q* may be written for the model as follows:

$$L_{CML,q}(\boldsymbol{\delta}) = \prod_{i=1}^{I-1} \prod_{g=i+1}^{I} \Pr(y_{qi} = m_{qi}, y_{qg} = m_{qg})$$

$$= \prod_{i=1}^{I-1} \prod_{g=i+1}^{I} \left[\begin{array}{l} \Phi_2\left(\theta_i^{m_{qi}} - \beta_i' x_{qi}, \theta_g^{m_{qg}} - \beta_g' x_{qg}, \rho_{ig}\right) - \Phi_2\left(\theta_i^{m_{qi}} - \beta_i' x_{qi}, \theta_g^{m_{qg}-1} - \beta_g' x_{qg}, \rho_{ig}\right) \\ -\Phi_2\left(\theta_i^{m_{qi}-1} \; - \beta_i' x_{qi}, \theta_g^{m_{qg}} - \beta_g' x_{qg}, \rho_{ig}\right) + \Phi_2\left(\theta_i^{m_{qi}-1} \; - \beta_i' x_{qi}, \theta_g^{m_{qg}-1} - \beta_g' x_{qg}, \rho_{ig}\right) \end{array} \right], \quad (3.9)$$

where $\Phi_2(.,.,\rho_{ig})$ is the standard bivariate normal cumulative distribution function with correlation ρ_{ig}. The pairwise marginal likelihood function is $L_{CML}(\boldsymbol{\delta}) = \prod_q L_{CML,q}(\boldsymbol{\delta})$.

As indicated in Chapter 2, the pairwise estimator $\hat{\boldsymbol{\delta}}_{CML}$ obtained by maximizing the logarithm of the pairwise marginal likelihood function with respect to the vector $\boldsymbol{\delta}$ is consistent and asymptotically normal distributed with asymptotic mean $\boldsymbol{\delta}$ and covariance matrix given by the inverse of Godambe's (1960) sandwich information matrix $\mathbf{G}(\boldsymbol{\delta})$ (see Zhao and Joe, 2005):

$\mathbf{V}_{CML}(\boldsymbol{\delta}) = [\mathbf{G}(\boldsymbol{\delta})]^{-1} = [\mathbf{H}(\boldsymbol{\delta})]^{-1}\mathbf{J}(\boldsymbol{\delta})[\mathbf{H}(\boldsymbol{\delta})]^{-1},$ where

$$\mathbf{H}(\boldsymbol{\delta}) = E\left[-\frac{\partial^2 \log L_{CML}(\boldsymbol{\delta})}{\partial\boldsymbol{\delta}\partial\boldsymbol{\delta}'} \right] \text{ and } \quad (3.10)$$

$$\mathbf{J}(\boldsymbol{\delta}) = E\left[\left(\frac{\partial \log L_{CML}(\boldsymbol{\delta})}{\partial\boldsymbol{\delta}} \right) \left(\frac{\partial \log L_{CML}(\boldsymbol{\delta})}{\partial\boldsymbol{\delta}'} \right) \right]$$

$\mathbf{H}(\boldsymbol{\delta})$ and $\mathbf{J}(\boldsymbol{\delta})$ can be estimated in a straightforward manner at the CML estimate $(\hat{\boldsymbol{\delta}}_{CML})$:

$$\hat{\mathbf{H}}(\hat{\boldsymbol{\delta}}) = -\left[\sum_{q=1}^{Q} \frac{\partial^2 \log L_{CML,q}(\boldsymbol{\delta})}{\partial\boldsymbol{\delta}\partial\boldsymbol{\delta}'}\right]_{\hat{\delta}}$$

$$= -\left[\sum_{q=1}^{Q}\sum_{i=1}^{I-1}\sum_{g=i+1}^{I} \frac{\partial^2 \log \Pr(y_{qi}=m_{qi}, y_{qg}=m_{qg})}{\partial\boldsymbol{\delta}\partial\boldsymbol{\delta}'}\right]_{\hat{\boldsymbol{\delta}}}, \text{ and} \qquad (3.11)$$

$$\hat{\mathbf{J}}(\hat{\boldsymbol{\delta}}) = \sum_{q=1}^{Q}\left[\left(\frac{\partial \log L_{CML,q}(\boldsymbol{\delta})}{\partial\boldsymbol{\delta}}\right)\left(\frac{\partial \log L_{CML,q}(\boldsymbol{\delta})}{\partial\boldsymbol{\delta}'}\right)\right]_{\hat{\boldsymbol{\delta}}}$$

In general, and as confirmed later in the simulation study, we expect that the ability to recover and pin down the parameters will be a little more difficult for the correlation parameters in $\boldsymbol{\Sigma}$ (when the correlations are low) than for the slope and threshold parameters, because the correlation parameters enter more non-linearly in the likelihood function.

3.4.3 Positive-Definiteness of the Implied Multivariate Correlation Matrix

A point that we have not discussed thus far in the CML approach is how to ensure the positive-definiteness of the symmetric correlation matrix $\boldsymbol{\Sigma}$. There are three ways that one can ensure the positive-definiteness of the $\boldsymbol{\Sigma}$ matrix. The first technique is to use Bhat and Srinivasan's technique of reparameterizing $\boldsymbol{\Sigma}$ through the Cholesky matrix, and then using these Cholesky-decomposed parameters as the ones to be estimated. Within the optimization procedure, one would then reconstruct the $\boldsymbol{\Sigma}$ matrix, and then "pick off" the appropriate elements of this matrix for the ρ_{ig} estimates at each iteration. This is probably the most straightforward and clean technique. The second technique is to undertake the estimation with a constrained optimization routine by requiring that the implied multivariate correlation matrix for any set of pairwise correlation estimates be positive definite. However, such a constrained routine can be extremely cumbersome. The third technique is to use an unconstrained optimization routine, but check for positive-definiteness of the implied multivariate correlation matrix. The easiest method within this third technique is to allow the estimation to proceed without checking for positive-definiteness at intermediate iterations, but check that the implied multivariate correlation matrix at the final converged pairwise marginal likelihood estimates is positive-definite. This will typically work for the case of a multivariate ordered-response model if one specifies exclusion restrictions (*i.e.*, zero correlations between some error terms) or correlation patterns that involve a lower dimension of effective parameters. Also, the number of correlation parameters in the full multivariate matrix explodes quickly as the dimensionality of the matrix increases, and estimating all these parameters becomes almost impossible (with any estimation technique) with the usual sample sizes available in practice. So, imposing exclusion restrictions is good econometric practice. However, if the above simple method of allowing the pairwise marginal estimation approach to proceed without checking for positive definiteness at intermediate iterations does not work, then one can check the implied multivariate correlation matrix for positive definiteness at each and every iteration. If the matrix is not positive-definite during a direction search at a given iteration, one can construct a "nearest" valid correlation matrix (see Ferdous *et al.*, 2010 for a discussion).

In the current study, we used an unconstrained optimization routine and ensured that the implied multivariate correlation matrix at convergence was positive-definite.

3.5 Experimental Design

To compare and evaluate the performance of the GHK and the CML estimation techniques, we undertake a simulation exercise for a multivariate ordered-response system with five ordinal variables. Further, to examine the potential impact of different correlation structures, we undertake the simulation exercise for a correlation structure with low correlations and another with high correlations. For each correlation structure, the experiment is carried out for 20 independent data sets with 1000 data points. Pre-specified values for the $\boldsymbol{\delta}$ vector are used to generate samples in each data set.

In the set-up, we use three exogenous variables in the latent equation for the first, third, and fifth ordered-response variables, and four exogenous variables for the second and fourth ordered-response variables. The values for each of the exogenous variables are drawn from a standard univariate normal distribution. A fixed coefficient vector $\boldsymbol{\beta}_{\mathbf{i}}$ $(i=1,2,3,4,5)$ is assumed on the variables, and the linear combination $\boldsymbol{\beta}'_{\mathbf{i}}\mathbf{x}_{\mathbf{qi}}$ (q = 1, 2, …, Q, Q = 1000; i = 1, 2, 3, 4, 5) is computed for each individual q and category i. Next, we generate Q five-variate realizations of the error term vector $(\varepsilon_{q1},\varepsilon_{q2},\varepsilon_{q3},\varepsilon_{q4},\varepsilon_{q5})$ with predefined positive-definite low error correlation structure ($\boldsymbol{\Sigma}_{\boldsymbol{low}}$) and high error correlation structure ($\boldsymbol{\Sigma}_{\boldsymbol{high}}$) as follows:

$$\boldsymbol{\Sigma}_{\boldsymbol{low}}=\begin{bmatrix} 1 & .30 & .20 & .22 & .15 \\ .30 & 1 & .25 & .30 & .12 \\ .20 & .25 & 1 & .27 & .20 \\ .22 & .30 & .27 & 1 & .25 \\ .15 & .12 & .20 & .25 & 1 \end{bmatrix}, \text{ and } \boldsymbol{\Sigma}_{\boldsymbol{high}}=\begin{bmatrix} 1 & .90 & .80 & .82 & .75 \\ .90 & 1 & .85 & .90 & .72 \\ .80 & .85 & 1 & .87 & .80 \\ .82 & .90 & .87 & 1 & .85 \\ .75 & .72 & .80 & .85 & 1 \end{bmatrix} \qquad (3.12)$$

The error term realization for each observation and each ordinal variable is then added to the systematic component $(\boldsymbol{\beta}'_{\mathbf{i}}\mathbf{x}_{\mathbf{qi}})$ as in Equation (3.1) and then translated to "observed" values of y_{qi} (0, 1, 2, ...) based on pre-specified threshold values. We assume four outcome levels for the first and the fifth ordered-response variables, three for the second and the fourth ordered-response variables, and five for the third ordered-response variable. Correspondingly, we pre-specify a vector of three threshold values [$(\boldsymbol{\theta}_{\mathbf{i}}=\theta_i^1,\theta_i^2,\theta_i^3)$, where i = 1 and 5] for the first and the fifth ordered-response equations, two for the second and the fourth equations [$(\boldsymbol{\theta}_{\mathbf{i}}=\theta_i^1,\theta_i^2)$, where i = 2 and 4], and four for the third ordered-response equation [$(\boldsymbol{\theta}_{\mathbf{i}}=\theta_i^1,\theta_i^2,\theta_i^3,\theta_i^4)$, where i = 3] .

As mentioned earlier, the above data generation process is undertaken 20 times with different realizations of the random error term to generate 20 different data sets. The CML estimation procedure is applied to each data set to estimate data-specific values of the $\boldsymbol{\delta}$ vector. The GHK simulator is applied to each dataset using 100 draws per individual of the randomized Halton sequence.[7] In addition, to assess and to quantify simulation variance, the GHK simulator is applied to each dataset 10 times with different (independent) randomized Halton draw sequences. This allows us to estimate simulation error by computing the standard deviation of estimated parameters among the 10 different GHK estimates on the same data set.

A few notes are in order here. We chose to use a setting with five ordinal variables so as to keep the computation time manageable for the maximum simulated likelihood estimations (going to, for example, 10 ordinal variables will increase computation time substantially, especially since more number of draws per individual may have to be used; note also that we have a total of 400 MSL estimation runs just for the five ordinal variable case in our experimental design). At the same time, a system of five ordinal variables leads to a large enough dimensionality of integration in the likelihood function where simulation estimation has to be used. Of course, one can examine the effect of varying the number of ordinal variables on the performance of the MSL and CML estimation approaches. In this study, we have chosen to focus on five dimensions, and examine the effects of varying correlation patterns and different model formulations corresponding to cross-sectional setting. A comparison with higher numbers of ordinal variables is left as a future exercise. However, in general, it is well known that MSL estimation gets more imprecise as the dimensionality of integration increases. On the other hand, our experience with CML estimation is that the performance does not degrade very much as the number of ordinal variables increases (see Ferdous *et al.*, 2010). Similarly, one can examine the effect of varying numbers of draws for MSL estimation. Our choice of 100 draws per individual was based on experimentation with different numbers of draws for the first data set. We found little improvement in ability to recover parameters or simulation variance beyond 100 draws per individual for this data

[7] Bhat (2001) used Halton sequence to estimate mixed logit models, and found that the simulation error in estimated parameters is lower with 100 Halton draws than with 1000 random draws (per individual). In our study, we carried out the GHK analysis of the multivariate ordered-response model with 100 randomized Halton draws as well as 500 random draws per individual, and found the 100 randomized Halton draws case to be much more accurate/efficient as well as much less time-consuming. So, we present only the results of the 100 randomized Halton draws case here.

set, and thus settled for 100 draws per individual for all data sets (as will be noted in the results section, the MSL estimation with 100 draws per individual indeed leads to negligible simulation variance). Finally, we chose to use three to four exogenous variables in our experimental design (rather than use a single exogenous variable) so that the resulting simulation data sets would be closer to realistic ones where multiple exogenous variables are employed.

3.6 Performance Comparison Between the MSL and CML Approaches

In this section, we first identify a number of performance measures and discuss how these are computed for the MSL approach and the CML approach. The subsequent sections present the simulation and computational results.

3.6.1 Performance Measures

As discussed earlier, we consider two correlation matrix patterns, one with low correlations and another with high correlations. The steps discussed below for computing performance measures are for a specific correlation matrix pattern.

MSL Approach

(1) Estimate the MSL parameters for each data set s (s = 1, 2, …, 20; *i.e.*, S = 20) and for each of 10 independent draws, and obtain the time to get the convergent values and the standard errors. Note combinations for which convergence is not achieved. Everything below refers to cases when convergence is achieved. Obtain the mean time for convergence (TMSL) and standard deviation of convergence time across the converged runs and across all data sets (the time to convergence includes the time to compute the covariance matrix of parameters and the corresponding parameter standard errors).

(2) For each data set s and draw combination, estimate the standard errors (s.e.) of parameters (using the sandwich estimator).

(3) For each data set *s*, compute the mean estimate for each model parameter across the draws. Label this as MED, and then take the mean of the MED values across the data sets to obtain **a mean estimate**. Compute the **absolute percentage bias** (APB) as: $APB = \left| \frac{\text{mean estimate - true value}}{\text{true value}} \right| \times 100$

(4) Compute the standard deviation of the MED values across the data sets and label this as the **finite sample standard error** (essentially, this is the empirical standard error).

(5) For each data set *s*, compute the median s.e. for each model parameter across the draws. Call this MSED, and then take the mean of the MSED values across the *S* data sets and label this as **the asymptotic standard error** (essentially this is the standard error of the distribution of the estimator as the sample size gets large). Note that we compute the median s.e. for each model parameter across the draws and label it as MSED rather than computing the mean s.e. for each model parameter across the draws. This is because, for some draws, the estimated standard errors turned out to be rather large relative to other independent standard error estimates for the same dataset. On closer inspection, this could be traced to the unreliability of the numeric Hessian used in the sandwich estimator computation. This is another bothersome issue with MSL – it is important to compute the covariance matrix using the sandwich estimator rather than using the inverse of the cross-product of the first derivatives (due to the simulation noise introduced when using a finite number of draws per individual in the MSL procedure; see McFadden and Train, 2000). Specifically, using the inverse of the cross-product of the first derivatives can substantially underestimate the covariance matrix. But coding the analytic Hessian (as part of computing the sandwich estimator) is extremely difficult, while using the numeric Hessian is very unreliable. Craig (2008) also alludes to this problem when he states that "(...) the randomness that is inherent in such methods [referring here to the GB algorithm, but applicable in general to MSL methods] is sometimes more than a minor nuisance." In particular, even when the log-likelihood

function is computed with good precision so that the simulation error in estimated parameters is very small, this is not always adequate to reliably compute the numerical Hessian. To do so, one will generally need to compute the log-likelihood with a substantial level of precision, which, however, would imply very high computational times even in low dimensionality situations. Finally, note that the mean asymptotic standard error is a theoretical approximation to the finite sample standard error, since, in practice, one would estimate a model on only one data set from the field.

(6) Next, for each data set *s*, compute the simulation standard deviation for each parameter as the standard deviation in the estimated values across the independent draws (about the MED value). Call this standard deviation as SIMMED. For each parameter, take the mean of SIMMED across the different data sets. Label this as the **simulation s.e.** for each parameter.

(7) For each parameter, compute a **simulation adjusted standard error** as follows: $\sqrt{(\text{asymptotic standard error})^2 + (\text{simulation standard error})^2}$

CML Approach

(1) Estimate the CML parameters for each data set *s* and obtain the time to get the convergent values (including the time to obtain the Godambe matrix-computed covariance matrix and corresponding standard errors). Determine the mean time for convergence (TCML) across the *S* data sets.[8]

(2) For each data set *s*, estimate the standard errors (s.e.) (using the Godambe estimator).

(3) Compute the **mean estimate** for each model parameter across the *R* data sets. Compute **absolute percentage bias** as in the MSL case.

[8] The CML estimator always converged in our simulations, unlike the MSL estimator.

(4) Compute the standard deviation of the CML parameter estimates across the data sets and label this as the **finite sample standard error** (essentially, this is the empirical standard error).

3.6.2 Simulation Results

Table 3.1a presents the results for the CMOP model with low correlations, and Table 3.1b presents the corresponding results for the CMOP model with high correlations. The results indicate that both the MSL and CML approaches recover the parameters extremely well, as can be observed by comparing the mean estimate of the parameters with the true values (see the column titled "parameter estimates"). In the low correlation case, the absolute percentage bias (APB) ranges from 0.03% to 15.95% (overall mean value of 2.21% - see last row of table under the column titled "absolute percentage bias") across parameters for the MSL approach, and from 0.00% to 12.34% (overall mean value of 1.92%) across parameters for the CML approach. In the high correlation case, the APB ranges from 0.02% to 5.72% (overall mean value of 1.22% - see last row of table under the column titled "absolute percentage bias") across parameters for the MSL approach, and from 0.00% to 6.34% (overall mean value of 1.28%) across parameters for the CML approach. These are incredibly good measures for the ability to recover parameter estimates, and indicate that both the MSL and CML perform about evenly in the context of bias. Further, the ability to recover parameters does not seem to be affected at all by whether there is low correlation or high correlation (in fact, the overall APB reduces from the low correlation case to the high correlation case). Interestingly, the absolute percentage bias values are generally much higher for the correlation (ρ) parameters than for the slope (β) and threshold (θ) parameters in the low correlation case, but the situation is exactly reversed in the high correlation case where the absolute percentage bias values are generally higher for the slope (β) and

Table 3.1a Evaluation of Ability to Recover "True" Parameters by the MSL and CML Approaches – With Low Error Correlation Structure

Parameter	True Value	MSL Approach						CML Approach				Relative Efficiency	
		Parameter Estimates		Standard Error Estimates				Parameter Estimates		Standard Error Estimates			
		Mean Estimate	Absolute Percentage Bias	Finite Sample Standard Error	Asymptotic Standard Error ($MASE_{MSL}$)	Simulation Standard Error	Simulation Adjusted Standard Error ($SASE_{MSL}$)	Mean Estimate	Absolute Percentage Bias	Finite Sample Standard Error	Asymptotic Standard Error ($MASE_{CML}$)	$\frac{MASE_{MSL}}{MASE_{CML}}$	$\frac{SASE_{MSL}}{MASE_{CML}}$
Coefficients													
β_{11}	0.5000	0.5167	3.34%	0.0481	0.0399	0.0014	0.0399	0.5021	0.43%	0.0448	0.0395	1.0109	1.0116
β_{21}	1.0000	1.0077	0.77%	0.0474	0.0492	0.0005	0.0492	1.0108	1.08%	0.0484	0.0482	1.0221	1.0222
β_{31}	0.2500	0.2501	0.06%	0.0445	0.0416	0.0010	0.0416	0.2568	2.73%	0.0252	0.0380	1.0957	1.0961
β_{12}	0.7500	0.7461	0.52%	0.0641	0.0501	0.0037	0.0503	0.7698	2.65%	0.0484	0.0487	1.0283	1.0311
β_{22}	1.0000	0.9984	0.16%	0.0477	0.0550	0.0015	0.0550	0.9990	0.10%	0.0503	0.0544	1.0100	1.0104
β_{32}	0.5000	0.4884	2.31%	0.0413	0.0433	0.0017	0.0434	0.5060	1.19%	0.0326	0.0455	0.9518	0.9526
β_{42}	0.2500	0.2605	4.19%	0.0372	0.0432	0.0006	0.0432	0.2582	3.30%	0.0363	0.0426	1.0149	1.0150
β_{13}	0.2500	0.2445	2.21%	0.0401	0.0346	0.0008	0.0346	0.2510	0.40%	0.0305	0.0342	1.0101	1.0104
β_{23}	0.5000	0.4967	0.66%	0.0420	0.0357	0.0021	0.0358	0.5063	1.25%	0.0337	0.0364	0.9815	0.9833
β_{33}	0.7500	0.7526	0.34%	0.0348	0.0386	0.0005	0.0386	0.7454	0.62%	0.0441	0.0389	0.9929	0.9930
β_{14}	0.7500	0.7593	1.24%	0.0530	0.0583	0.0008	0.0583	0.7562	0.83%	0.0600	0.0573	1.0183	1.0184
β_{24}	0.2500	0.2536	1.46%	0.0420	0.0486	0.0024	0.0487	0.2472	1.11%	0.0491	0.0483	1.0067	1.0079
β_{34}	1.0000	0.9976	0.24%	0.0832	0.0652	0.0017	0.0652	1.0131	1.31%	0.0643	0.0633	1.0298	1.0301
β_{44}	0.3000	0.2898	3.39%	0.0481	0.0508	0.0022	0.0508	0.3144	4.82%	0.0551	0.0498	1.0199	1.0208
β_{15}	0.4000	0.3946	1.34%	0.0333	0.0382	0.0014	0.0382	0.4097	2.42%	0.0300	0.0380	1.0055	1.0061
β_{25}	1.0000	0.9911	0.89%	0.0434	0.0475	0.0016	0.0475	0.9902	0.98%	0.0441	0.0458	1.0352	1.0358
β_{35}	0.6000	0.5987	0.22%	0.0322	0.0402	0.0007	0.0402	0.5898	1.69%	0.0407	0.0404	0.9959	0.9961
Correlation Coefficients													
ρ_{12}	0.3000	0.2857	4.76%	0.0496	0.0476	0.0020	0.0476	0.2977	0.77%	0.0591	0.0467	1.0174	1.0184
ρ_{13}	0.2000	0.2013	0.66%	0.0477	0.0409	0.0019	0.0410	0.2091	4.56%	0.0318	0.0401	1.0220	1.0231
ρ_{14}	0.2200	0.1919	12.76%	0.0535	0.0597	0.0035	0.0598	0.2313	5.13%	0.0636	0.0560	1.0664	1.0682
ρ_{15}	0.1500	0.1739	15.95%	0.0388	0.0439	0.0040	0.0441	0.1439	4.05%	0.0419	0.0431	1.0198	1.0239
ρ_{23}	0.2500	0.2414	3.46%	0.0546	0.0443	0.0040	0.0445	0.2523	0.92%	0.0408	0.0439	1.0092	1.0133
ρ_{24}	0.3000	0.2960	1.34%	0.0619	0.0631	0.0047	0.0633	0.3013	0.45%	0.0736	0.0610	1.0342	1.0372
ρ_{25}	0.1200	0.1117	6.94%	0.0676	0.0489	0.0044	0.0491	0.1348	12.34%	0.0581	0.0481	1.0154	1.0194
ρ_{34}	0.2700	0.2737	1.37%	0.0488	0.0515	0.0029	0.0516	0.2584	4.28%	0.0580	0.0510	1.0094	1.0110
ρ_{35}	0.2000	0.2052	2.62%	0.0434	0.0378	0.0022	0.0378	0.1936	3.22%	0.0438	0.0391	0.9662	0.9678
ρ_{45}	0.2500	0.2419	3.25%	0.0465	0.0533	0.0075	0.0538	0.2570	2.78%	0.0455	0.0536	0.9937	1.0034

Table 3.1a (Continued) Evaluation of Ability to Recover “True” Parameters by the MSL and CML Approaches – With Low Error Correlation Structure

Parameter	True Value	MSL Approach						CML Approach				Relative Efficiency	
		Parameter Estimates		Standard Error Estimates				Parameter Estimates		Standard Error Estimates			
		Mean Estimate	Absolute Percentage Bias	Finite Sample Standard Error	Asymptotic Standard Error $(MASE_{MSL})$	Simulation Standard Error	Simulation Adjusted Standard Error $(SASE_{MSL})$	Mean Estimate	Absolute Percentage Bias	Finite Sample Standard Error	Asymptotic Standard Error $(MASE_{CML})$	$\frac{MASE_{MSL}}{MASE_{CML}}$	$\frac{SASE_{MSL}}{MASE_{CML}}$
Threshold Parameters													
θ_1^1	-1.0000	-1.0172	1.72%	0.0587	0.0555	0.0007	0.0555	-1.0289	2.89%	0.0741	0.0561	0.9892	0.9893
θ_1^2	1.0000	0.9985	0.15%	0.0661	0.0554	0.0011	0.0554	1.0010	0.10%	0.0536	0.0551	1.0063	1.0065
θ_1^3	3.0000	2.9992	0.03%	0.0948	0.1285	0.0034	0.1285	2.9685	1.05%	0.1439	0.1250	1.0279	1.0282
θ_2^1	0.0000	-0.0172	-	0.0358	0.0481	0.0007	0.0481	-0.0015	-	0.0475	0.0493	0.9750	0.9751
θ_2^2	2.0000	1.9935	0.32%	0.0806	0.0831	0.0030	0.0831	2.0150	0.75%	0.0904	0.0850	0.9778	0.9784
θ_3^1	-2.0000	-2.0193	0.97%	0.0848	0.0781	0.0019	0.0781	-2.0238	1.19%	0.0892	0.0787	0.9920	0.9923
θ_3^2	-0.5000	-0.5173	3.47%	0.0464	0.0462	0.0005	0.0462	-0.4968	0.64%	0.0519	0.0465	0.9928	0.9928
θ_3^3	1.0000	0.9956	0.44%	0.0460	0.0516	0.0011	0.0516	1.0014	0.14%	0.0584	0.0523	0.9877	0.9879
θ_3^4	2.5000	2.4871	0.52%	0.0883	0.0981	0.0040	0.0982	2.5111	0.44%	0.0735	0.1002	0.9788	0.9796
θ_4^1	1.0000	0.9908	0.92%	0.0611	0.0615	0.0031	0.0616	1.0105	1.05%	0.0623	0.0625	0.9838	0.9851
θ_4^2	3.0000	3.0135	0.45%	0.1625	0.1395	0.0039	0.1396	2.9999	0.00%	0.1134	0.1347	1.0356	1.0360
θ_5^1	-1.5000	-1.5084	0.56%	0.0596	0.0651	0.0032	0.0652	-1.4805	1.30%	0.0821	0.0656	0.9925	0.9937
θ_5^2	0.5000	0.4925	1.50%	0.0504	0.0491	0.0017	0.0492	0.5072	1.44%	0.0380	0.0497	0.9897	0.9903
θ_5^3	2.0000	2.0201	1.01%	0.0899	0.0797	0.0017	0.0798	2.0049	0.24%	0.0722	0.0786	1.0151	1.0154
Overall mean value across parameters		-	2.21%	0.0566	0.0564	0.0022	0.0564	-	1.92%	0.0562	0.0559	1.0080	1.0092

Table 3.1b Evaluation of Ability to Recover "True" Parameters by the MSL and CML Approaches – With High Error Correlation Structure

Parameter	True Value	MSL Approach						CML Approach				Relative Efficiency	
		Parameter Estimates		Standard Error Estimates				Parameter Estimates		Standard Error Estimates			
		Mean Estimate	Absolute Percentage Bias	Finite Sample Standard Error	Asymptotic Standard Error ($MASE_{MSL}$)	Simulation Standard Error	Simulation Adjusted Standard Error ($SASE_{MSL}$)	Mean Estimate	Absolute Percentage Bias	Finite Sample Standard Error	Asymptotic Standard Error ($MASE_{CML}$)	$\frac{MASE_{MSL}}{MASE_{CML}}$	$\frac{SASE_{MSL}}{MASE_{CML}}$
Coefficients													
β_{11}	0.5000	0.5063	1.27%	0.0300	0.0294	0.0020	0.0294	0.5027	0.54%	0.0292	0.0317	0.9274	0.9294
β_{21}	1.0000	1.0089	0.89%	0.0410	0.0391	0.0026	0.0392	1.0087	0.87%	0.0479	0.0410	0.9538	0.9560
β_{31}	0.2500	0.2571	2.85%	0.0215	0.0288	0.0017	0.0289	0.2489	0.42%	0.0251	0.0290	0.9943	0.9961
β_{12}	0.7500	0.7596	1.27%	0.0495	0.0373	0.0028	0.0374	0.7699	2.65%	0.0396	0.0395	0.9451	0.9477
β_{22}	1.0000	1.0184	1.84%	0.0439	0.0436	0.0036	0.0437	1.0295	2.95%	0.0497	0.0463	0.9419	0.9451
β_{32}	0.5000	0.5009	0.17%	0.0343	0.0314	0.0023	0.0315	0.5220	4.39%	0.0282	0.0352	0.8931	0.8955
β_{42}	0.2500	0.2524	0.96%	0.0284	0.0294	0.0021	0.0294	0.2658	6.34%	0.0263	0.0315	0.9318	0.9343
β_{13}	0.2500	0.2473	1.08%	0.0244	0.0233	0.0015	0.0234	0.2605	4.18%	0.0269	0.0251	0.9274	0.9293
β_{23}	0.5000	0.5084	1.67%	0.0273	0.0256	0.0020	0.0256	0.5100	2.01%	0.0300	0.0277	0.9221	0.9248
β_{33}	0.7500	0.7498	0.02%	0.0302	0.0291	0.0019	0.0291	0.7572	0.96%	0.0365	0.0318	0.9150	0.9170
β_{14}	0.7500	0.7508	0.11%	0.0416	0.0419	0.0039	0.0420	0.7707	2.75%	0.0452	0.0450	0.9302	0.9341
β_{24}	0.2500	0.2407	3.70%	0.0311	0.0326	0.0033	0.0327	0.2480	0.80%	0.0234	0.0363	0.8977	0.9022
β_{34}	1.0000	1.0160	1.60%	0.0483	0.0489	0.0041	0.0491	1.0000	0.00%	0.0360	0.0513	0.9532	0.9566
β_{44}	0.3000	0.3172	5.72%	0.0481	0.0336	0.0028	0.0337	0.3049	1.62%	0.0423	0.0368	0.9133	0.9165
β_{15}	0.4000	0.3899	2.54%	0.0279	0.0286	0.0026	0.0288	0.4036	0.90%	0.0274	0.0301	0.9516	0.9554
β_{25}	1.0000	0.9875	1.25%	0.0365	0.0391	0.0036	0.0393	1.0008	0.08%	0.0452	0.0398	0.9821	0.9862
β_{35}	0.6000	0.5923	1.28%	0.0309	0.0316	0.0030	0.0317	0.6027	0.45%	0.0332	0.0329	0.9607	0.9649
Correlation Coefficients													
ρ_{12}	0.9000	0.8969	0.34%	0.0224	0.0177	0.0034	0.0180	0.9019	0.21%	0.0233	0.0183	0.9669	0.9845
ρ_{13}	0.8000	0.8041	0.51%	0.0174	0.0201	0.0035	0.0204	0.8009	0.11%	0.0195	0.0203	0.9874	1.0023
ρ_{14}	0.8200	0.8249	0.60%	0.0284	0.0265	0.0061	0.0272	0.8151	0.60%	0.0296	0.0297	0.8933	0.9165
ρ_{15}	0.7500	0.7536	0.49%	0.0248	0.0243	0.0046	0.0247	0.7501	0.01%	0.0242	0.0251	0.9678	0.9849
ρ_{23}	0.8500	0.8426	0.87%	0.0181	0.0190	0.0081	0.0207	0.8468	0.38%	0.0190	0.0198	0.9606	1.0438
ρ_{24}	0.9000	0.8842	1.75%	0.0187	0.0231	0.0097	0.0251	0.9023	0.26%	0.0289	0.0244	0.9484	1.0284
ρ_{25}	0.7200	0.7184	0.22%	0.0241	0.0280	0.0072	0.0289	0.7207	0.09%	0.0295	0.0301	0.9298	0.9600
ρ_{34}	0.8700	0.8724	0.27%	0.0176	0.0197	0.0036	0.0200	0.8644	0.65%	0.0208	0.0220	0.8972	0.9124
ρ_{35}	0.8000	0.7997	0.04%	0.0265	0.0191	0.0039	0.0195	0.7988	0.15%	0.0193	0.0198	0.9645	0.9848
ρ_{45}	0.8500	0.8421	0.93%	0.0242	0.0231	0.0128	0.0264	0.8576	0.89%	0.0192	0.0252	0.9156	1.0480

Table 3.1b (Continued) Evaluation of Ability to Recover "True" Parameters by the MSL and CML Approaches – With High Error Correlation Structure

Parameter	True Value	MSL Approach						CML Approach				Relative Efficiency	
		Parameter Estimates		Standard Error Estimates				Parameter Estimates		Standard Error Estimates			
		Mean Estimate	Absolute Percentage Bias	Finite Sample Standard Error	Asymptotic Standard Error $(MASE_{MSL})$	Simulation Standard Error	Simulation Adjusted Standard Error $(SASE_{MSL})$	Mean Estimate	Absolute Percentage Bias	Finite Sample Standard Error	Asymptotic Standard Error $(MASE_{CML})$	$\frac{MASE_{MSL}}{MASE_{CML}}$	$\frac{SASE_{MSL}}{MASE_{CML}}$
Threshold Parameters													
θ_1^1	-1.0000	-1.0110	1.10%	0.0600	0.0520	0.0023	0.0520	-1.0322	3.22%	0.0731	0.0545	0.9538	0.9548
θ_1^2	1.0000	0.9907	0.93%	0.0551	0.0515	0.0022	0.0515	1.0118	1.18%	0.0514	0.0528	0.9757	0.9766
θ_1^3	3.0000	3.0213	0.71%	0.0819	0.1177	0.0065	0.1179	2.9862	0.46%	0.1185	0.1188	0.9906	0.9921
θ_2^1	0.0000	-0.0234	-	0.0376	0.0435	0.0028	0.0436	0.0010	-	0.0418	0.0455	0.9572	0.9592
θ_2^2	2.0000	2.0089	0.44%	0.0859	0.0781	0.0066	0.0784	2.0371	1.86%	0.0949	0.0823	0.9491	0.9525
θ_3^1	-2.0000	-2.0266	1.33%	0.0838	0.0754	0.0060	0.0757	-2.0506	2.53%	0.0790	0.0776	0.9721	0.9752
θ_3^2	-0.5000	-0.5086	1.73%	0.0305	0.0440	0.0030	0.0441	-0.5090	1.80%	0.0378	0.0453	0.9702	0.9725
θ_3^3	1.0000	0.9917	0.83%	0.0516	0.0498	0.0035	0.0499	0.9987	0.13%	0.0569	0.0509	0.9774	0.9798
θ_3^4	2.5000	2.4890	0.44%	0.0750	0.0928	0.0066	0.0930	2.5148	0.59%	0.1144	0.0956	0.9699	0.9724
θ_4^1	1.0000	0.9976	0.24%	0.0574	0.0540	0.0050	0.0542	1.0255	2.55%	0.0656	0.0567	0.9526	0.9566
θ_4^2	3.0000	3.0101	0.34%	0.1107	0.1193	0.0125	0.1200	3.0048	0.16%	0.0960	0.1256	0.9498	0.9550
θ_5^1	-1.5000	-1.4875	0.84%	0.0694	0.0629	0.0056	0.0632	-1.5117	0.78%	0.0676	0.0649	0.9699	0.9737
θ_5^2	0.5000	0.4822	3.55%	0.0581	0.0465	0.0041	0.0467	0.4968	0.64%	0.0515	0.0472	0.9868	0.9906
θ_5^3	2.0000	1.9593	2.03%	0.0850	0.0741	0.0064	0.0744	2.0025	0.12%	0.0898	0.0761	0.9735	0.9771
Overall mean value across parameters		-	1.22%	0.0429	0.0428	0.0044	0.0432	-	1.28%	0.0455	0.0449	0.9493	0.9621

threshold (θ) parameters compared to the correlation (ρ) parameters (for both the MSL and CML approaches). This is perhaps because the correlation parameters enter more non-linearly in the likelihood function than the slope and threshold parameters, and need to be particularly strong before they start having any substantial effects on the log-likelihood function value. Essentially, the log-likelihood function tends to be relatively flat at low correlations, leading to more difficulty in accurately recovering the low correlation parameters. But, at high correlations, the log-likelihood function shifts considerably in value with small shifts in the correlation values, allowing them to be recovered accurately.[9]

The standard error measures provide several important insights. First, the finite sample standard error and asymptotic standard error values are quite close to one another, with very little difference in the overall mean values of these two columns (see last row). This holds for both the MSL and CML estimation approaches, and for both the low and high correlation cases, and confirms that the inverses of the sandwich information estimator (in the case of the MSL approach) and the Godambe information matrix estimator (in the case of the CML approach) recover the finite sample covariance matrices remarkably well. Second, the empirical and asymptotic standard errors for the threshold parameters are higher than for the slope and correlation parameters (for both the MSL and CML cases, and for both the low and high correlation cases). This is perhaps because the threshold parameters play a critical role in the partitioning of the underlying latent variable into ordinal outcomes (more so than the slope and correlation parameters), and so are somewhat more difficult to pin down. Third, a comparison of the standard errors across the low and high correlation cases reveals that the empirical and asymptotic standard errors are much lower for the correlation parameters in the latter case than in the former case. This reinforces the finding earlier that the correlation parameters

[9] One could argue that the higher absolute percentage bias values for the correlation parameters in the low correlation case compared to the high correlation case is simply an artifact of taking percentage differences from smaller base correlation values in the former case. However, the sum of the absolute values of the deviations between the mean estimate and the true value is 0.0722 for the low correlation case and 0.0488 for the high correlation case. Thus, the correlation values are indeed being recovered more accurately in the high correlation case compared to the low correlation case.

are much easier to recover at high values because of the considerable influence they have on the log-likelihood function at high values; consequently, not only are they recovered accurately, but they are also recovered more precisely at high correlation values. Fourth, across all parameters, there is a reduction in the empirical and asymptotic standard errors for both the MSL and CML cases between the low and high correlation cases (though the reduction is much more for the correlation parameters than for the non-correlation parameters). Fifth, the simulation error in the MSL approach is negligible to small. On average, based on the mean values in the last row of the table, the simulation error is about 3.9% of the sampling error for the low correlation case and 10.3% of the sampling error for the high correlation case. The higher simulation error for the high correlation case is not surprising, since we use the same number of Halton draws per individual in both the low and high correlation cases, and the multivariate integration is more involved with a high correlation matrix structure. Thus, as the levels of correlations increase, the evaluation of the multivariate normal integrals can be expected to become less precise at a given number of Halton draws per individual. However, overall, the results suggest that our MSL simulation procedure is well tuned, and that we are using adequate numbers of Halton draws per individual for the accurate evaluation of the log-likelihood function and the accurate estimation of the model parameters (this is also reflected in the negligible difference in the simulation-adjusted standard error and the mean asymptotic standard error of parameters in the MSL approach).

The final two columns of each of Tables 3.1a and 3.1b provide a relative efficiency factor between the MSL and CML approaches. The first of these columns provides the ratio of the asymptotic standard error of parameters from the MSL approach and the asymptotic standard error of the corresponding parameters from the CML approach. The second of these columns provides the ratio of the simulation-adjusted standard error of parameters from the MSL approach and the asymptotic standard error of parameters from the CML approach. As expected, the second column provides slightly higher values of efficiency, indicating that CML efficiency increases when one also considers the presence of simulation standard error in the MSL estimates. However, this efficiency increase is negligible in the current context because of very small MSL simulation error. The more important and interesting point though is that the relative efficiency of the CML approach is as good as the MSL approach in the low correlation case. This is different from the relative efficiency results obtained in Renard *et al.* (2004), Zhao and Joe (2005), and Kuk and Nott (2000) in other model contexts, where the CML has been shown to lose efficiency relative to a maximum likelihood approach. However, note that all these other earlier studies focus on a comparison of a CML approach vis-à-vis a maximum likelihood (ML) approach, while, in our setting, we must resort to MSL to approximate the likelihood function. To our knowledge, this is the first comparison of the CML approach to an MSL approach, applicable to situations when the full information maximum likelihood estimator cannot be evaluated analytically. In this regard, it is not clear that the earlier theoretical result that the difference between the asymptotic covariance matrix of the CML estimator (obtained as the inverse of the Godambe matrix) and of the ML estimator (obtained as the inverse of the cross-product matrix of derivatives) should be positive semi-definite would extend to our case because the asymptotic covariance of MSL is computed as the inverse of the sandwich information matrix.[10] Basically, the presence of simulation noise, even if very small in the estimates of the parameters as in our case, can lead to a significant drop in the amount of information available in the sandwich matrix, resulting in increased standard errors of parameters when using MSL. Our results regarding the efficiency of individual parameters suggests that any reduction in efficiency of the CML (because of using only pairwise likelihoods rather than the full likelihood) is balanced by the reduction in efficiency because of using MSL rather than ML, so that there is effectively no loss in

[10] McFadden and Train (2000) indicate, in their use of independent number of random draws across observations, that the difference between the asymptotic covariance matrix of the MSL estimator obtained as the inverse of the sandwich information matrix and the asymptotic covariance matrix of the MSL estimator obtained as the inverse of the cross-product of first derivatives should be positive definite for finite number of draws per observation. Consequently, for the case of independent random draws across observations, the relationship between the MSL sandwich covariance matrix estimator and the CML Godambe covariance matrix is unclear. The situation gets even more unclear in our case because of the use of Halton or Lattice point draws that are not based on independent random draws across observations.

asymptotic efficiency in using the CML approach (relative to the MSL approach) in the CMOP model for low correlation. However, for the high correlation case, the MSL does provide slightly better efficiency than the CML. However, even in this case, the relative efficiency of parameters in the CML approach ranges between 90%-99% (mean of 95%) of the efficiency of the MSL approach, without considering simulation standard error. When considering simulation error, the relative efficiency of the CML approach is even better at about 96% of the MSL efficiency (on average across all parameters). Overall, there is little to no drop in efficiency because of the use of the CML approach in the cross-sectional multivariate ordered-response probit model system context.

3.6.3 Non-Convergence and Computational Time

The simulation estimation of multivariate ordered-response model can involve numerical instability because of possible unstable operations such as large matrix inversions and imprecision in the computation of the Hessian. This can lead to convergence problems. On the other hand, the CML approach is a straightforward approach that should be easy to implement and should not have any convergence-related problems. In the current empirical study, we classified any estimation run that had not converged in 5 hours as being non-convergent.

We computed the non-convergence rate for the MSL approach in terms of the starting seeds that led to failure in a complete estimation of 10 simulation runs (using different randomized Halton sequences) for each data set. If a particular starting seed led to failure in convergence for any of the 10 simulation runs, that seed was classified as a failed seed. Otherwise, the seed was classified as a successful seed. This procedure was applied for each of the 20 data sets generated for each of the low and high correlation matrix structures until we had a successful seed.[11] The non-convergence rate was then

[11] Note that we use the terminology "successful seed" to simply denote if the starting seed led to success in a complete estimation of the 10 simulation runs. In MSL estimation, it is not uncommon to obtain non-convergence (because of a number of reasons) for some sets of random sequences. There is, however, nothing specific to be learned here in terms of what starting seeds are likely to be successful and what starting seeds are likely to be unsuccessful. The intent is to use the terminology "successful seed" simply as a measure of non-convergence rates.

computed as the number of failed seeds divided by the total number of seeds considered. Note that this would be a good reflection of non-convergence rates if the analyst ran the simulation multiple times on a single data set to recognize simulation noise in statistical inferences. The results indicated a non-convergence rate of 28.5% for the low correlation case and 35.5% for the high correlation case. For both the low and high correlation cases, we always obtained convergence with the CML approach.

Next, we examined the time to convergence per converged estimation run for the MSL and CML procedures (the time to convergence included the time to compute the standard error of parameters). For the MSL approach, we had a very well-tuned and efficient procedure with an analytic gradient (written in Gauss matrix programming language). The CML procedure, which is very easy to code relative to the MSL, was also undertaken in the Gauss language. For both approaches we used naïve independent probit starting values. The estimations were carried out on a desktop machine.

Here, we only provide a relative computational time factor (RCTF), computed as the mean time needed for an MSL run divided by the mean time needed for a CML run. In addition, we present the standard deviation of the run times as a percentage of mean run time (SDR) for the MSL and CML estimations. The RCTF for the case of the low correlation matrix is 18, and for the case of the high correlation matrix is 40. The substantially higher RCTF for the high correlation case is because of an increase in the mean MSL time between the low and high correlation cases; the mean CML time hardly changed. The MSL SDR for the low correlation case is 30% and for the high correlation case is 47%, while the CML SDR is about 6% for both the low and high correlation cases. The computation time results do very clearly indicate the advantage of the CML over the MSL approach – the CML approach estimates parameters in much less time than the MSL, and the stability in the CML computation time is substantially higher than the stability in the MSL computation times. As the number of ordered-response outcomes increase, one can only expect a further increase in the computational time advantage of the CML over the MSL estimation approach.

3.7 Conclusions

This study compared the performance of the composite marginal likelihood (CML) approach with the maximum-simulated likelihood (MSL) approach in multivariate ordered-response situations. We used simulated data sets with known underlying model parameters to evaluate the two estimation approaches in the context of a cross-sectional ordered-response setting. The ability of the two approaches to recover model parameters was examined, as was the sampling variance and the simulation variance of parameters in the MSL approach relative to the sampling variance in the CML approach. The computational costs of the two approaches were also presented.

Overall, the simulation results demonstrate the ability of the Composite Marginal Likelihood (CML) approach to recover the parameters in a multivariate ordered-response choice model context, independent of the correlation structure. In addition, the CML approach recovers parameters as well as the MSL estimation approach in the simulation contexts used in the current study, while also doing so at a substantially reduced computational cost and improved computational stability. Further, any reduction in the efficiency of the CML approach relative to the MSL approach is in the range of non-existent to small. All these factors, combined with the conceptual and implementation simplicity of the CML approach, makes it a promising and simple approach not only for the multivariate ordered-response model considered here but also for other analytically-intractable econometric models. Also, as the dimensionality of the model explodes, the CML approach remains practical and feasible, while the MSL approach becomes impractical and/or infeasible.

Part II

Chapter 4

A Multivariate Ordered-Response Model System for Adults' Weekday Activity Episode Generation by Activity Purpose and Social Context

4.1 Introduction

4.1.1 Motivation

The emphasis of the activity-based approach to travel modeling is on understanding the activity participation characteristics of individuals within the context of their demographic attributes, activity-travel environment, and social interactions. In the activity-based approach, activity episodes rather than trip episodes take the center stage, with the focus being on activity episode generation and scheduling over a specified time period (Jones *et al.*, 1990, Bhat and Koppelman, 1999, Pendyala and Goulias, 2002, Arentze and Timmermans, 2004, and Pinjari and Bhat, 2011 provide extensive reviews of the activity-based approach). Several operational analytic frameworks for this activity analysis approach have also been formulated, and many metropolitan areas in the U.S. have implemented these frameworks (see Pinjari *et al.*, 2008 for a recent review). These frameworks have focused on a "typical" weekday frame of analysis, and follow a general structure where out-of-home work-related decisions (employed or not, duration of work, location of work, and timing of work) are modeled first followed by the generation and scheduling of out-of-home non-work episodes (in the rest of this study, we will use the term "non-work episodes" to refer to out-of-home non-work episodes).

The generation and scheduling of non-work episodes entails the determination of the number of non-work episodes by purpose, along with various attributes of each episode and the sequencing of these non-work episodes relative to work and in-home episodes. In the context of episode attributes, one dimension that has been receiving substantial attention recently is the "with whom" dimension (or the social context). This is motivated by the recognition that individuals usually do not make their activity engagement decisions in isolation. For instance, within a household, an individual’s activity participation decisions are likely to be dependent on other members of the household because of the possible sharing of household maintenance responsibilities, joint activity participation in discretionary activities, and pick-up/drop-off of household members with restricted mobility (Gleibe and Koppelman, 2002, Kapur and Bhat, 2007). In a similar vein, outside the confines of the household, an individual’s activity participation might be influenced by non-household members because of car-pooling arrangements, social engagements, and joint recreational pursuits. In fact, Srinivasan and Bhat (2008), in their descriptive study of activity patterns, found that about 30% of individuals undertake one or more out-of-home (OH) activity episodes with household members on weekdays, and about 50% pursue OH activity episodes with non-household companions on weekdays. These interactions in activity decisions across household and non-household members are important to consider to accurately predict activity-travel patterns. For instance, the spatial and temporal joint participation in dinner at a restaurant of a husband and a wife are necessarily linked. Thus, considering the husband’s and wife’s activity-travel patterns independently without maintaining the linkage in time and space in their patterns will necessarily result in less accurate activity travel pattern predictions for each one of them. Further, there is a certain level of rigidity in such joint activity participations (since such participations necessitate the synchronization of the schedules of multiple individuals in time and space), because of which the responsiveness to transportation control measures such as pricing schemes may be less than what would be predicted if each individual were considered in isolation (see Vovsha and Bradley, 2006 and Timmermans and Zhang, 2009 for extensive discussions of the importance of considering inter-individual interactions for accurately evaluating land-use and transportation policy actions).

To be sure, several recent studies have focused on explicitly accommodating inter-individual interactions in activity-travel modeling. The reader is referred to a special issue of *Transportation* edited by Bhat and Pendyala (2005), as well as a special issue of

Transportation Research Part B edited by Timmermans and Zhang (2009), for recent papers on this topic. While these and other earlier studies have contributed in very important ways, they focus on intra-household interactions, and mostly on the interactions between the household heads (see, for example, Wen and Koppelman, 1999, Scott and Kanaroglou, 2002, Meka *et al.*, 2002, Srinivasan and Bhat, 2005, and Kato and Matsumoto, 2009). On the other hand, as discussed earlier in this study, there is a significant amount of activity episode participations in the wider social network beyond the household (see also Goulias and Kim, 2005, Axhausen, 2005, Arentze and Timmermans, 2008, and Carrasco and Miller, 2009). Many earlier intra-household interaction studies in the literature also confine their attention to the single activity category of maintenance-oriented activities (see Srinivasan and Athuru, 2005, and Wang and Li, 2009). But, as indicated by PBQD (2000), over 75% of non-work episodes on a typical weekday are for discretionary purposes and, as pointed out by Srinivasan and Bhat (2008), a high percentage of these discretionary episodes involve one or more companions. This suggests the important need to consider inter-individual interactions in discretionary activity too (and not just in maintenance-oriented activity). Further, a significant fraction of existing studies on inter-individual interactions focus on daily *time allocations* or *joint time-use* in activities over a certain time period (an extensive review of these time allocation/time-use studies is provided in Vovsha *et al.*, 2003, and Kato and Matsumoto, 2009). This is also true of the recent studies by Bhat and colleagues (Kapur and Bhat, 2007, Sener and Bhat, 2007) that use the multiple discrete-continuous extreme value (MDCEV) model to examine household and non-household companionship arrangement for each of several types of activities. While providing important insights, these studies of daily time-use do not directly translate to information regarding out-of-home episodes. On the other hand, it is the scheduling and sequencing of out-of-home episodes that get manifested in the form of travel patterns (Doherty and Axhausen, 1999, Scott and Kanaroglou, 2002, Vovsha *et al.*, 2003). Finally, even among those studies that consider inter-individual interactions at an episode level, almost all of them have adopted a framework that first generates activity episodes by activity purpose, and subsequently "assigns" each of these purpose-specific episodes to a certain accompaniment type (for example, alone versus joint), typically using a discrete choice model (see, for example, Wen and Koppelman, 1999, Gliebe and Koppelman, 2002, and Bradley and Vovsha, 2005). Unfortunately, such a sequential framework cannot accommodate general patterns of observed and unobserved variable effects that are specific to each activity purpose-accompaniment type combination (see also Scott and Kanaroglou, 2002).

4.1.2 The Current Study

The objective of the current study, motivated by the discussion above, is to propose and estimate a joint modeling system for adult individuals' (aged 15 years or over) non-work activity episodes (or simply "episodes" from hereon) by purpose that also explicitly incorporates companionship arrangement information. The six activity purpose categories considered in the study are: (1) family care (including child care), (2) maintenance shopping (grocery shopping, purchasing gas/food, and banking), (3) non-maintenance shopping (window shopping, cloth shopping, electronics shopping, *etc.*), (4) meals, (5) physically active recreation (sports, exercise, walking, bicycling, *etc.*), and (6) physically inactive recreation (social, relaxing, movies, and attending religious/cultural/sports events).[12] The companionship arrangement for episodes is considered in five categories: (1) alone, (2) only family (including children, spouse, and unmarried partner), (3) only relatives (parents, siblings, grandchild, *etc.*), (4) only friends (including friends, colleagues, neighbors, co-workers, peers, and other acquaintances), and (5) mixed company (a combination of family, extended family, and friends).[13] The total number of

[12] There is obviously some subjectivity in the classification adopted here, though the overall consideration was to accommodate differences between the disaggregate activity purposes along such contextual dimensions as location of participation, physical intensity level, duration of participation, amount of structure in activity planning, and company type of participation (see Srinivasan and Bhat, 2005).

[13] While we consider the companionship arrangement for episodes, the reader will note that we still consider the generation of episodes at the individual-level. Future efforts should consider the generation of episodes at a higher level, such as a household-level or a neighborhood level, so that there is consistency in activity episode generation across individuals. Thus, for example, if a husband has a joint out-of-home (OH) activity episode with his wife, it must also be true that the wife has a joint OH activity episode with her husband.

activity purpose-companionship type categories is 30, and the model system developed here jointly considers the number of episodes in each of these 30 categories. The data used in the empirical analysis is drawn from the American Time Use Survey (ATUS), which collects detailed individual-level activity information for one day from a randomly selected adult (15 years or older) in each of a subset of households responding to the Current Population Survey (CPS).

The study uses a multivariate ordered-response model system for analyzing the number of episodes of each activity purpose-companionship type. In this system, we allow dependence between the number of episodes of different purpose-companionship types due to both observed exogenous variables as well as unobserved factors. The inclusion of dependence generated by unobserved factors allows *complementarity* and *substitution* effects in activity participation decisions (even after controlling for observed effects). For instance, individuals who are "go-getters" and "dynamic" in their lifestyle may have a higher participation propensity in sports-type activities ("physically active recreation") and also in cultural/social activities ("physically inactive recreation"). This would constitute a complementary relationship between these two activity purpose categories. Similarly, individuals who are "sociable" may be more likely to participate in activity episodes with friends, but not alone. This represents a substitution relationship in the company types of 'friends" and "alone". Besides, the presence of common unobserved factors among combination categories that share the same activity purpose or that share the same companionship type can also generate complementary effects. Thus, an individual who is "sociable" by personality may have a higher propensity to participate in dining out-with friends as well as a higher propensity to participate in physically-inactive recreation with friends. Overall, the extent of complementary and substitution relationships may be specific to the combinations of activity purpose category and company type, which is the general case modeled in the current study.

The econometric challenge in estimating a joint multivariate ordered-response system with a large number of categories is dealt with by applying the technique of composite marginal likelihood approach. The rest of the chapter is organized as follows.

Section 4.2 presents the model structure. Section 4.3 summarizes the data source and sample preparation procedure. Section 4.4 discusses the estimated results and the final section concludes the chapter by summarizing the salient features and findings of the study.

4.2 The Model Structure

4.2.1 Background

Employing an ordered-response system in the current context allows us the use of a general covariance matrix for the underlying latent variables, which translates to a flexible correlation pattern among the observed count outcomes (number of episodes across purpose types and companionship types in the current case). On the other hand, the traditional approach in the econometric literature to address correlated counts is to start with a Poisson or negative binomial distribution for each univariate count and add a random component to the conditional mean specification. If these random components are allowed to be correlated across equations, the net result is a mixed count model that allows correlation across outcomes. Such a model can be estimated using classical or Bayesian simulation techniques (Egan and Herriges, 2006, Chib and Winkelmann, 2001). An important problem with this approach, however, is that the use of the Poisson or negative binomial distribution as the underlying kernel for mixing restricts "the amount of probability mass that can be accommodated at any one point" (see Herriges *et al.*, 2008). Thus, in cases with a high fraction of '0' values, as in the current empirical context of the number of episodes in each activity purpose-companionship type combination, the count mixing models are not able to provide good predictions. The alternative of adding zero-inflated approaches to accommodate the high number of '0' values, while easy to undertake in a univariate count model, becomes difficult in the multivariate count case.

Of course, the use of an ordered-response system for count outcomes is certainly not new in the transportation literature. In fact, it has a long history of use for modeling such travel count dimensions as household car ownership levels (Kitamura, 1987, 1988, Golob and van Wissen, 1989, Golob, 1990, Bhat and Guo, 2007) and trip generation/stop-making (see Meurs, 1989, Agyemang-Duah *et al.*, 1995, Agyemang-Duah and Hall, 1997, Bhat, 1999, Bricka and Bhat, 2006, and Carrasco and Miller, 2009 to list just a few). While the traditional ordered-response model was initially developed for the case of ordinal responses, and while count outcomes are cardinal, this distinction is really irrelevant for the use of the ordered-response system for count outcomes. This is particularly the case when the count outcome takes few discrete values, as in the current empirical case, but is also not much of an issue when the count outcome takes a large number of possible values. A perceived problem in the latter case may be that the ordered-response model entails the estimation of *K*-1 threshold values that horizontally partition the underlying continuous variable to map into the observed count values, where *K* is the largest possible count value. But, as has been demonstrated by Meyer (1990), there is little loss of efficiency due to the estimation of a large number of thresholds in the ordered-response model structure. As long as there are even a few observations in each of the *K* categories under consideration, it is straightforward to estimate the ordered-response structure.

The ordered-response applications in the transportation literature discussed above all focus on a univariate count outcome. Three earlier multivariate count studies using a multivariate ordered-response structure that are directly relevant to the current study are Scott and Kanaroglou (2002), Bhat and Srinivasan (2005), and Herriges *et al.* (2008). These are discussed in turn below.

Scott and Kanaroglou use a trivariate normal distribution for the underlying latent continuous variables for three count outcomes, which correspond to the daily number of non-work episodes in couple households made by the male head, the female head, and jointly by both the heads. This leads to a trivariate integral for the probability expression for each household, which can be computed in a straightforward way using trivariate cumulative normal distribution functions. The restriction to three outcomes obviates the need for simulation, but also constrains the authors to consider all non-work episodes together without differentiating between activity types. Besides, the interaction in activity participation is confined to the household heads.

Bhat and Srinivasan appear to be the first to have proposed a modeling system and estimation approach that can conceptually accommodate any number of count outcomes. The authors use a logistic error term in each univariate ordered-response specification, and then also add a normally distributed mixing error term in the latent continuous equation. By allowing the mixing terms to be distributed multivariate normal, they effectively generate a flexible correlation structure across the outcome categories. They use a maximum simulated likelihood approach for evaluating the multi-dimensional integral in the resulting probability expression, using quasi-Monte Carlo simulation methods proposed by Bhat (2001; 2003). In addition, they develop a method to parameterize the likelihood function in terms of the elements of the Cholesky decomposed-matrix of the correlation matrix of the mixing normally distributed elements to ensure the positive definiteness of the matrix, and further parameterize the diagonal elements of the Cholesky matrix to guarantee unit values along the diagonal. Bhat and Srinivasan apply their model system to analyze the number of episodes of participation of individuals in seven different activity purposes, but they do not focus on accompaniment type. While their simulation approach can be extended in principle to any number of count outcomes, numerical stability, convergence, and precision problems start surfacing as the number of dimensions increase.

Herriges *et al.* (2008) recently have proposed an alternate estimation approach for the multivariate ordered-response system based on the posterior mode in an objective Bayesian approach as in Jeliazkov *et al.* (2008).[14] The approach of Herriges *et al.* (2008) is based on assuming prior distributions on the non-threshold parameters, reparameterizing the threshold parameters, imposing a standard conjugate prior on the reparameterized version of the error covariance matrix and a flat prior on the transformed

[14] It is interesting that Herriges *et al.* appear to be "discovering" the use of an ordered-response structure for count outcomes, while such a structure has in fact been used extensively in the past for count outcomes in the transportation literature. Further, Herriges *et al.* do not seem to have been aware of the work of Bhat and Srinivasan (2005), which develops a frequentist inference approach for correlated counts.

threshold, obtaining an augmented posterior density using Baye's Theorem for the reparameterized model, and fitting the model using a Markov Chain Monte Carlo (MCMC) method. Unfortunately, the method remains very cumbersome, requires extensive simulation, and is time-consuming. Further, convergence assessment becomes very difficult as the number of dimensions increase. In this regard, both the MSL and the Bayesian approach are "brute force" simulation techniques that are not straightforward to implement and can create convergence assessment problems. Herriges *et al.* apply their Bayesian estimation approach to examine the annual number of trips made by Iowa households to each of 29 lakes in the state.

In the current study, we consider and use a third inference approach – the Composite Marginal Likelihood (CML) approach. In the next sections, we discuss the mathematical formulation of the model and the composite marginal likelihood function (*i.e.*, the pairwaise marginal likelihood function).

4.2.2 Mathematical Formulation

Let q be an index for individuals (q = 1, 2, …, Q), and let i be the index for episode category (i = 1, 2, …, I, where I denotes the total number of episode categories for each individual; in the current study, I = 30). Let the number of episode count values for category i be K_i + 1 (*i.e.*, the discrete levels, indexed by k, belong in {0, 1, 2, …, K_i} for category i). In the usual ordered-response framework notation, we write the latent propensity (y^*_{qi}) for each episode category as a function of relevant covariates and relate this latent propensity to the observed count outcome (y_{qi}) through threshold bounds (see McKelvey and Zavoina, 1975).[15]

$$y^*_{qi} = \boldsymbol{\beta}'_{\mathbf{i}}\mathbf{x}_{\mathbf{qi}} + \varepsilon_{qi}, y_{qi} = k \text{ if } \theta^k_i < y^*_{qi} < \theta^{k+1}_i, \tag{4.1}$$

[15] Note that the model structure presented in this section is identical to the model structure presented in the previous chapter. However, the notations and symbols used to specify the model have different interpretations, since the context of the two studies are different. Thus, it is convenient to replicate the model system in the current empirical context.

where $\mathbf{x_{qi}}$ is a $(L\times 1)$ vector of exogenous variables (not including a constant), $\boldsymbol{\beta}_\mathbf{i}$ is a corresponding $(L\times 1)$ vector of coefficients to be estimated, ε_{qi} is a standard normal error term, and θ_i^k is the lower bound threshold for count level k of episode category i $(\theta_i^0 < \theta_i^1 < \theta_i^2 ... < \theta_i^{K_i+1};\ \theta_i^0 = -\infty,\ \theta_i^{K_i+1} = +\infty$ for each category i). The ε_{qi} terms are assumed independent and identical across individuals (for each and all *i*). For identification reasons, the variance of each ε_{qi} term is normalized to 1. However, we allow correlation in the ε_{qi} terms across episode categories *i* for each individual *q*. Specifically, define $\boldsymbol{\varepsilon}_\mathbf{q} = (\varepsilon_{q1}, \varepsilon_{q2}, \varepsilon_{q3}, \ldots, \varepsilon_{qI})'$. Then, $\boldsymbol{\varepsilon}_\mathbf{q}$ is multivariate normal distributed with a mean vector of zeros and a correlation matrix as follows:

$$\boldsymbol{\varepsilon}_\mathbf{q} \sim N\left[\begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho_{12} & \rho_{13} & \cdots & \rho_{1I} \\ \rho_{21} & 1 & \rho_{23} & \cdots & \rho_{2I} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ \rho_{I1} & \rho_{I2} & \rho_{I3} & \cdots & 1 \end{pmatrix}\right], \tag{4.2}$$

$$\boldsymbol{\varepsilon}_\mathbf{q} \sim N[\mathbf{0}, \boldsymbol{\Sigma}]$$

The off-diagonal terms of $\boldsymbol{\Sigma}$ capture the error covariance across the underlying latent continuous variables of the different episode categories; that is, they capture the effect of common unobserved factors influencing the propensity of choice of count level for each episode category. Thus, if ρ_{12} is positive, it implies that individuals with a higher than average propensity in their peer group to participate in the first episode category are also likely to have a higher than average propensity to participate in the second episode category. Of course, if all the correlation parameters (*i.e.*, off-diagonal elements of $\boldsymbol{\Sigma}$), which we will stack into a vertical vector $\boldsymbol{\Omega}$, are identically zero, the model system in Equation (4.1) collapses to independent ordered-response probit models for each episode category.

4.2.3 The Pairwise Marginal Likelihood Inference Approach

The parameter vector of the multivariate probit model is $\boldsymbol{\delta} = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2', ..., \boldsymbol{\beta}_I';\ \boldsymbol{\theta}_1', \boldsymbol{\theta}_2', ..., \boldsymbol{\theta}_I';\ \boldsymbol{\Omega}')'$, where $\boldsymbol{\theta}_\mathbf{i} = (\theta_i^1, \theta_i^2, ..., \theta_i^{K_i})'$ for $i = 1, 2, ..., I$. Let the actual observed count level for individual *q* and episode category *i* be m_{qi}. Then, the likelihood function for individual *q* may be written as follows:

$$L_q(\boldsymbol{\delta}) = \Pr(y_{q1} = m_{q1}, y_{q2} = m_{q2}, ..., y_{qI} = m_{qI})$$

$$L_q(\boldsymbol{\delta}) = \int_{v_1=\theta_1^{m_{q1}}-\beta_1' x_{q1}}^{\theta_1^{m_{q1}+1}-\beta_1' x_{q1}} \int_{v_2=\theta_2^{m_{q2}}-\beta_2' x_{q2}}^{\theta_2^{m_{q2}+1}-\beta_2' x_{q2}} \cdots \int_{v_I=\theta_I^{m_{qI}}-\beta_I' x_{qI}}^{\theta_I^{m_{qI}+1}-\beta_I' x_{qI}} \phi_I(v_1, v_2, ..., v_I \mid \Omega) dv_1 dv_2 ... dv_I \qquad (4.3)$$

The likelihood function above requires the computation of an *I*-dimensional rectangular integral. While there are maximum simulated likelihood (MSL) approaches that can evaluate such multidimensional normal integrals using the Geweke-Hajivassiliou-Keane simulator (Hajivassiliou *et al.*, 1996), as noted previously, they can become problematic even for moderate *I* in terms of computational effort. Thus, in this study, we employ a pairwise marginal likelihood estimation approach. The pairwise marginal likelihood function for individual *q* may be written as follows:

$$L_{CML,q}(\boldsymbol{\delta}) = \prod_{i=1}^{I-1} \prod_{g=i+1}^{I} \Pr(y_{qi} = m_{qi}, y_{qg} = m_{qg})$$

$$= \prod_{i=1}^{I-1} \prod_{g=i+1}^{I} \left[\begin{array}{l} \Phi_2\left(\theta_i^{m_{qi}+1} - \beta_i' x_{qi}, \theta_g^{m_{qg}+1} - \beta_g' x_{qg}, \rho_{ig}\right) - \Phi_2\left(\theta_i^{m_{qi}+1} - \beta_i' x_{qi}, \theta_g^{m_{qg}} - \beta_g' x_{qg}, \rho_{ig}\right) \\ -\Phi_2\left(\theta_i^{m_{qi}} - \beta_i' x_{qi}, \theta_g^{m_{qg}+1} - \beta_g' x_{qg}, \rho_{ig}\right) + \Phi_2\left(\theta_i^{m_{qi}} - \beta_i' x_{qi}, \theta_g^{m_{qg}} - \beta_g' x_{qg}, \rho_{ig}\right) \end{array} \right], (4.4)$$

and $L_{CML}(\boldsymbol{\delta}) = \prod_q L_{CML,q}(\boldsymbol{\delta})$

The pairwise likelihood function above is easily maximized, and the effort involved is no more difficult than in a usual bivariate ordered probit model. The pairwise estimator $\hat{\boldsymbol{\delta}}_{CML}$ is obtained by maximizing the logarithm of the function in Equation (4.4) with respect to the vector $\boldsymbol{\delta}$. The $\mathbf{H}(\boldsymbol{\delta})$ matrix and the $\mathbf{J}(\boldsymbol{\delta})$ matrix of the covariance, which is given by the inverse of Godambe's sandwich information matrix

$(\mathbf{V}_{CML}(\boldsymbol{\delta}) = [\mathbf{G}(\boldsymbol{\delta})]^{-1} = [\mathbf{H}(\boldsymbol{\delta})]^{-1}\mathbf{J}(\boldsymbol{\delta})[\mathbf{H}(\boldsymbol{\delta})]^{-1})$, can be estimated in a straightforward manner at the CML estimate $(\hat{\boldsymbol{\delta}}_{CML})$ as follows:

$$\hat{\mathbf{H}}(\hat{\boldsymbol{\delta}}) = -\left[\sum_{q=1}^{Q} \frac{\partial^2 \log L_{CML,q}(\boldsymbol{\delta})}{\partial\boldsymbol{\delta}\partial\boldsymbol{\delta}'}\right]_{\hat{\boldsymbol{\delta}}}$$

$$= -\left[\sum_{q=1}^{Q}\sum_{i=1}^{I-1}\sum_{g=i+1}^{I} \frac{\partial \log \Pr(y_{qi} = m_{qi}, y_{qg} = m_{qg})}{\partial\boldsymbol{\delta}} \frac{\partial \log \Pr(y_{qi} = m_{qi}, y_{qg} = m_{qg})}{\partial\boldsymbol{\delta}'}\right]_{\hat{\boldsymbol{\delta}}}, \text{and} \quad (4.5)$$

$$\hat{\mathbf{J}}(\hat{\boldsymbol{\delta}}) = \sum_{q=1}^{Q}\left[\left(\frac{\partial \log L_{CML,q}(\boldsymbol{\delta})}{\partial\boldsymbol{\delta}}\right)\left(\frac{\partial \log L_{CML,q}(\boldsymbol{\delta})}{\partial\boldsymbol{\delta}'}\right)\right]_{\hat{\boldsymbol{\delta}}}$$

4.3 Data

4.3.1 Data Source

The data used for the empirical analysis in the study is drawn from the 2007 American Time Use Survey (ATUS). The ATUS is a national level survey conducted and processed by the U.S. Census Bureau for the Bureau of Labor Statistics (ATUS, 2008). The household sample for the ATUS is drawn from the set of households that completed the Current Population Survey (CPS). Next, from each sampled CPS household, the ATUS randomly selects one individual of age 15 or over, and collects information on all episodes the individual participates in over the course of a single day. The episode-level information collected in the ATUS includes activity episode purpose, start and end time, location of participation (for example, grocery store, library, *etc.*), and 'with whom' participated in. In addition, data on individual and household socio-demographics, individual labor force participation and employment-related characteristics, and regional location and characteristics of the survey day are also collected.

4.3.2 Sample Formation and Description

The 2007 ATUS micro data were processed in several steps to obtain the sample for the current analysis. First, only individuals who were surveyed on a weekday that was not a holiday were selected, because the focus of the current study is to study individuals' activity participation patterns on a typical weekday. Second, all work, work-related, education, education-related, travel, sleep, and in-home activity episodes (such as phone call, grooming, *etc.*) were removed from the list of activity episodes undertaken by the respondents on the survey day. Third, all out-of-home activity episodes, originally documented in over four hundred fine activity purpose types, were aggregated into six broad activity purpose type categories: (1) personal/family care (including personal care, caring for children in the household, pick-up/drop-off of children/adults, and caring for extended family members; for the sake of brevity, we will refer to personal/family care activities simply as "family care" activities from hereon), (2) maintenance shopping (such as grocery shopping, purchasing gas/food, and banking), (3) non-maintenance shopping, (4) meals, (5) physically active recreation (including sports, exercise, recreational and volunteer activities), and (6) physically inactive recreation (including social, relaxing, movies, and attending religious/sporting/recreational events). Subsequently, the companion types for each episode were classified into five mutually exclusive and collectively exhaustive categories: (1) alone, (2) only family (includes children, spouse or unmarried partner), (3) only relatives (parents, sibling, grandchild, *etc.*), (4) only friends (friends, co-workers, neighbors, *etc.*), and (5) mixed company (a combination of family, relatives, and friends). The activity type and companion type classification resulted in thirty episode categories. Fourth, the number of episodes undertaken during the survey day by an individual in each of the episode categories is obtained by aggregating all episodes of that category for the person. Fifth, data on household and individual socio-demographics, residential location, and zonal characteristics were appended to the person-level file. Finally, several screening and consistency checks were performed and records with missing or inconsistent data were eliminated.

The final sample for analysis includes out-of-home non-mandatory episode participation information for 4143 individuals (workers and non-workers, aged 15 years or older) on a typical weekday. Table 4.1 presents the percentage distribution of individuals' participation in episodes by activity type and companionship type. For example, the first entry in Table 4.1 indicates that 91.3% of individuals do not undertake family care activities alone. Across all the categories, we find that meals with friends is the most frequently undertaken episode category on weekdays, with over 27% of individuals in the sample participating in one or more episodes of this category. Other categories with relatively frequent participation (across individuals) include maintenance shopping alone, family care with family, meals alone, and physically inactive recreation with friends. The last of these is also the activity purpose that individuals are most likely (relative to other activity purposes) to undertake with relatives (8.9%) or with mixed company (7.2%).

4.4 Empirical Analysis

4.4.1 Variable Specification

Several types of variables were considered in the model specification. These included (1) individual socio-demographics (gender, age, race, education level, employment status, student status, and indication of any disability), (2) household socio-demographics (household structure, presence of children, family income, and employment status of spouse/partner)[16], and (3) day of the week and seasonal effect variables.

In addition to the three groups of variable discussed above, we also considered several interaction effects among the variables. The final specification was based on a systematic process of removing statistically insignificant variables and combining variables when their effects were not significantly different. The specification process was also guided by prior research and intuitiveness/parsimony considerations. We should also note here that, for the continuous variables in the data (such as age and income limits), we tested alternative functional forms that included a linear form, a spline (or piece-wise linear) form, and dummy variables for different ranges.

[16] The ATUS survey does not collect information on household vehicle ownership. As a result, this variable is not available for use in the empirical analysis.

Table 4.1 Percentage of Individuals in Each Number of Episodes Category by 'With Whom' and Activity Types (Weekday)

'With Whom' Dimension	Number of Episodes	Activity Type Dimension					
		Family care	Maintenance shopping	Non-maintenance shopping	Meals	Physically active recreation	Physically inactive recreation
Alone	0	91.3	73.7	86.7	79.1	90.2	87.4
	1	7.3	18.6	10.9	18.6	7.9	9.7
	2	1.4	5.6	2.4	2.3	1.9	2.8
	≥ 3	-	2.2	-	-	-	-
Only family (children/spouse/partner)	0	78.8	90.5	92.5	91.6	96.4	95.7
	1	11.2	7.6	5.6	7.4	3.0	3.6
	2	6.5	1.9	1.9	1.0	0.6	0.7
	3	2.1	-	-	-	-	-
	≥ 4	1.4	-	-	-	-	-
Only relatives (includes parents, brother, sister, and other related persons)	0	91.5	95.8	96.5	93.5	97.5	91.1
	1	5.6	3.4	2.9	5.6	2.1	6.6
	≥ 2	3.0	0.8	0.6	0.9	0.4	2.3
Only friends (includes friends, co-workers, neighbors, *etc.*)	0	96.7	96.5	98.5	72.9	94.9	81.6
	1	2.4	2.9	1.5	22.3	4.2	12.9
	2	0.9	0.5	-	4.8	0.9	4.1
	≥ 3	-	-	-	-	-	1.4
Mixed company (*i.e.*, with family and/or relatives and/or friends)	0	94.2	98.3	99.3	95.6	97.9	92.8
	1	4.0	1.7	0.7	4.4	2.1	5.8
	≥ 2	1.8	-	-	-	-	1.4

4.4.2 Model Estimation Results

Table 4.2 presents the model estimation results. The columns in the table correspond to the explanatory variables, while the rows correspond to the episode categories. An empty cell indicates that the corresponding column exogenous variable does not have a statistically significant effect on the corresponding row episode category participation propensity. The t-statistic for each coefficient is provided beneath the coefficient in parentheses. The base category is listed in the heading of the column corresponding to that variable. The coefficients in the table indicate the effects of variables on the latent propensity of participation in each episode category (that is, they represent elements of the $\boldsymbol{\beta}_{\mathbf{i}}$ vector in Equation (4.1)). Since all the variables in the model are dummy variables, the relative magnitudes of the coefficients also provide an estimate of the importance of the variables in influencing participation propensities and participation probabilities. The marginal impact of variables on the participation probabilities for each combination of number of episodes for the different episode categories varies across individuals because of the non-linear structure of the ordered probit formulation. Aggregate level marginal effects may be computed for each dummy variable by changing the value of the variable to one for the subsample of observations for which the variable takes a value of zero and to zero for the subsample of observations for which the variable takes the value of one. We can then sum the shifts in expected aggregate shares in the two subsamples after reversing the sign of the shifts in the second subsample and compute an effective marginal change in expected aggregate shares in the entire sample due to a change in the dummy variable from 0 to 1. We are not showing these marginal effects here because there are as many as 80 trillion aggregate marginal effects (one for each combination of episode levels across all the 30 episode categories) for each variable. But in Chapter 7, we demonstrate the application of the model due to changes in two variables. In the following sections, we discuss the effect of variables on the latent participation propensities by variable category.

Table 4.2 Model Estimation Results (t-statistics in parentheses)

		Individual socio-demographics variables									
		Male (base: female)	Age (base: age ≥ 60)		Caucasian (base: non-Caucasian)	Education level (base: high school graduate)		Employment status (base: not employed)		Student (base: not student)	Have a disability (base: no disability)
			Age <40	40≤ Age <60		Education < bachelors	Education ≥ bachelors	Full time employed	Part time employed		
Family care	Alone	-0.179 (-3.05)	-0.122 (-2.12)					-0.318 (-5.05)	-0.174 (-2.03)		
	Only family	-0.509 (-10.31)	0.339 (3.76)	0.438 (4.76)							
	Only relatives	-0.195 (-3.24)	-0.347 (-4.68)	-0.220 (-2.88)				-0.327 (-5.00)			0.406 (3.02)
	Only friends										
	Mixed company	-0.398 (-5.81)	0.314 (5.00)					-0.113 (-1.62)	-0.182 (-1.81)		
Maintenance shopping	Alone		-0.363 (-8.55)			0.277 (5.47)	0.298 (6.23)				
	Only family	-0.218 (-3.82)	0.305 (3.62)	0.376 (4.37)				-0.278 (-4.69)			
	Only relatives	-0.226 (-3.00)						-0.323 (-4.34)			0.398 (2.67)
	Only friends							-0.126 (-1.72)			
	Mixed company	-0.210 (-2.07)	0.303 (3.13)					-0.190 (-1.95)			
Non-maintenance shopping	Alone	-0.101 (-1.97)	-0.430 (-6.45)	-0.188 (-2.76)		0.241 (3.92)	0.162 (2.65)	-0.198 (-3.19)	-0.134 (-1.62)		
	Only family	-0.233 (-3.84)			0.322 (3.45)						
	Only relatives	-0.310 (-3.89)									0.419 (2.33)
	Only friends										
	Mixed company	-0.289 (-2.01)	0.394 (2.91)								
Meals	Alone	0.290 (6.34)						1.075 (16.02)	0.600 (6.57)		
	Only family		0.161 (1.85)	0.131 (1.52)							
	Only relatives										
	Only friends	0.088 (2.10)			0.107 (1.89)			0.867 (16.53)	0.447 (6.10)	0.489 (7.36)	
	Mixed company		0.287 (4.19)		0.310 (2.95)						
Physically active recreation	Alone	0.190 (3.54)	-0.261 (-3.57)	-0.242 (-3.38)		0.136 (1.91)	0.451 (7.21)				
	Only family		0.168 (2.16)		0.254 (2.12)						0.589 (3.21)
	Only relatives						0.189 (2.20)				
	Only friends		-0.281 (-3.00)	-0.403 (-4.56)			0.156 (2.25)			0.461 (4.59)	
	Mixed company		0.205 (2.10)							0.331 (2.62)	
Physically inactive recreation	Alone	0.080 (1.58)			-0.188 (-2.98)			0.386 (7.18)			
	Only family		0.212 (2.87)					-0.465 (-6.34)			
	Only relatives							-0.250 (-4.54)			
	Only friends	0.123 (2.73)						0.225 (4.78)		0.377 (5.50)	
	Mixed company	-0.205 (-3.35)	0.284 (4.85)					-0.192 (-3.25)			

Table 4.2 (Continued) Model Estimation Results (t-statistics in parentheses)

		Household socio-demographics variables								Day-of-the-week and seasonal effect variables	
		Household (HH) structure (base: "other" HH)			Presence of children (base: age ≤ 4)		HH income (base: < 30k)		Spouse/partner employed (base: unemployed)	Friday (base: other days of the week)	Summer (base: fall, spring, and winter)
		Nuclear family HH	Couple HH	Single individual HH	4< Age ≤ 10	10< Age ≤ 15	30k ≤ Income < 75k	Income ≥ 75k			
Family care	Alone								0.161 (2.82)		
	Only family	0.381 (6.39)	-0.418 (-4.60)		0.579 (10.98)	0.308 (5.87)			0.537 (9.78)		
	Only relatives									0.183 (2.80)	
	Only friends								-0.477 (-5.20)		
	Mixed company										
Maintenance shopping	Alone									0.098 (1.92)	
	Only family	0.487 (6.81)	0.364 (4.92)		0.161 (2.27)					0.210 (3.26)	
	Only relatives	-0.423 (-4.59)								0.184 (2.23)	
	Only friends										
	Mixed company										
Non-maintenance shopping	Alone										
	Only family	0.614 (9.04)	0.393 (4.96)							0.202 (2.93)	
	Only relatives	-0.321 (-3.38)								0.237 (2.81)	
	Only friends						-0.216 (-2.02)				
	Mixed company										
Meals	Alone			0.278 (5.31)						-0.143 (-2.42)	
	Only family	0.576 (8.14)	0.532 (7.15)				0.345 (4.36)	0.294 (3.49)		0.124 (1.83)	
	Only relatives	-0.451 (-5.82)	-0.215 (-2.77)								
	Only friends	-0.166 (-3.37)		0.202 (3.78)		0.189 (3.60)				0.141 (2.80)	
	Mixed company									0.275 (3.47)	
Physically active recreation	Alone			0.169 (2.67)							
	Only family	0.357 (3.99)	0.187 (1.80)				0.206 (1.99)	0.348 (3.20)			0.222 (2.77)
	Only relatives										
	Only friends	-0.191 (-2.29)									0.157 (2.09)
	Mixed company						0.384 (2.70)	0.546 (3.85)			
Physically inactive recreation	Alone			0.146 (2.56)				-0.273 (-4.70)			
	Only family	0.405 (4.73)	0.210 (2.12)				0.327 (3.19)	0.411 (3.86)		0.161 (2.01)	
	Only relatives	-0.462 (-6.63)								0.150 (2.26)	
	Only friends	-0.225 (-4.09)		0.247 (4.37)		0.342 (6.16)				0.088 (1.63)	
	Mixed company									0.275 (4.05)	

4.4.2.1 Effect of Individual Socio-Demographic Variables

The results indicate the presence of distinct gender effects in activity type participation and accompaniment. Specifically, men are less likely than women to participate, across all companion types, in family care activities (except with “only friends”), maintenance activities (except “alone” and with “only friends”) and non-maintenance shopping activities (except with “only friends”). These results reinforce the gender stereotype of women being more responsible for, and/or more vested and interested in, family care and shopping activities, a recurring finding in the literature (for example, see Yamamoto and Kitamura, 1999, and Frusti *et al.*, 2003). However, men have a higher propensity than women to (a) participate alone in discretionary activities (*i.e*., meals out, physically active recreation, and physically inactive recreation), and (b) participate with “only friends” in meals out and physically inactive recreation. This is again consistent with the results found by Srinivasan and Bhat (2006) and Carrasco and Miller (2009), and suggests that men are more likely to undertake active and inactive leisure activities either alone or with friends on a weekday. Finally, men pursue physically inactive recreation with “mixed company” less than do women, potentially a reflection of the combination of family-centric responsibilities and social network level interactions of women relative to men (see Kapur and Bhat, 2007 for a similar result).

The effect of individual age on activity purpose and accompaniment type is accommodated in a non-linear fashion by introducing age in three categories: age less than 40 years, age 40 years or above but less than 60 years, and age 60 years or above (the base age category). The results suggest that, in general, individuals younger than 60 years are more disposed toward pursuing activities with “only family”, and are less likely to participate in physically active recreation with “only friends”. Further, individuals below the age of 40 years are the least likely (relative to other age groups) to participate in activity episodes alone and most likely to participate in episodes with mixed company. Overall, these patterns suggest a combination of the family orientation and larger social networks of younger individuals, perhaps due to household life cycle characteristics. For instance, compared to older individuals, younger individuals are likely to have more family responsibilities, have more social interactions with friends and co-workers, and also have a larger pool of individuals to interact as part of their extended family (parents, siblings, grandparents, *etc.*). Finally, individuals who are older than 60 years are most likely to participate in family care activities with “only relatives”, as evidenced by the negative coefficients corresponding to the age < 40 and $40 \leq age < 60$ columns for the “Family care-Only relatives” row of Table 4.2. This result may be attributable to such activities as care received by senior parents from their children, or child care provided by grandparents to grandchildren.

The race-related coefficients reveal that Caucasians are more likely than non-Caucasians to (1) participate in non-maintenance shopping and physically active recreation with “only family”, and (2) undertake meal episodes with only friends or with friends and family (we did not find statistically significant race differences in the group of non-Caucasians, and hence represent race differences by a simple binary representation between Caucasians and non-Caucasians). The above results are consistent with earlier studies that suggest that Caucasians have higher levels of participation in meals/recreational pursuits (see Bhat and Gossen, 2004 and Mallett and McGuckin, 2000), though our current study also introduces the “with whom” element that earlier studies do not. In this regard, our results also indicate that Caucasians tend to participate less than non-Caucasians in physically inactive recreation “alone”.

Education level also has an impact on the type of episodes pursued and accompaniment type. Specifically, individuals with an education level beyond high school have a higher propensity (than individuals with only a high school degree which is the base category) to participate alone in shopping activities (maintenance and non-maintenance) and physically active recreation. These results may be indicative of the tighter time constraints among individuals with high education, because of which it is easier to schedule shopping and physically active recreational activities (such as going to the gym) alone. Further, the results suggest that individuals with a bachelor’s degree or higher are more likely to pursue physically active recreation with relatives, and with friends. Overall, the results suggest an increased awareness among highly educated adults of the benefits of investing in health and fitness-enhancing pursuits, highlighting the importance of a good education for a healthy society.

Employment status, in the current study, is characterized as employed full-time, employed part-time and unemployed. The several negative coefficients in the "family care" and "maintenance shopping" panels of the table corresponding to the full-time employed variable reflect the lower propensity of full-time employees to pursue these activities (relative to other individuals). The same is true for non-maintenance shopping, though this is confined to the "alone" accompaniment type. Overall, full-time employed individuals have tight time constraints, which may explain their reduced participation in family care and shopping pursuits (see Goulias and Kim, 2001, for a similar result). However, full-time employed individuals have a high propensity to have meals out and physically inactive recreation episodes alone or with friends. The result regarding meals out alone or with friends is perhaps a manifestation of lunch activity participation alone or with co-workers. Finally, full-time employees are less likely to participate in physically inactive recreation with "only family", "only relatives", and "mixed company", potentially another reflection of tight time constraints (see also Yamamoto *et al.*, 2004). The results for part-time workers provide similar results as for full time workers, except for participation in maintenance shopping and physically inactive recreation.

The next variable in the table corresponds to student status. In this analysis, we defined an individual who is enrolled in high school, college, or university as a student. As expected, students have a high propensity to participate in discretionary activities (meals, physically active recreation, and physically inactive recreation) with friends, potentially a reflection of the combination of social opportunities to interact with friends as well as the social pressures to "fit in" within their peer group.

As one would expect, physical disability significantly affects activity episode participation. Individuals with a physical disability are likely to need assistance from their relatives or immediate family for activity participation, as indicated by the positive coefficients in the "only relatives" or "only family" rows of Table 4.2.

4.4.2.2 Effects of Household Socio-Demographic Variables

Household structure effects were considered by including several types of households, including nuclear family households (two adults of opposite/same sex with one or more children), couple families (two adults of opposite/same sex), single individual households, and "other" households (roommate households, returning young adult households, other related individual households, and all other types of households). The results show that adults in nuclear and couple family households are much more likely than adults in other households to pursue non-family care activities with their immediate family (as reflected in the positive coefficients for nuclear and couple families in the "only family" row for all non-family care activity purposes). Further, nuclear households are less likely than other households to participate in non-family care activities with friends or relatives. These results indicate the high levels of intra-household interactions within nuclear family households and, to a somewhat lesser degree, in couple family households. On the other hand, the results for "single individual" households shows that there is a relatively higher propensity of inter-household interactions with friends in the meal and physically inactive recreation activities of individuals who live alone (these individuals also participate more in meals and recreation alone). Overall, the results reinforce the need to explicitly consider intra-household and inter-household interactions in activity-travel pattern modeling, as discussed in the first section of this Chapter. Clearly, the nature of the interactions varies by household structure, which also needs to be considered in the modeling. Besides, earlier studies, such as Bhat and Srinivasan (2005), indicate that nuclear and couple family households have a higher participation propensity in shopping and physically active and inactive recreation activities as a whole, but our current study reveals that this is the case only for episodes with the immediate family. In fact, as just indicated above, nuclear family households have a lower propensity than other households to participate in shopping and discretionary (meals/recreation) activities with friends and relatives. This underscores the need to consider accompaniment type at the level of generation of episodes (as done in this study), and not further downstream in the modeling process where episodes are first generated purely by activity purpose and then assigned to one of many accompaniment types.

The effect of age of children is introduced in the model in three categories: presence of children 4 years old or younger (the base category), presence of children aged between 5 to 10 years, and presence of children aged between 11 to 15 years. As expected, adults in households with older children (aged 5 years or more) are more likely than adults in households with young children (less than 5 years of age) to have family care episodes with "only family", a clear reflection of the chauffeuring of children to/from school and other non-school activities as children grow older (sometimes labeled in the popular press as the "soccer mom" and "tennis dad" responsibilities). Adults in households with children in the 5-10 age group partake more in maintenance shopping episodes with "only family", which may be attributed to one or both parents pursuing maintenance shopping with the child "in tow". This effect is not statistically significant for the oldest child group since these children have acquired a certain level of independence and do not need child care at all times. Besides, there is evidence from the social psychology literature that pre-teenagers and teenagers would rather not be seen with parents, since this is considered "uncool" (Thornton *et al.*, 1995, Williams, 2003).[17] Of course, the independence levels of children in the pre-teens and teens also enables the participation of parents in meals and physically inactive recreation activities with friends, as reflected by the positive coefficients in the "meals-only friends" and "physically inactive recreation-only friends" rows of Table 4.2.

The effect of income is captured using dummy variables for different income categories, which enables the accommodation of nonlinear impacts on the propensity to participate in episodes (the dummy variable representation was found to be superior to a continuous linear income effect in our specifications). The results in Table 4.2 show that household income influences participation in meals, physically active recreation, and

[17] This finding is also supported by message boards and parent blogs posted on a number of websites such as life.familyeducation.com, www.ParentsConnect.com, www.theparentreport.com, family.go.com, all dedicated to address and deal with pre-teen and teenage issues.

physically inactive recreation. As expected, individuals in high-income households have a higher propensity to participate in these activity episodes because of their higher expenditure potential for discretionary pursuits. However, this is only true for episodes participated with "only family". In fact, individuals in highest income group are less likely than individuals in other income groups to pursue physically inactive recreation alone (perhaps attributable to time constraints due to the level and intensity of work activity). Also, middle income individuals have a lower propensity to participate in non-maintenance shopping with "only friends", a result that is not immediately intuitive and needs exploration in future studies. But, overall, such differential episode generation rates by accompaniment type can only be accommodated if accompaniment type is considered at the generation level, rather than later on in the modeling hierarchy.

Finally, in the category of household demographics, individuals in a household with a working spouse contribute more (less) than individuals without a working spouse to family care episodes alone or with immediate family (with friends).

4.4.2.3 Day of Week and Season Variables

The variables considered in this category include day of week variables and season variables (categorized as summer, fall, spring and winter). Clearly, there is a higher propensity of participation on Fridays in almost all non-physically active combinations of activity purpose and accompaniment. Further, it is unlikely that individuals pursue meals out activities alone on Fridays. For other activity purposes except maintenance shopping, there is no difference between Fridays and other days for solo-participation in episodes. Overall, individuals pursue more non-physical activity episodes on Fridays relative to other days of the week, and generally participate with family and friends.

The seasonal effects reflect a higher propensity to participate in physically active recreation with family and friends over the summer compared to other seasons. This may be attributable to better weather conditions for outdoor activities, more daylight time, and more schedule opportunities to pursue activities with family and friends.

4.4.2.4 Threshold Parameters

The threshold parameters are not shown in the table, but are available on request from the authors. These parameters represent the cut-off points that map the latent propensity of individuals to participate in each activity purpose-accompaniment type category to the reported number of episodes for each category. As such, they do not have any substantive behavioral interpretations.

4.4.2.5 Correlation Estimates

As indicated earlier in Chapter 3, it is not practical to estimate the parameters of the full correlation matrix (in the current case, the number of parameters in the full correlation matrix is 435). In our analysis, we specified several initial exclusion restrictions based on (1) intuitive considerations (for example, there is no reason why unobserved factors influencing participation in maintenance shopping with family should be correlated with unobserved factors influencing participation in physically active recreation with friends), and (2) the estimation of bivariate models for pairs of episode categories to determine if the corresponding correlations were statistically significant. These initial exclusion restrictions were used to estimate several alternative model specifications using the pairwise procedure proposed, and the final correlation matrix specification was obtained based on statistical fit and parsimony considerations.

The estimated covariances and their t-statistics (in parentheses) are shown in Table 4.3. Only the upper diagonal terms in the variance-covariance matrix are shown since the matrix is symmetric. As mentioned before, the variance of the error terms are set to one to normalize the scale (see Section 4.2.2). The covariance (correlation) matrix indicates several statistically significant correlations among the stop-making propensities of different activity type-accompaniment categories, highlighting the importance of accounting for common unobserved factors in modeling episode participation frequency. For the sake of conciseness, we focus only on the salient aspects of the covariance matrix structure in the discussion here. Specifically, the following observations may be made

Table 4.3 Correlation in Unobserved Propensities Across the Choice Dimension (t-statistics in parentheses)

		Family care					Maintenance shopping					Non-maintenance shopping					Meals					Physically active recreation					Physically inactive recreation				
		Alone	Only family	Only relatives	Only friends	Mixed company	Alone	Only family	Only relatives	Only friends	Mixed company	Alone	Only family	Only relatives	Only friends	Mixed company	Alone	Only family	Only relatives	Only friends	Mixed company	Alone	Only family	Only relatives	Only friends	Mixed company	Alone	Only family	Only relatives	Only friends	Mixed company
Family care	Alone	1	0.087 (2.88)	0.087 (2.88)	0.275 (5.67)																										
	Only family		1																												
	Only relatives			1					0.363 (1.03)					0.460 (15.95)					0.460 (15.95)					0.228 (5.04)					0.363 (1.03)		
	Only friends				1					0.476 (7.49)					0.476 (7.49)					0.290 (5.58)					0.248 (7.53)					0.377 (6.79)	
	Mixed company					1					0.377 (9.91)										0.377 (9.91)					0.352 (3.22)					0.395 (16.28)
Maintenance shopping	Alone						1	-0.042 (-1.33)				0.278 (11.99)																			
	Only family							1					0.473 (22.84)					0.473 (22.84)					0.282 (5.26)					0.282 (5.26)			0.257 (4.60)
	Only relatives								1					0.492 (8.58)					0.442 (9.46)					0.228 (5.04)				0.286 (3.10)	0.357 (27.57)		
	Only friends									1					0.429 (2.44)					0.347 (23.95)					0.248 (7.53)					0.347 (23.95)	
	Mixed company										1										0.437 (2.62)			0.337 (2.75)	0.313 (7.83)	0.341 (14.28)		0.331 (3.84)	0.305 (6.29)		0.394 (9.61)
Non-maintenance shopping	Alone											1																			
	Only family												1					0.430 (17.31)					0.303 (4.11)					0.285 (6.56)			0.232 (6.16)
	Only relatives													1					0.509 (11.47)					0.228 (5.04)				0.298 (3.00)	0.357 (27.57)		0.265 (7.41)
	Only friends														1					0.320 (13.32)					0.248 (7.53)					0.320 (13.32)	0.289 (2.34)
	Mixed company															1	0.308 (4.94)	0.339 (4.89)	0.340 (5.25)		0.426 (7.35)		0.365 (2.80)	0.364 (4.21)	0.344 (10.48)	0.366 (16.50)		0.363 (2.91)	0.326 (5.60)	0.319 (6.59)	0.366 (16.50)
Meals	Alone																1	-0.192 (-13.92)	-0.032 (-0.47)	-0.526 (-57.74)	-0.192 (-13.92)						0.191 (2.66)				
	Only family																	1		-0.131 (-2.27)			0.389 (8.48)			0.302 (3.74)		0.531 (22.70)			0.238 (4.45)
	Only relatives																		1	-0.131 (-3.14)				0.228 (5.04)					0.605 (31.21)		
	Only friends																			1					0.300 (5.89)		-0.179 (-6.82)		-0.155 (-3.07)	0.407 (29.45)	
	Mixed company																				1					0.341 (14.28)					0.567 (8.88)
Physically active recreation	Alone																					1									
	Only family																						1			0.354 (2.58)		0.341 (9.11)			0.275 (5.32)
	Only relatives																							1				0.302 (3.44)	0.322 (13.77)		
	Only friends																								1					0.255 (8.12)	
	Mixed company																									1		0.331 (2.96)			0.340 (8.04)
Physically inactive recreation	Alone																										1				
	Only family																											1			0.308 (8.77)
	Only relatives																												1	-0.076 (-2.18)	
	Only friends																													1	
	Mixed company																														1

from Table 4.3. First, the shaded matrices along the diagonal of the correlation matrix do not have many off-diagonal elements. This suggests the absence of common unobserved factors that affect participation across accompaniment types for any given activity purpose category. Thus, for example, a higher than average propensity to participate alone in non-maintenance activity (due to unobserved factors) does not increase or decrease the propensity to participate in non-maintenance activity with others. The main exception to this general observation is for meal activities, where there are significant substitution effects across accompaniment types. That is, an individual's propensity to pursue dining out with a particular companion type is negatively correlated with the individual's propensity to pursue dining out with other companion types. Second, the large number of parameters significant and consistently positive along the diagonals in each off-diagonal matrix of Table 4.3 highlights the preference for sticking to the same accompaniment (social) group for undertaking different types of activities. For example, individuals predisposed to participating in maintenance shopping activity with "only family" tend to participate in other activity purposes too with "only family". This preference (or stickiness) to pursue all types of activities with the same accompaniment group is particularly strong for the non-alone accompaniment categories. Third, some of the highest correlation values may be observed along the diagonal of the matrix corresponding to meals (row entry) and physically inactive recreation (column entry), suggesting that meals out and physically inactive recreation episodes are frequently combined (for instance, dinner out and a movie, or dinner out and a cultural event). This is reinforced by the fact that individuals who tend to have meals with "only friends" are not very likely to pursue physically inactive recreation alone or with "only relatives". In any case, there is a general complementary relationship between the propensities to participate in meals out and physically inactive recreation. Fourth, there are also quite high correlation values along the diagonals of the matrices corresponding to maintenance shopping and non-maintenance shopping, maintenance shopping and meals, and non-maintenance shopping and meals, highlighting the strong complementary tendencies among shopping/meal activities with the same accompaniment type. Fifth, the most number of off-diagonal correlation elements may be found in the matrix for non-maintenance shopping and physically inactive recreation, indicating substantial complementary effects in participation propensities for these two activity purpose categories across all types of accompaniment arrangements. Sixth, rather than the common perception that there is a substitution effect between physically active and physically inactive recreation propensities, there is in fact a complementary effect. That is, individuals who participate more in physically inactive recreation are also more likely (after controlling for observed factors) to participate in physically active recreation. Finally, there is a general complementary relationship between participation with "mixed company" and participation with other company types for the non-maintenance shopping activity and other discretionary activity purposes (meals, physically inactive recreation, and physically active recreation).

4.4.2.6 Overall Measures of Fit

The log-composite likelihood value for the independent ordered-response probit model (that is, independent ordered-response probit models for each episode category) with only the threshold parameters is –1,136,772.91. The corresponding value at convergence for the fully specified independent ordered-response probit model (IORP) is –1,083,191.5 and that for the fully specified multivariate ordered-response probit model (MORP) is –1,081,484.6. The composite likelihood ratio test (*CLRT*) statistic for comparing the MORP model with the IORP model is 3413.83. However, the *CLRT* statistic does not have the standard chi-squared asymptotic distribution under the null hypothesis as in the case of the maximum likelihood inference procedure. In the current study, we use bootstrapping to obtain the precise distribution of the *CLRT* statistic (see Section 2.5 for details on the procedure for bootstrapping).

The estimated p-value based on 50 bootstrap samples is 0.0196 for the test between the MORP and IORP models. This low p-value rejects the null hypotheses of absence of correlations across the propensities of participation for the different episode categories, and highlights the value of the MORP model estimated in the current study.

Of course, this should also be obvious from the many statistically significant parameters in the correlation matrix in Table 4.3.

Another more intuitive, but aggregate, approach to obtain a sense of measure of fit would be to compare the predicted versus the actual number of out-of-home episodes for each activity purpose-accompaniment combination level. In this study, and to illustrate the data fit of the models while also conserving on space, we present the results only for the episode level combinations of two categories: meals with friends and physically inactive recreation with friends. These are two of the most common episode categories participated in during weekdays, as observed earlier in Section 4.3.2. Also, we select these two episode categories because they are helpful in demonstrating the application of the model in response to changes in socio-demographic variables (see next section). Table 4.4 presents the results, where the numbers in underlined font correspond to the actual number of individuals participating in each level of the two episode categories. The numbers in plain font are the predicted values from the MORP model, while the italicized numbers are the predicted values from the IORP model. A visual comparison of these numbers indicates the superiority in data fit of the MORP model. To quantify this, we develop a weighted mean absolute percentage error statistic that is computed as the absolute percentage error for each cell weighted by the fraction of individuals in each cell (based on the actual numbers in each cell). This statistic is 4.5% for the MORP model and 17.8% for the IORP model. One can also compute a more traditional root mean-squared error (RMSE) statistic between the predicted and actual values across all the cells for each of the MORP and IORP models. This statistic is 17.8 for the MORP and 76.4 for the IORP.

Overall, from the perspectives of both disaggregate and aggregate measures of fit, the MORP model clearly outperforms the IORP model.

Table 4.4 Number of Individuals Choosing “Meals with Friends” and “Physically Inactive Recreation with Friends” Episodes

Number of “meals with friends” episodes	Number of “physically inactive recreation with friends” episodes			
	0	1	2	3
0	2667.00[a] 2650.14[b] *2501.61*[c]	269.00 285.22 *375.46*	67.00 70.17 *117.54*	19.00 16.87 *38.47*
1	597.00 638.12 *729.13*	207.00 188.62 *127.76*	92.00 69.20 *42.60*	28.00 24.84 *14.92*
2	117.00 100.39 *152.35*	58.00 55.14 *29.27*	12.00 28.72 *10.16*	10.00 15.57 *3.72*

[a] The actual number of individuals participating in each combination level of episode category.

[b] The predicted number of individuals from the MORP model participating in each combination level of episode category.

[c] The predicted number of individuals from the IORP model participating in each combination level of episode category.

4.5 Summary and Conclusions

This chapter proposes a multivariate ordered-response system framework to model the interactions in activity episode decisions across household and non-household members at the fundamental level of activity generation. Such a system recognizes the dependence in the number of episodes generated for different purposes as well as with different accompaniment types, and explicitly allows *complementary* and *substitution* effects in activity episode participation decisions. The econometric challenge in estimating such a joint multivariate ordered-response system with a large number of episode categories is addressed by resorting to the technique of composite marginal likelihood.

The empirical analysis in the study uses data drawn from the 2007 American Time Use Survey (ATUS). Unlike conventional activity-travel surveys, the ATUS survey explicitly collects information on all accompanying family and non-family members for all activity episode participations. Thus, it is an ideal dataset for exploring the social context of adults' activity episode participations.[21] The empirical results provide important insights into the determinants of adults' weekday activity episode generation behavior. For instance, the results indicate the presence of distinct gender effects in activity type participation and accompaniment, with women being more responsible for, and/or more vested and interested in, family care and shopping activities, and men being more likely to undertake active and inactive leisure activities either alone or with friends. Further, there are also clear age-related effects. Individuals below the age of 40 years are the least likely (relative to other age groups) to participate in activity episodes alone and most likely to participate in episodes with mixed company, suggesting a combination of the family orientation and larger social networks of younger individuals. Race, education level, employment and student status, household structure and presence of children, household income, the day of week, and season of the year also have important effects on adults' weekday activity episodes by purpose and the social context of participation. In

[21] A limitation of ATUS is that it does not collect locational information on household residences or activity episode participation locations. Hence, our analysis is unable to include built environment and locational effects on episode generation behavior. If available, this information can be incorporated as additional attributes in our multivariate ordered-response system.

addition to estimating the coefficients of explanatory variables, the CML approach allows us to estimate the parameters underlying the correlation due to unobserved factors in the propensity to participate in the 30 different purpose-accompaniment episode categories. Accommodating these unobserved correlation effects leads to a statistically superior data fit in the empirical context of this study and also provides useful insights into complementary and substitution effects among activity type and companionship type dimensions. Overall, the empirical estimation results underscore the ability of the CML approach to specify and estimate behaviorally rich structures to analyze inter-individual interactions in activity episode generation.

In summary, the results underscore the substantial linkages in the activity episode generation of adults based on activity purpose and accompaniment type. The extent of this linkage varies by individual demographics, household demographics, day of the week, and season of the year. These inter- and intra-family linkages, and their variations across individuals, need to be accommodated within the framework of activity-based travel modeling for accurate travel forecasting and reliable transportation policy analysis.

Chapter 5

Modeling the Influence of Family, Social Context, and Spatial Proximity on Non-Motorized Transport Mode Use

5.1 Introduction

5.1.1 Motivation

In recent years, the study of individual and household choices of non-motorized travel modes for activity participation has received increasing attention at the interface of transportation and related fields, such as environmental sustainability, accident analysis and prevention, urban design and planning, sociology, child and adolescence development, and public health. In the context of transportation, an analysis of the 2009 National Household Travel Survey (NHTS) data indicates that nearly 20 percent of all trips undertaken in the USA are one mile or shorter, and just over 40 percent of all trips are three miles or less. These statistics suggests that walking and bicycling could conceivably be viable modes for a larger extent of trip making than is currently the case.[22] There is also evidence that projects such as "Walking School Bus" and "KidsWalk-to-School" help children develop social skills and promote social vibrancy within communities (Kingham and Ussher, 2007). From an environmental perspective, an increase in the use of non-motorized transportation will lead to an overall reduction in mobile source emissions, pollution exposure, and potential health risk such as respiratory dysfunction and cardiopulmonary disease (Tonne *et al*., 2007, de Nazelle and Rodríguez, 2009). From a safety standpoint, studies have shown that a positive shift in walking and cycling has a negative influence on fatality and injury rates for pedestrians and cyclists (Jacobsen, 2003, Robinson, 2005).

Another area where participation in walking and bicycling activities has received significant attention is in the context of public health concern. Data from the U.S.

[22] In the rest of this chapter, we shall use the terms walking and bicycling and non-motorized (transport) modes interchangeably.

National Health Interview Survey (NHIS) suggests that in 2009, only 34.7% adults (aged 18 or over) participated in regular leisure-time physical activity. Low participation in regular physical activity has a number of implications on individual's health and well being because of the strong relationship between lack of physical activity and obesity.[23] For instance, obesity has been linked as a serious risk factor for health problems such as coronary heart diseases, type 2 diabetes, liver and gallbladder disease, osteoarthritis, and depression (Swallen *et al.*, 2005, WHO, 2006). In addition, it is now well documented that overweight/obese children are more likely to suffer from low self-esteem and/or be victim of bullying (Lumeng *et al.*, 2010). The problem of overweight and obesity also has significant economic consequences on the US health care system. In 2001, the average health-care cost of an obese individual was estimated to be $1,069 more than a normal-weight individual (Thorpe *et al.*, 2004). Wang *et al.* (2008) predict that, if the current trend continues, overweight/obesity related health care cost is likely to double every decade and by 2030 this cost may be as much as $956.9 billions. In addition to the healthcare related cost, obesity has been associated with other socioeconomic costs. For example, Jacobson and King (2009) found that, in the USA, overweight non-commercial vehicle passengers contribute to an additional billion gallons of gasoline consumption every year. While there are several factors that affect obesity in children and adults, it is now well established that lack of/a low level of physical activity is a common contributing factor (Haskell *et al.*, 2007, Bassett *et al.* 2008). In fact, studies have shown that regular participation in physical activities (such as walking and bicycling) has beneficial effects on all-cause mortality for individuals of all age groups (Andersen *et al.*, 2000). In children and adolescents, additional benefits of regular participation in physical

[23] Data from 2007-2008 National Health and Nutrition Examination Survey (NHANES) indicates that 18.7% children (age 6 to 19 years) and 33.8% adults (age 20 years or above) in the USA are considered obese. Among children, the obesity rates are 19.6% and 18.1% for age groups 6-11 years and 12-19 years respectively. Among adults, the obesity rates are 32.2% and 35.5% for men and women respectively (see Ogden *et al.*, 2010 and Flegal *et al.*, 2010 for more detailed breakdown on overweight and obesity rates among children and adults respectively).

activities include healthy musculoskeletal development, maintaining blood pressure, bone strength, and improvement in academic performance (Strong *et al.* 2005).

The above discussion clearly indicates that many studies in a number of disciplines have tried to quantify and understand the factors that influence walking, bicycling, and physical activity participation among children and adults. For instance, several studies have examined overall physical activity participation among adults and children in the context of the built environment, but these studies do not explicitly separate walking and bicycling from other physically active episodes of participation (*e.g.*, Badland and Schofield, 2005, Frank *et al.*, 2005). Several other studies have lumped walking and bicycling together into a single category of non-motorized mode use, without sufficiently recognizing that there may be trade-offs across the use of these two modes of transport and important differences in the factors that influence their use (*e.g.*, Cao *et al.*, 2009). Yet other studies have examined walking or bicycling in isolation of the other, thus preventing the ability to model or understand the use of these non-motorized physically active modes in a holistic perspective. There are numerous studies exclusively dedicated to the study of the choice of walking (*e.g.*, McGinn *et al.*, 2007, Forsyth *et al.*, 2009, and Agrawal and Schimek, 2007), and others that exclusively focus on bicycling (*e.g.*, Rietveld and Daniel, 2004, Hunt and Abraham, 2007, and Xing *et al.*, 2010).

The importance of considering walking and bicycling mode use in a unified framework has not gone unrecognized in the literature. However, many of these studies have restricted their focus to examining walking and bicycling habits of either children/adolescents, particularly in the context of their travel to and from school (*e.g.*, Cooper *et al.*, 2006), or adults in the context of their commute or short-distance trip making (*e.g.*, Plaut, 2005, Kim and Ulfarsson, 2008). Ogilvie *et al.* (2004), Pikora *et al.* (2003), and Saelens *et al.* (2003) provide more extensive reviews of studies in this topic area. In general, past research considers specific demographic segments, and describes or models non-motorized mode use of individuals in isolation of their social, familial, and spatial context. Sener *et al.* (2009) jointly considered physical activity participation of all members in a family, but their analysis was limited by the consideration of all physical activities together as a single choice.

5.1.2 The Current Study

The objective of this study is to propose and estimate a joint model system of walking and bicycling activity duration that recognizes the presence of both observed and unobserved variables, and explicitly incorporates dependence between walking and cycling activity durations due to: (a) individual-specific factors, (b) family-level influence, (c) social group to which the individual belongs, and (d) spatial effects of residential neighborhood. The total time spent walking and bicycling by individuals aged 5 years or above over a period of one week is considered here as a measure of the amount of non-motorized mode use.[24] The data used in this study is drawn from the San Francisco Bay Area subsample of the 2009 National Household Travel Survey (NHTS 2009). In addition to individual- and household-level socio-demographic information, the NHTS 2009 California add-on data set includes detailed attitudinal information on walking and bicycling. This makes the California-specific NHTS 2009 data set particularly appealing for this study.

The current study uses a hazard-based duration model structure. Specifically, a proportional hazard formulation is employed to capture walking and bicycling activity participation behavior of individuals.[25] The model system specified here recognizes the presence of individual-specific unobserved factors that can affect the amount of non-motorized mode use as a whole, as well as the amount of time specifically allocated to bicycling vis-à-vis walking. The model incorporates the effects of unobserved common household-specific attributes that can influence walking and cycling activity durations of

[24] Participation in walking is defined as an activity undertaken for a specific purpose such as walking to/from public transportation stop, for exercise, walking the dog, *etc*. (*i.e.*, walking as part of daily household chores is not included in the activity duration).

[25] Duration models are being increasingly used in transportation field in recent years. The reader is referred to Hensher and Mannering (1994) for a review of applications of duration models in transportation research in the past. Also, see Bhat and Pinjari (2008) for a list of recent applications of duration models in this area.

all individuals in a household. Similarly, social group-specific and spatial-cluster specific unobserved factors that impact walking and cycling activity durations are also included in the model system. A specification that captures these multiple effects and interactions leads to a multi-level cross-cluster structure that recognizes and preserves between-cluster heterogeneity. That is, the proposed model explicitly recognizes that, in addition to observed exogenous variables, walking and bicycling activity durations of an individual depend on common unobserved individual-, household-, social-, and spatial-specific factors. Further, the dependences between walking and cycling activity durations within and across individuals are correlated through these unobserved factors. For example, consider individuals from a "health conscious" household. Individuals from this household are likely to have a higher propensity to engage in walking and bicycling activities for longer time periods. Also, between these two activities, if the household has an intrinsic preference for walking over bicycling, then the participation durations of walking activity of all individuals in the household are likely to be affected. Thus, ignoring unobserved common household-specific factors and considering only observed exogenous variables is likely to result in inconsistent parameter estimates. This, in turn, can lead to less accurate assessment of the responsiveness of policy measures designed to promote walking and/or bicycling at individual as well as household levels. In general, ignoring heterogeneity due to multi-level clustering effects will result in inconsistent parameter estimates (Jones and Duncan, 1996, Bhat, 2000).

The multivariate cross-cluster model system proposed in the current study requires the evaluation of a more than thousand-dimensional integral (the number of individuals in the data set multiplied by the number of activity types). As using the usual estimation techniques could become computationally prohibitive, a composite marginal likelihood (CML) approach is employed.

The rest of the chapter is organized as follows. The detailed modeling methodology is presented in Section 5.2. Section 5.3 provides an overview of the data used in the study. Section 5.4 presents model estimation results. The salient features and findings of the study are summarized and concluding thoughts are offered in the final section of the chapter.

5.2 The Model Structure

5.2.1 Mathematical Formulation

Let $\lambda_{qijlm}(\tau)$ represent the hazard at continuous time τ of ending time investment in activity type *m* (*m* = 1, …, *M*) for the q^{th} (*q* = 1, 2, …, *Q*) individual belonging to household *i* (*i* = 1, 2, …, *I*), social cluster *j* (*j* = 1, 2, …, *J*), and spatial cluster *l* (*l* = 1, 2, …, *L*). That is, $\lambda_{qijlm}(\tau)$ represents the conditional probability that individual *q* will stop investing additional time in activity type *m* during an infinitesimally small time period after time τ, given that the individual has not yet stopped investing time in activity type *m* until time τ:

$$\lambda_{qijlm}(\tau) = \lim_{\Delta \to 0^+} \frac{\Pr(\tau < T_{qijlm} < \tau + \Delta \mid T_{qijlm} > \tau)}{\Delta}, \qquad (5.1)$$

where T_{qijlm} is the index representing the continuous time of participation in activity *m* for individual *q* belonging to household *i*, social cluster *j*, and spatial cluster *l*. Next, the hazard rate $\lambda_{qijlm}(\tau)$ may be written using a proportional hazard formulation as a function of a vector of covariates $\mathbf{x_{qm}}$ specific to individual *q* and activity type *m*:

$$\lambda_{qijlm}(\tau) = \lambda_{m0}(\tau)\exp(\boldsymbol{\beta}'_{\mathbf{m}}\mathbf{x_{qm}} + \alpha_{qijlm} + \omega_{qm}), \qquad (5.2)$$

where $\boldsymbol{\beta}_{\mathbf{m}}$ is a vector of covariate coefficients specific to activity *m*, α_{qijlm} is a scalar term associated with individual *q*, household *i*, social cluster *j*, spatial cluster *l*, and activity type *m*, and ω_{qm} is an unobserved idiosyncratic factor affecting the hazard for individual *q* and activity *m* (ω_{qm} may represent factors such as the q^{th} individual's intrinsic liking or aversion for activity type *m*). ω_{qm} is assumed to be independent of x_{qm} and α_{qijlm}, and normally distributed with a mean of zero (an innocuous normalization for identification purposes) and variance σ_m^2.[26]

Equation (5.2) represents the micro-level model for individual *q* in household *i*, belonging to social cluster *j* and spatial cluster *l*, participating in activity *m*. We next allow the scalar term α_{qijlm} to vary across individuals, households, social clusters, and spatial clusters in a higher-level macro-model:

$$\alpha_{qijlm} = \boldsymbol{\varsigma}'\mathbf{h_{qijl}} + v_q + u_i + u_{im} + w_j + w_{jm} + z_l + z_{lm}, \qquad (5.3)$$

where $\mathbf{h_{qijl}}$ is a vector of observed variables specific to individual *q* or household *i* or social cluster *j* or spatial cluster *l* or to the combination of these higher level macro-units, $\boldsymbol{\varsigma}$ is a corresponding parameter vector to be estimated, v_q is an individual-specific random term that captures unobserved variation across individuals in the hazard function for all activity types (v_q may include intrinsic individual-specific factors such as motivation for physical activity that affects the duration of participation of the individual in all types of walking and bicycling activities), u_i is a household-specific random term that captures unobserved variation across households in the hazard function for all activity types (u_i may include intrinsic household-specific lifestyle factors impacting all individuals in the household in their attitudes and perspectives toward all types of walking and bicycling activities), u_{im} is another household-specific random term that captures unobserved variation across households in the hazard function specific to activity type *m* (u_{im} includes intrinsic household-specific factors that makes individuals in a household more inclined to participate in specific types of physical activity such as bicycling), w_j and w_{jm} are similar social-cluster specific random terms, and z_l and z_{lm} are similar spatial-cluster specific random terms. Consider that the above random terms are

[26] It is quite typical to assume that $c_{qm} = \exp(\omega_{qm})$ is gamma distributed rather than assuming ω_{qm} to be normally distributed. This is because when there are no cluster effects, the gamma distribution assumption leads to a mixing structure that results in a closed form likelihood expression. However, in the current study where there are cross-cluster effects, it is more convenient to adopt a normal distribution in the estimation, as we indicate later.

realizations from independent and identically normally distributed terms across individuals (for v_q), across households (for u_i and u_{im}), across social clusters (for w_j and w_{jm}), and across spatial clusters (for z_l and z_{lm}). Thus, the distributions of the error terms are:

$v_q \sim N\left[0, \theta^2\right]$, $u_i \sim N\left[0, \mu^2\right]$, $u_{im} \sim N\left[0, \mu_m{}^2\right]$, $w_j \sim N\left[0, \eta^2\right]$, $w_{jm} \sim N\left[0, \eta_m{}^2\right]$, $z_l \sim N\left[0, \delta^2\right]$, and $z_{lm} \sim N\left[0, \delta_m{}^2\right]$

Next, define $\boldsymbol{\gamma}_{\mathbf{m}} = (\boldsymbol{\beta}'_{\mathbf{m}}, \varsigma')'$ and $\mathbf{d}_{\mathbf{qijlm}} = (\mathbf{x}'_{\mathbf{qm}}, \mathbf{h}'_{\mathbf{qijl}})'$. Then, the micro- and macro-models of Equations (5.2) and (5.3) can be combined into a single equation as follows:

$$\lambda_{qijlm}(\tau) = \lambda_{m0}(\tau)\exp(\boldsymbol{\gamma}'_{\mathbf{m}}\mathbf{d}_{\mathbf{qijlm}} + v_q + u_i + u_{im} + w_j + w_{jm} + z_l + z_{lm} + \omega_{qm}), \qquad (5.4)$$

The proportional hazard formulation of Equation (5.4) can be written equivalently in terms of the logarithm of the integrated hazard at continuous time T_{qijlm} as follows (see Bhat and Pinjari, 2008):

$$s^*_{qijlm} = -\ln \Lambda_0(T_{qijlm}) = -\ln \int_{\tau=0}^{T_{qijlm}} \lambda_{m0}(\tau)d\tau = \boldsymbol{\gamma}'_{\mathbf{m}}\mathbf{d}_{\mathbf{qijlm}} + v_q + u_i + u_{im} + w_j + w_{jm} + z_l + z_{lm} + \omega_{qm} + \varepsilon_{qm} \;, \qquad (5.5)$$

ε_{qm} in the above equation occurs because of the intrinsic probabilistic nature of the hazard function. Further, when the relationship between the hazard function and covariates takes the proportional hazard form of Equation (5.4), it is straightforward to show that ε_{qm} is standard extreme value distributed: $\Pr(\varepsilon_{qm} < a) = G(a) = \exp[-\exp(-a)]$. In Equation (5.5), since each individual q is uniquely identified with a particular household i, social cluster j, and spatial cluster l, it is convenient from a presentation standpoint to suppress the indices i, j, and l in T_{qijlm} and d_{qijlm}. Thus, hereafter we will use the notation T_{qm} for T_{qijlm}, and d_{qm} for d_{qijlm}.

Now, consider the case where time T_{qm} is unobservable on the continuous scale, but is observed in grouped (or discrete) intervals t_{qm}. In the empirical context of the current study, this grouping is a result of individuals rounding off activity durations when reporting time-use patterns in activity-travel surveys. That individuals do so has now been well established in earlier studies (see Bhat, 1996, Hautsch, 1999). For instance, individuals tend to round off to the closest five minutes for activity participations of durations less than an hour, and then round off to the closest 10-15 minutes beyond an hour. The net result of such rounding is that there is clumping or "ties" in the data at durations of time that are integer multiples of five minutes. The presence of such ties renders usual parametric continuous baseline hazard models inappropriate, since these models use density function terms in the likelihood function that are appropriate only for continuous duration data. In particular, if the typical continuous hazard model frameworks are directly applied to model grouped data durations, the resulting estimates would generally be inconsistent (Prentice and Gloeckler, 1978). Thus, it is important to explicitly recognize the interval-level data arising from the grouping of underlying continuous times during the estimation process. To do so, consider k as an index for grouped time intervals (*i.e.*, t_{qm}= 0, 1, 2,…, k,…, K_m). In the analysis, we will assume that the first grouped time interval (t_{qm}= 0) corresponds to non-participation in the activity type, and we assign a low duration upper bound of continuous time (say $b_{m,1}$) for this first grouped interval.[27] Note that $b_{m,1}$ also constitutes the lower bound for the second

[27] Another option would be to explicitly model participation in activity type *m*, along with the discrete interval of participation in activity type *m* conditional on a positive participation decision. One can pursue such an exercise either by using separate models of discrete-continuous choice systems for each activity type *m* (see, for example, Bhat and Eluru (2009) and Genius and Strazzera (2008) for recent general frameworks for these modeling systems) or by employing a multiple discrete-continuous extreme value (MDCEV) model (see, for example, Bhat 2008). The first approach ignores the inter-relationship in time-use across activity types. To be sure, jointness may be added across activity types within this framework, but the structure gets extremely cumbersome in doing so. The second MDCEV approach is a convenient way to handle discrete-continuous choices across multiple activities, but is relatively limited in the richness of substitution structures allowed across activity types. It also gets somewhat unwieldy when trying to incorporate complementary effects across activity types in participations and participation durations. Further, in both these approaches, it is practically infeasible to incorporate random unobserved clustering effects along the multi-level and cross-level dimensions associated with the individual, the household, the social grouping, and the spatial grouping. On the other hand, the focus of this study is on accommodating such multi-level and cross-level clustering effects. At the same time, as we discuss later in the data section, individuals who invest some time in walking and bicycling activity over the course of the week do so for a reasonably high minimum duration. Thus, it is not unreasonable to assign a low duration threshold as the upper bound of the first time interval category, which is designated as the non-participation category. Also,

time interval, while the value of zero constitutes the lower bound for the first (non-participation time interval). More generally, let $b_{m,k+1}$ be the upper bound on the continuous time scale corresponding to the grouped time interval *k*. Then, we may write equation (5.5) in an equivalent grouped response form as follows:

$$s^*_{qijlm} = -\ln\Lambda_0(T_{qm}) = \boldsymbol{\gamma}'_{\mathbf{m}}\mathbf{d}_{\mathbf{qm}} + v_q + u_i + u_{im} + w_j + w_{jm} + z_l + z_{lm} + \omega_{qm} + \varepsilon_{qm}\,,\;\; t_{qm} = k$$

$$\text{if } \psi_{m,K_m-k} < s^*_{qm} < \psi_{m,K_m+1-k}\,, \qquad (5.6)$$

where $\psi_{m,K_m+1-k} = -\ln\Lambda_0(b_{m,k})$ is the upper bound for interval *k* for activity type *m* $(\psi_{m,0} < \psi_{m,1} < \psi_{m,2} \ldots < \psi_{m,K_m+1};\ \psi_{m,0} = -\infty, \psi_{m,K_m+1} = +\infty)$.[28]

In the above specification, if θ^2 (variance of v_q), μ^2 (variance of u_i), μ_m^2 (variance of u_{im}; $m = 1,\ldots,M$), η^2 (variance of w_j), η_m^2 (variance of $w_{jm}; m = 1,\ldots,M$), δ^2 (variance of z_l), and δ_m^2 (variance of $z_{lm}; m = 1,\ldots,M$) are all simultaneously equal to zero, then it implies that there is no variation in the activity durations for different activity types based on unobserved factors that are specific to the individual, the household, the social cluster to which the individual belongs, and the spatial cluster to which the individual belongs. In

the non-linear nature of the grouped duration model structure we use in the current study is flexible enough to accommodate large fractions of individuals falling in the non-participation category. Overall, the grouped duration modeling structure adopted in the current study is ideal for the empirical analysis at hand and lends itself well to estimation using the composite marginal likelihood approach.

[28] Note that once the threshold bounds are estimated, the analyst can work backwards from there to obtain the baseline hazard shape by using the relationship $\psi_{m,K_m+1-k} = -\ln\Lambda_0(b_{m,k})$. Specifically, assume a constant hazard for all continuous time durations τ_{mk} that fall in interval *k* for activity type *m* $(b_{m,k} < \tau_{mk} < b_{m,k+1})$, and $\Delta\tau_{mk}$ be the length of the time interval *k* for activity type *m*. Then, $\hat{\lambda}_0(\tau_{mk}) = \dfrac{\exp(-\hat{\psi}_{m,K_m-k}) - \exp(-\hat{\psi}_{m,K_m+1-k})}{\Delta\tau_{mk}}, k = 1,2,\ldots,K_m;\ m = 1,\ldots,M.$ Also, because the continuous time bounds for each grouped time interval are known a priori, and the estimated thresholds in the ordered-response structure of Equation (6) are (negative of) the logarithm of the integrated baseline hazard values, there is no need for any normalization of the scale associated with the underlying "latent" variable s^*_{qijlm}, as in a typical ordered-response model (see Meyer, 1990, and Bhat and Pinjari, 2008). That is, it is possible to estimate the variance of ω_{qm} $(i.e., \sigma_m^2)$.

this case, the cross-random grouped response (CRGR) model of Equation (5.6) collapses to the standard grouped response (SGR) model. The implication is that all unobserved heterogeneity is due to overall idiosyncratic factors associated with the propensity to participate in each activity type, and there are no common unobserved individual, household, social group, and spatial cluster factors impacting durations of participation in the activity types. Note also that the specification of Equation (5.6) generates a rich covariance pattern structure among the hazard functions for participation in different activity types. The (log) integrated hazards (LIHs) for any pair of activity types *m* and m' ($m \neq m'$) for the same individual have a covariance of $\theta^2 + U_q \mu^2 + \eta^2 + \delta^2$, where $U_q = 1$ if the individual is in a household with more than one individual and zero otherwise. For two different individuals q and q', the covariance in the LIHs between any pairing of activity types m and m' across the two individuals is equal to $H_{qq'}\mu^2 + H_{qq'mm'}\mu_m^2 + R_{qq'}\eta^2 + R_{qq'mm'}\eta_m^2 + G_{qq'}\delta^2 + G_{qq'mm'}\delta_m^2$, where $H_{qq'}$ = 1 if individuals q and q' are in the same household, $H_{qq'mm'}$ = 1 if individuals q and q' are in the same household and *m* and m' are the same activity type, $R_{qq'}$ = 1 if individuals q and q' are in the same social cluster, $R_{qq'mm'}$ = 1 if individuals q and q' are in the same social cluster and *m* and m' are the same activity type, $G_{qq'}$ = 1 if individuals q and q' are in the same spatial cluster, and $G_{qq'mm'}$ = 1 if individuals q and q' are in the same spatial cluster and *m* and m' are the same activity type. The indicator variables above take the value of zero otherwise.

5.2.2 Estimation Approach

Let y_{qm} be the q^{th} individual's observed activity participation time (obtained in the grouped intervals) in activity type *m*. The conditional likelihood function for individual *q*'s participation duration in activity type *m* (conditional on $v_q, u_i, u_{im}, w_j, w_{jm}, z_l, z_{lm}$ and ω_{qm}) can be written as:

$$L_{qm}\big|v_q, u_i, u_{im}, w_j, w_{jm}, z_l, z_{lm}, \omega_{qm} = G[\psi_{K_m+1-y_{qm}} - B_{qm}] - G[\psi_{K_m-y_{qm}} - B_{qm}]$$

Where $B_{qm} = \boldsymbol{\gamma}'_{\mathbf{m}}\mathbf{d}_{\mathbf{qm}} + v_q + u_i + u_{im} + w_j + w_{jm} + z_l + z_{lm} + \omega_{qm}$

The likelihood function unconditional on ω_{qm} is:

$$L_{qm}\big|v_q, u_i, u_{im}, w_j, w_{jm}, z_l, z_{lm} = \int_{\omega_{qm}} (G[\psi_{K_m+1-y_{qm}} - B_{qm}] - G[\psi_{K_m-y_{qm}} - B_{qm}])dF(\omega_{qm}),$$

where $F(\omega_{qm})$ is the univariate cumulative normal distribution function corresponding to ω_{qm}. The likelihood function of the entire sample cannot be broken down as the product of the likelihood functions for each individual's choices of grouped time interval for each activity *m*, because the underlying latent values s^*_{qijlm} are correlated across individuals *q* and activities *m* (due to the presence of the $v_q, u_i, u_{im}, w_j, w_{jm}, z_l,$ and z_{lm} error terms). Further, since the various clusters are not hierarchical (*i.e.*, one cluster is not nested within the other), the analyst needs to consider the entire set of $(Q \times M)$ observations (*q* = 1, 2, …, *Q*; *m* = 1, …, *M*) as a single cluster in developing the likelihood function. To do so, stack the s^*_{qijlm} values together vertically in the vector $\mathbf{s}^*$, and let the implied variance-covariance of $\mathbf{s}^*$ due to the $v_q, u_i, u_{im}, w_j, w_{jm}, z_l,$ and z_{lm} (but not considering ω_{qm} and ε_{qm}) error terms be $\boldsymbol{\Omega}$. Thus $\boldsymbol{\Omega}$ is a $[(M \times Q) \times (M \times Q)]$ variance-covariance matrix whose elements are parameterized based on $\theta^2, \mu^2, \mu_m^2, \eta^2, \eta_m^2, \delta^2,$ and δ_m^2. Define a multivariate normally distributed variable vector $g \sim MVN_{Q \times M}(\mathbf{0}, \boldsymbol{\Omega})$. Then the likelihood function may be written as:

$$L = \int_g \prod_{q=1}^{q} \prod_{m=1}^{M} (L_{qm}\big|v_q, u_i, u_{im}, w_j, w_{jm}, z_l, z_{lm})dF_{Q \times M}(g\big|\boldsymbol{\Omega})$$

The likelihood function above entails the evaluation of an integral of the order of $(Q \times M)$. The usual simulation techniques become impractical, if not infeasible, to evaluate such a multidimensional integral for even small to moderate *Q*, as discussed earlier. Thus, we employ the composite marginal likelihood (CML) technique in the current study. For this, define the parameter vector to be estimated as:

$\boldsymbol{\kappa} = (\boldsymbol{\gamma}'_1, ..., \boldsymbol{\gamma}'_{\mathbf{M}};\ \boldsymbol{\psi}'_1, ..., \boldsymbol{\psi}'_{\mathbf{M}}; \sigma_1, ..., \sigma_M; \mu_1, ..., \mu_M; \eta_1, ..., \eta_M; \delta_1, ..., \delta_M, \theta, \mu, \eta, \delta)'$, where $\boldsymbol{\psi}_{\mathbf{m}} = (\psi_{m,1}, \psi_{m,2}, ..., \psi_{m,K_m})'$. The pairwise marginal likelihood function includes two main components – one component that represents the likelihood of all pairs of activity type combinations within individuals, and the second component that represents the likelihood of pairs of individual-activity type combinations across individuals:

$$L_{CML}(\boldsymbol{\kappa}) =$$

$$\left[\prod_{q=1}^{Q}\prod_{m=1}^{M-1}\prod_{m'=m+1}^{M}\int_{\varepsilon_{qm'}=-\infty}^{+\infty}\int_{\varepsilon_{qm}=-\infty}^{+\infty}\left[\Phi_2(m_{y_{qm}+1}, m_{y_{qm'}+1}, \theta_{qmm'}) - \Phi_2(m_{y_{qm}}, m_{y_{qm'}+1}, \theta_{qmm'}) - \Phi_2(m_{y_{qm}+1}, m_{y_{qm'}}, \theta_{qmm'}) + \Phi_2(m_{y_{qm}}, m_{y_{qm'}}, \theta_{qmm'})\right]dF(\varepsilon_{qm})dF(\varepsilon_{qm'})\right] \times$$
$$\left[\prod_{q=1}^{Q-1}\prod_{q'=q+1}^{Q}\prod_{m=1}^{M}\prod_{m'=1}^{M}\int_{\varepsilon_{q'm'}=-\infty}^{+\infty}\int_{\varepsilon_{qm}=-\infty}^{+\infty}\left[\Phi_2(m_{y_{qm}+1}, m_{y_{q'm'}+1}, \theta_{qq'mm'}) - \Phi_2(m_{y_{qm}}, m_{y_{q'm'}+1}, \theta_{qq'mm'}) - \Phi_2(m_{y_{qm}+1}, m_{y_{q'm'}}, \theta_{qq'mm'}) + \Phi_2(m_{y_{qm}}, m_{y_{q'm'}}, \theta_{qq'mm'})\right]dF(\varepsilon_{qm})dF(\varepsilon_{q'm'})\right] \tag{5.7}$$

where

$$m_{y_{qm}} = \left[\frac{\psi_{K_m - y_{qm}} - \boldsymbol{\gamma}'_{\mathbf{m}}\mathbf{d}_{\mathbf{qm}} - \varepsilon_{qm}}{\sqrt{\theta^2 + U_q\mu^2 + \eta^2 + \delta^2 + U_q\mu_m^2 + \eta_m^2 + \delta_m^2 + \sigma_m^2}}\right],$$

$$\theta_{qmm'} = \left[\frac{\theta^2 + U_q\mu^2 + \eta^2 + \delta^2}{\left(\sqrt{\theta^2 + U_q\mu^2 + \eta^2 + \delta^2 + U_q\mu_m^2 + \eta_m^2 + \delta_m^2 + \sigma_m^2}\right)\left(\sqrt{\theta^2 + U_q\mu^2 + \eta^2 + \delta^2 + U_q\mu_{m'}^2 + \eta_{m'}^2 + \delta_{m'}^2 + \sigma_{m'}^2}\right)}\right], m \neq m'$$

and

$$\theta_{qq'mm'} = \left[\frac{H_{qq'}\mu^2 + H_{qq'mm'}\mu_m^2 + R_{qq'}\eta^2 + R_{qq'mm'}\eta_m^2 + G_{qq'}\delta^2 + G_{qq'mm'}\delta_m^2}{\left(\sqrt{\theta^2 + U_q\mu^2 + \eta^2 + \delta^2 + U_q\mu_m^2 + \eta_m^2 + \delta_m^2 + \sigma_m^2}\right)\left(\sqrt{\theta^2 + U_{q'}\mu^2 + \eta^2 + \delta^2 + U_{q'}\mu_{m'}^2 + \eta_{m'}^2 + \delta_{m'}^2 + \sigma_{m'}^2}\right)}\right], q \neq q'.$$

In the CML function above, $F(.)$ is the univariate cumulative standard type I extreme value distribution. In Equation (5.7), the integration can be carried out using quadrature techniques or simulation techniques. However, an alternative is to use the normal scale mixture (NSM) representation of the extreme value distribution. That is, we remove the non-normality of the error term ε by replacing it with a weighted mixture of normally distributed variables (see Choy and Chan, 2008 and Bhat, 2011 for explanations and recent applications of NSM. Also, the reader is referred to a special issue of *Computational Statistics & Data Analysis* edited by Böhning and Seidel (2003) for recent developments on this topic). The NSM technique can be applied using standard statistical software packages, and is very efficient. In the context of the current model, Equation (5.5) can be re-written for the *r*^th component of the normal scale mixture of the extreme value distribution ε_{qm} as follows:

$$s^{*r}_{qijlm} = -\ln \Lambda^{r}_{0}(T_{qm}) = \boldsymbol{\gamma}'_{\mathbf{m}}\mathbf{d}_{\mathbf{qm}} + v_q + u_i + u_{im} + w_j + w_{jm} + z_l + z_{lm} + \omega_{qm} + \pi_{qmr} \ , \quad (5.8)$$

where $\pi_{qmr} \sim N(a_{qmr}, b^2_{qmr})$

Assume that the weight for this *r*^th component is $p_r \ (\sum_{r=1}^{R} p_r = 1)$. Then, following through on the usual CML approach, the CML likelihood function contribution for each activity type pairing from the same individual *q* for the *r*^th normal scale mixture component for ε_{qm} and the *e*^th normal scale mixture component for $\varepsilon_{qm'}$ may be written as:

$$L^{re}_{qmm',CML}(\boldsymbol{\kappa}) = \left[\Phi_2(m^r_{y_{qm}+1}, m^e_{y_{qm'}+1}, \theta^{re}_{qmm'}) - \Phi_2(m^r_{y_{qm}}, m^e_{y_{qm'}+1}, \theta^{re}_{qmm'}) - \Phi_2(m^r_{y_{qm}+1}, m^e_{y_{qm'}}, \theta^{re}_{qmm'}) + \Phi_2(m^r_{y_{qm}}, m^e_{y_{qm'}}, \theta^{re}_{qmm'})\right] \quad (5.9)$$

where

$$m^r_{y_{qm}} = \left[\frac{\psi_{K_m - y_{qm}} - \boldsymbol{\gamma}'_{\mathbf{m}}\mathbf{d}_{\mathbf{qm}} - a_{qmr}}{\sqrt{\theta^2 + U_q\mu^2 + \eta^2 + \delta^2 + U_q\mu^2_m + \eta^2_m + \delta^2_m + \sigma^2_m + b^2_{qmr}}}\right],$$

$$m^e_{y_{qm'}} = \left[\frac{\psi_{K_{m'} - y_{qm'}} - \boldsymbol{\gamma}'_{\mathbf{m'}}\mathbf{d}_{\mathbf{qm'}} - a_{qm'e}}{\sqrt{\theta^2 + U_q\mu^2 + \eta^2 + \delta^2 + U_q\mu^2_{m'} + \eta^2_{m'} + \delta^2_{m'} + \sigma^2_{m'} + b^2_{qm'e}}}\right],$$

$$\theta^{re}_{qmm'} = \left[\frac{\theta^2 + U_q\mu^2 + \eta^2 + \delta^2}{\left(\sqrt{\theta^2 + U_q\mu^2 + \eta^2 + \delta^2 + U_q\mu^2_m + \eta^2_m + \delta^2_m + \sigma^2_m + b^2_{qmr}}\right)\left(\sqrt{\theta^2 + U_q\mu^2 + \eta^2 + \delta^2 + U_q\mu^2_{m'} + \eta^2_{m'} + \delta^2_{m'} + \sigma^2_{m'} + b^2_{qm'e}}\right)}\right],$$

The likelihood function contribution from each activity type pairing of the same individual then may be obtained by taking the weighted average of Equation (5.9) over all {*r*,*e*} mixture components as follows:

$$L_{qmm',CML}(\boldsymbol{\kappa}) = \sum_{e=1}^{R}\sum_{r=1}^{R} p_e p_r L^{re}_{qmm',CML}(\boldsymbol{\kappa}), \ m \neq m'.$$

Similarly, the likelihood contribution from each activity type pairing across individuals may be obtained as follows:

$$L_{qq'mm',CML}(\boldsymbol{\kappa}) = \sum_{e=1}^{R}\sum_{r=1}^{R} p_e p_r L^{re}_{qq'mm',CML}(\boldsymbol{\kappa}), q \neq q',$$

where

$$L^{re}_{qq'mm',CML}(\boldsymbol{\kappa}) = \left[\Phi_2(m^r_{y_{qm}+1}, m^e_{y_{q'm'}+1}, \theta^{re}_{qq'mm'}) - \Phi_2(m^r_{y_{qm}}, m^e_{y_{q'm'}+1}, \theta^{re}_{qq'mm'}) - \Phi_2(m^r_{y_{qm}+1}, m^e_{y_{q'm'}}, \theta^{re}_{qq'mm'}) + \Phi_2(m^r_{y_{qm}}, m^e_{y_{q'm'}}, \theta^{re}_{qq'mm'})\right] \quad ,$$

and

$$\theta^{re}_{qq'mm'} = \left[\frac{H_{qq'}\mu^2 + H_{qq'mm'}\mu_m^2 + R_{qq'}\eta^2 + R_{qq'mm'}\eta_m^2 + G_{qq'}\delta^2 + G_{qq'mm'}\delta_m^2}{\left(\sqrt{\theta^2 + U_q\mu^2 + \eta^2 + \delta^2 + U_q\mu_m^2 + \eta_m^2 + \delta_m^2 + \sigma_m^2 + b_{qmr}^2}\right)\left(\sqrt{\theta^2 + U_{q'}\mu^2 + \eta^2 + \delta^2 + U_{q'}\mu_{m'}^2 + \eta_{m'}^2 + \delta_{m'}^2 + \sigma_{m'}^2 + b_{q'm'e}^2}\right)}\right], q \neq q'.$$

Finally, the overall CML function may be written as:

$$L_{CML}(\boldsymbol{\kappa}) = \left[\prod_{q=1}^{Q}\prod_{m=1}^{M-1}\prod_{m'=m+1}^{M} L_{qmm',CML}(\boldsymbol{\kappa})\right] \times \left[\prod_{q=1}^{Q-1}\prod_{q'=q+1}^{Q}\prod_{m=1}^{M}\prod_{m'=1}^{M} L_{qq'mm',CML}(\boldsymbol{\kappa})\right] \qquad (5.10)$$

In the current study, we use 5 mixture components to approximate the extreme value error term, based on the weights and normal distribution approximation for each mixture provided by Frühwirth-Schnatter and Wagner (2005).[29]

The pairwise estimator $\hat{\boldsymbol{\kappa}}_{CML}$ is obtained by maximizing the logarithm of the likelihood function given in Equation (5.10). As indicated earlier, the covariance matrix, given by the inverse of Godambe's sandwich information matrix $(\mathbf{G}(\boldsymbol{\kappa}))$, is as follows:

$$\mathbf{V}_{CML}(\boldsymbol{\kappa}) = [\mathbf{G}(\boldsymbol{\kappa})]^{-1} = [\mathbf{H}(\boldsymbol{\kappa})]^{-1}\mathbf{J}(\boldsymbol{\kappa})[\mathbf{H}(\boldsymbol{\kappa})]^{-1},$$

The $\mathbf{H}(\boldsymbol{\kappa})$ matrix can be estimated as:

[29] We carried out a preliminary analysis which indicated that 5 mixture components to approximate the extreme value error term are adequate for the current analysis. However, the methodology provided in this study is generic and can be applied with any number of mixture components. The values for p_r, a_{qmr}, and b_{qmr} parameters are provided in the paper by Frühwirth-Schnatter and Wagner (FSW, 2005). Note, FSW present the results for minimum extreme value type I distribution while the current study uses maximum extreme value type I distribution. We apply the parameters provided in FSW after switching the sign of a_{qmr} for each mixture component.

$$\hat{\mathbf{H}}(\hat{\boldsymbol{\kappa}}) = -\left[\sum_{q=1}^{Q}\sum_{m=1}^{M-1}\sum_{m'=m+1}^{M}\frac{\partial^2 \log L_{qmm',CML}(\boldsymbol{\kappa})}{\partial\boldsymbol{\kappa}\partial\boldsymbol{\kappa}'} + \sum_{q=1}^{Q-1}\sum_{q'=q+1}^{Q}\sum_{m=1}^{M}\sum_{m'=1}^{M}\frac{\partial^2 \log L_{qq'mm',CML}(\boldsymbol{\kappa})}{\partial\boldsymbol{\kappa}\partial\boldsymbol{\kappa}'}\right]_{\hat{\boldsymbol{\kappa}}}$$

However, the estimation of the $\mathbf{J}(\boldsymbol{\kappa})$ matrix is more difficult, since the term $\left[\sum_{q=1}^{Q-1}\sum_{q'=q+1}^{Q}\sum_{m=1}^{M}\sum_{m'=1}^{M}\frac{\partial \log L_{qq'mm',CML}(\boldsymbol{\kappa})}{\partial\boldsymbol{\kappa}}\right]$ vanishes when evaluated at the CML estimate $\hat{\boldsymbol{\kappa}}_{CML}$. Further, one cannot estimate $\mathbf{J}(\boldsymbol{\kappa})$ as the sampling variance of individual contributions to the composite score function because of the underlying household-level, social, and spatial dependence in the observations. Hence we resort to pure Monte Carlo computation to estimate the $\mathbf{J}(\boldsymbol{\kappa})$ matrix. In this approach, we generate *B* data sets (T^1, T^2,..., T^B) where each dataset T^b (*b*=1,2,…, *B*) is a $(Q \times M)$ matrix of the dependent variables generated using the exogenous variables and the CML estimates $(\hat{\boldsymbol{\kappa}}_{CML})$. Once these datasets are generated, the estimate of $\mathbf{J}(\boldsymbol{\kappa})$ is given by:

$$\hat{\mathbf{J}}(\hat{\boldsymbol{\kappa}}) = \frac{1}{B}\sum_{b=1}^{B}\left[\left(\frac{\partial \log L_{CML}(\boldsymbol{\kappa})}{\partial\boldsymbol{\kappa}}\right)_{T^b}\left(\frac{\partial \log L_{CML}(\boldsymbol{\kappa})}{\partial\boldsymbol{\kappa}'}\right)_{T^b}\right]_{\hat{\boldsymbol{\kappa}}}$$

The above computation is not very demanding because the model in Equation (5.6) can be generated in a straight-forward manner. We tested various values of *B* to ensure the stability and a reasonable level of accuracy in the estimation of $\mathbf{J}(\boldsymbol{\kappa})$ matrix.

5.3 Data

5.3.1 Data Source

The data set used in the current study is drawn from the 2009 National Household Travel Survey (NHTS, 2009). The NHTS is a national survey that collects information on all trips undertaken by all individuals (age 5 years or older) in a large sample of households from across the United States for a 24-hour period. The 2009 NHTS collected such information for all individuals in a sample of more than 150,000 households. Information collected includes, for example, trip start and end time, purpose, mode of travel, composition of travel party, and trip length. In addition, individual- and household-level socio-demographics, data on internet use, regional location, and characteristics of the survey day are also collected.

In the current study, data collected for households drawn from Marin, Solano, and Sonoma counties in the San Francisco Bay Area is used. We used California-specific NHTS data because the NHTS add-on survey instrument for California collected detailed information on walking and bicycling activity duration for all individuals (5 years of age or above) over a period of one week. Such information was not available from the general (non-California) NHTS sample. Also, as indicated earlier, the NHTS California data set contains attitudinal information on individuals participating in walking and bicycling activities (more on this in Section 5.3.3). Within the California data set, the sample from the three specific counties listed above was selected for analysis because the we have access to extensive secondary data on built environment attributes for these locations.

5.3.2 Sample Formation

Several steps were necessary to extract information from the NHTS data set and obtain the final sample for the analysis.[30] First, only individuals who were at least 5 years old were selected, because the walking and bicycling activity durations were collected only for individuals in this age group. Second, all individuals who did not participate in at least one activity (*i.e.*, either walking or bicycling) over a period of one week were removed from the data file. Third, the walking and bicycling activity durations, which were reported in hours and minutes, were converted to minutes. Then, the continuous activity durations were divided into grouped intervals and indexed appropriately (see Section 5.2.1 for details). Fourth, an indicator variable was generated to identify individuals from the same family/household. Fifth, we considered a number of demographic factors such as individual's age, household structure, and household income to define social grouping.[31] However, a preliminary analysis indicated that, because of

[30] Note that the NHTS "person" file was used as the source file for the current analysis.

[31] In the current study, we employ an egocentric approach to define social cluster. In this approach, individuals are divided in to social groups based on their demographics and/or attitudes toward joint

the similarity in activity participation patterns among individuals within a certain age range, using age to define social groups would give the best model specification. All individuals were grouped in to one of nine mutually exclusive and collectively exhaustive social groups defined as follows: 5≤ age ≤10, 11≤ age ≤13, 14≤ age ≤15, 16≤ age ≤25, 26≤ age ≤35, 36≤ age ≤45, 46≤ age ≤55, 56≤ age ≤65, and age> 65. Note that the first three social groups correspond to the age groups for elementary, middle, and high school going children. Sixth, the traffic analysis zone (TAZ) was used for spatial clustering. [32] The residential location of each household was geo-located to a TAZ; thus, all households that reside in a TAZ belong to the same spatial cluster. Seven, data on household socio-demographics, built environment characteristics, and information on the survey day were appended to the data file. Finally, several screening and consistency checks were performed and records with missing or inconsistent data were eliminated.

5.3.3 Attitudinal Variables

Individuals who participated in walking and/or bicycling activity over a period of one week were asked a series of questions designed to reveal their attitudes/beliefs towards these activities. Information was collected for walking activity and bicycling activity separately. Collected data includes information on individuals' lifestyle, health condition, available walking (or bicycling) facilities/environment in the neighborhood, traffic and crime related safety concerns, air pollution, and attitude towards motorized traffic (see NHTS 2009 for a complete list of the questions).

activities (the reader is referred to Dugundji and Walker, 2005 and Carrasco *et al.*, 2008 for more information on egocentric approach). Thus, this approach has the advantage of being able to use readily available individual-level demographic information to define clustering scheme. Of course, the methodology proposed in the current study is generic and can accommodate any types of social clustering scheme.

[32] In our analysis, we used the 1990 MTC Travel Analysis Zones system for the San Francisco Bay Area (http://www.mtc.ca.gov/maps_and_data/GIS/data.htm). Please note that for confidentiality-related reasons, information on residential location was available only at the Census tract level. As a result, some assumptions/aggregations were necessary to definite the spatial clustering scheme.

A factor analysis was performed to reduce the number of variables and to obtain a more compact set of influential factors. The factor analysis was undertaken using principal components estimation and varimax rotation (Gorsuch, 1983, Kline, 1994). After the factor analysis was performed and the principal components were identified, we discarded the factor loadings and assigned a unit weight to each identified component, which is subsequently distributed equally between the relevant factors. This approach allowed us to retain all the attitudinal information with sufficient number of records while keeping the number of identified components at a manageable level (hereafter, we shall refer to the identified principal components as attitudinal variables). Tables 5.1a and 5.1b present the definition of the attitudinal variables (identified through the factor analysis) for walking and bicycling activity, respectively.

5.3.4 Sample Description

The final sample for analysis comprises of 882 individuals (age 5 years or above) from 561 households. Of these individuals, 96.1% participate in some walking activity and 18.9% participate in some bicycling activity over a period of one week. Individuals who participate in these activities spend, on average, 204 minutes and 130 minutes per week in walking and bicycling activity, respectively.

Table 5.2 provides information on walking and bicycling activity durations for individuals who participate in these activities. The lengths of the discrete periods used in estimation (presented in the third column) increase for larger activity durations until termination for all individuals (except for the first period for walking activity which is 10 minutes long). The number of discrete periods used for walking is higher than for bicycling because of the more extensive number of individuals walking in the sample, thus providing adequate number of individuals in finer time periods. For the final discrete period, all spells longer than 840 minutes for walking and 240 minutes for bicycling are collapsed to a single period.

The discrete-period sample hazards (the sixth column) are estimated using Kaplan-Meier non-parametric estimator (Kiefer, 1988). The hazards are transformed to

Table 5.1a Definition of Attitudinal Variables – Reasons for Not Walking More Frequently

Factor	Attitudinal Variable				
	Absence of "attractions" and busy life style related factors	Inconvenience	Unavailability of walk-friendly environment/ facilities	(Lack of) Walking conditions due to the motorized vehicles related factors	(Lack of) Safety
You're too busy?	0.34				
You have things to carry?	0.33				
No shops or other interesting places to go?	0.33				
No one to walk with?		0.50			
You have small children along?		0.50			
No nearby paths or trails?			0.25		
No nearby parks?			0.25		
No sidewalks or the sidewalks are in poor condition?			0.25		
Not enough people walking around?			0.25		0.25
There are too many cars?				0.50	
Of fast traffic?				0.50	
Not enough light at night?					0.25
You fear street crime?					0.25
Street crossings are unsafe?					0.25

Table 5.1b Definition of Attitudinal Variables – Reasons for Not Bicycling More Frequently

Factor	Attitudinal Variable			
	Busy life style and absence of bicycle paths/trails	Inconvenience and lack of paved bicycle facilities	Unavailability of proper cycling facilities/ conditions	(Lack of) Safety
You're too busy?	0.33			
You have small children along?	0.33			
No nearby paths or trails?	0.34			
Not enough bike or wide curb lanes?		0.50		
You have too many things to carry?		0.50		
Not enough light at night?			0.25	
No sidewalks or the sidewalks are in poor condition?			0.25	
There are too many cars?			0.25	0.25
Of fast traffic?			0.25	0.25
Street crossings are unsafe?				0.25
You have no one to bike with?				0.25

continuous-time sample hazards and are plotted in Figure 5.1.[33] These plots show that the sample hazards are higher for bicycling activity duration compared to walking activity duration in the first 45 minutes. This implies that individuals who participate in walking activity tend to commit a certain minimum amount of time to pursue this activity. Also, walking activity duration hazards exhibit more widespread “peaks” than bicycling activity duration hazards. This indicates a more even distribution of walking activity durations across participating individuals in comparison to bicycling activity durations. Hazard function for walking duration exhibits three highest “peaks” at integer multiples of 1-hour (*i.e.*, at time periods containing 1-hour, 2-hour, and 3-hour walking activity durations per week). Other “peaks” in the plot of hazard function for walking can be observed at multiples of 30 minutes intervals. A similar trend, but to a lesser degree can be observed in the plot of the hazard function for bicycling duration. This pattern of hazard functions highlights the discrete interval nature of reporting of the underlying continuous time variable and the need to adopt an appropriate framework that can explicitly recognize this feature. The model system proposed in the current study incorporates this ability.

5.4 Empirical Analysis

5.4.1 Variable Specification

Several types of variables were considered in the model specification. These included individual socio-demographics, household socio-demographics, and attitudinal variables. In addition to these three groups of variables, different function forms and interaction

[33] The discrete-period sample hazards cannot be compared directly across period due to variation in the length of time period. So, we convert them to continuous-time sample hazard under the assumption that hazard is constant within each period *k*. Thus, continuous-time sample hazard $\hat{\lambda}_{m0}(k)$ can be estimated as follows: $\hat{\lambda}_{m0}(k) = -\dfrac{\ln(1-\hat{\lambda}^{*}_{m0}(k))}{\Delta t(k)}$,

where $\hat{\lambda}^{*}_{m0}(k)$ is the discrete-period sample hazard in period *k* and $\Delta t(k)$ is the length of the period *k*.

Table 5.2 Walking and Bicycling Activity Durations and the Discrete Period Sample Hazards

Discrete time period k[34]	Time interval t (mins)	Interval length (mins)	No. of individuals terminating activity participation in this time period (F_k)	No. of individuals "at risk" of terminating activity participation in this time period (R_k)	Discrete-period hazard $\left(H_k = \frac{F_k}{R_k}\right)$	Standard error of H_k[35]
Walking activity duration						
1	$0 < t \leq 10$	10	7	848	0.008	0.003
2	$10 < t \leq 15$	5	8	841	0.010	0.003
3	$15 < t \leq 20$	5	22	833	0.026	0.006
4	$20 < t \leq 30$	10	37	811	0.046	0.007
5	$30 < t \leq 40$	10	16	774	0.021	0.005
6	$40 < t \leq 50$	10	25	758	0.033	0.006
7	$50 < t \leq 60$	10	106	733	0.145	0.013
8	$60 < t \leq 80$	20	11	627	0.018	0.005
9	$80 < t \leq 100$	20	85	616	0.138	0.014
10	$100 < t \leq 120$	20	116	531	0.218	0.018
11	$120 < t \leq 150$	30	36	415	0.087	0.014
12	$150 < t \leq 180$	30	90	379	0.237	0.022
13	$180 < t \leq 210$	30	27	289	0.093	0.017
14	$210 < t \leq 240$	30	47	262	0.179	0.024
15	$240 < t \leq 300$	60	55	215	0.256	0.030
16	$300 < t \leq 360$	60	47	160	0.294	0.036
17	$360 < t \leq 420$	60	31	113	0.274	0.042
18	$420 < t \leq 480$	60	14	82	0.171	0.042
19	$480 < t \leq 600$	120	34	68	0.500	0.061
20	$600 < t \leq 720$	120	9	34	0.265	0.076
21	$720 < t \leq 840$	120	10	25	0.400	0.098
22	$840 < t$	> 120	15	15	1.000	-
Bicycling activity duration						
1	$0 < t \leq 15$	15	8	167	0.048	0.017
2	$15 < t \leq 30$	15	22	159	0.138	0.027
3	$30 < t \leq 45$	15	17	137	0.124	0.028
4	$45 < t \leq 60$	15	25	120	0.208	0.037
5	$60 < t \leq 90$	30	20	95	0.211	0.042
6	$90 < t \leq 120$	30	20	75	0.267	0.051
7	$120 < t \leq 180$	60	26	55	0.473	0.067
8	$180 < t \leq 240$	60	10	29	0.345	0.088
9	$240 < t$	> 60	19	19	1.000	-

[34] Note that in the estimated model *k* starts from 0 which represents non-participation in the activity.

[35] Standard error of H_k is estimated using Greenwood's formula.

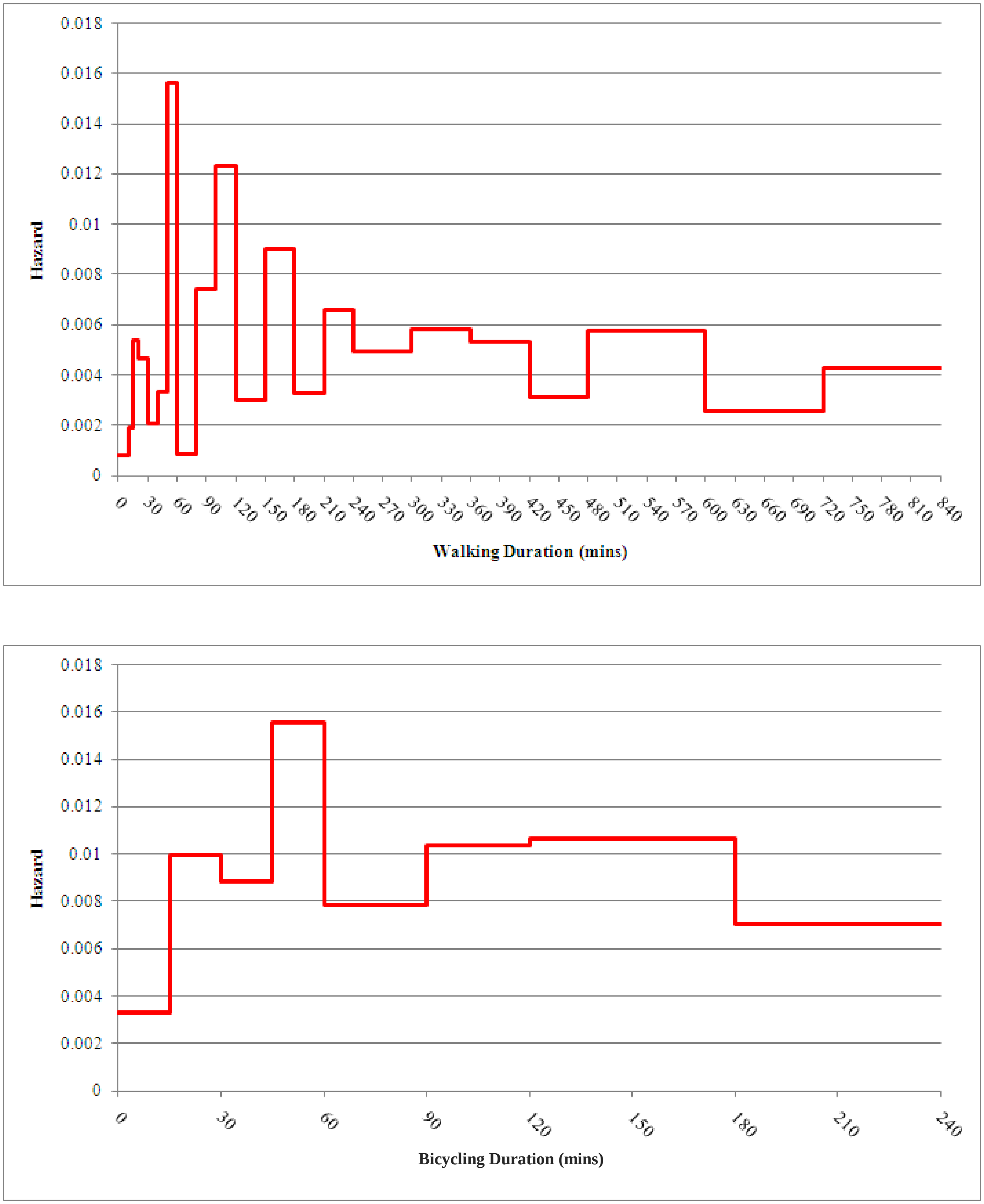


Figure 5.1 Continuous-Time Sample Hazard Functions

effects among the variables were also considered.[36] The final specification was based on intuitive considerations, insights from previous literature, and statistical fit/significance considerations. The final specification includes some variables that are not statistically significant at the usual 5% level of significance. We do not discard them because the effects of these variables are intuitive and have the potential to guide future research.

5.4.2 Model Estimation Results

Table 5.3 presents the model estimation results. The rows in the table correspond to the explanatory variables, while the columns correspond to the activity categories. Each activity category column has two sub-columns: the first sub-column provides the estimated coefficient corresponding to the row explanatory variable and the second sub-column provides the t-statistic for that coefficient. The base category is listed either next to that variable or in the heading of the row corresponding to that variable. The coefficients in the table indicate the effects of variables on the duration hazard for walking and cycling activity. A "-" cell entry indicates that the corresponding row exogenous variable does not have a statistically significant effect on the corresponding column activity hazard rate. A positive (negative) coefficient implies that the corresponding explanatory variable increases (decreases) the hazard rate and decreases (increases) the activity duration. In the following sections, we discuss the effects of variables on the activity duration hazards by variable category.

[36] None of the many built environment variables considered entered into the final model specification. This is because the attitudinal variables potentially capture the effects of the built environment.

Table 5.3 Model Estimation Results

	Walking activity		**Bicycling activity**	
	Estimates	t-stat	Estimates	t-stat
Effects of individual and household socio-demographic variables				
Age (base: age > 65 years)				
5 years ≤ Age ≤ 10 years	2.146	1.31	-	-
11 years ≤ Age ≤ 15 years	2.732	2.16	1.868	1.50
Other individual and household characteristics				
Male (base: female)	-	-	-1.361	-2.07
Full-time employed (base: not employed)	-	-	-1.224	-1.51
Non-motorized modes are used for work (base: motorized modes are used for work)	-	-	-1.931	-1.57
Presence of children aged 5 to 10 years in the HH (base: no children in the HH)	1.728	1.54	2.402	2.22
Effects of attitudinal variables				
Walking				
Inconvenience	10.079	3.47	-	-
(Lack of) Walking conditions due to motorized vehicles related factors	6.922	3.18	-	-
(Lack of) Safety	14.938	4.56	-	-
Bicycling				
Busy life style and absence of bicycle paths/trails	-	-	8.514	3.47
Inconvenience and lack of paved bicycle facilities	-	-	7.459	3.88
(Lack of) Safety	-	-	5.995	1.83
Heterogeneity parameters (standard deviation)				
Individual-specific heterogeneity				
Overall (θ)	2.934	1.92	2.934	1.92
Activity-specific (σ_m)	6.397	6.87	0.068	2.69
Social group-specific heterogeneity				
Overall (η)	0.469	3.26	0.469	3.26
Activity-specific (η_m)	-	-	0.138	2.82
Spatial cluster-specific heterogeneity				
Overall (δ)	1.496	3.60	1.496	3.60
Activity-specific (δ_m)	3.408	6.22	0.058	2.48

5.4.2.1 Individual and Household Socio-Demographic Variables

The results indicate that children who are 15 years of age or younger tend to spend less time walking compared to senior adults (*i.e.*, adults over the age of 65 years). Children in the 11 to 15 years age group are also less inclined to spend their time bicycling. This is consistent with the results of earlier studies which found that children tend to have a lower propensity to participate in physical activities (Sallis *et al.*, 2000, Sener *et al.*, 2009). This may also be attributed to parent(s) using car as the main mode of transportation to chauffeur children to/from school and organized leisure activities (Hjorthol and Fyhri, 2009). Compared to females, males tend to allocate more time to pursue bicycling activity. This result may be a reflection of distinct gender effect in terms of risk aversion, and reinforces the earlier findings that women are more likely to be concerned about bicycling in traffic and in the presence of aggressive motorist than men (Garrard *et al.*, 2006). Employment status also has an important effect on the bicycling activity duration. The results suggest that full-time employed adults are likely to allocate more time for bicycling compared to unemployed and part-time employed adults. In this context the results also suggest that when non-motorized modes are used for traveling to/from work, individuals tend to allocate more time for bicycling compared to when motorized modes are used for work. These two findings taken together imply that, among employed individuals, full-time workers using non-motorized modes for work are likely to allocate most time for bicycling (possibly bicycling for recreation or to “decompress” after work as well). After them, the next group of employed individuals who are likely to allocate most time bicycling is the part-time workers who use non-motorized modes to access work. The final group of employed bicycle users is the full-time workers who use motorized modes for all activities or non-motorized modes for all non-work activities.

The next variable captures the effect of the presence of 5-to-10-year-old children in the household. The positive sign of the co-efficient reflects lower tendency among individuals in households with young children to allocate time for walking and bicycling activities, presumably because children in that age group require higher child care/more attention, leaving other individuals with less time to pursue walking and bicycling activities.

5.4.2.2 Attitudinal Variables

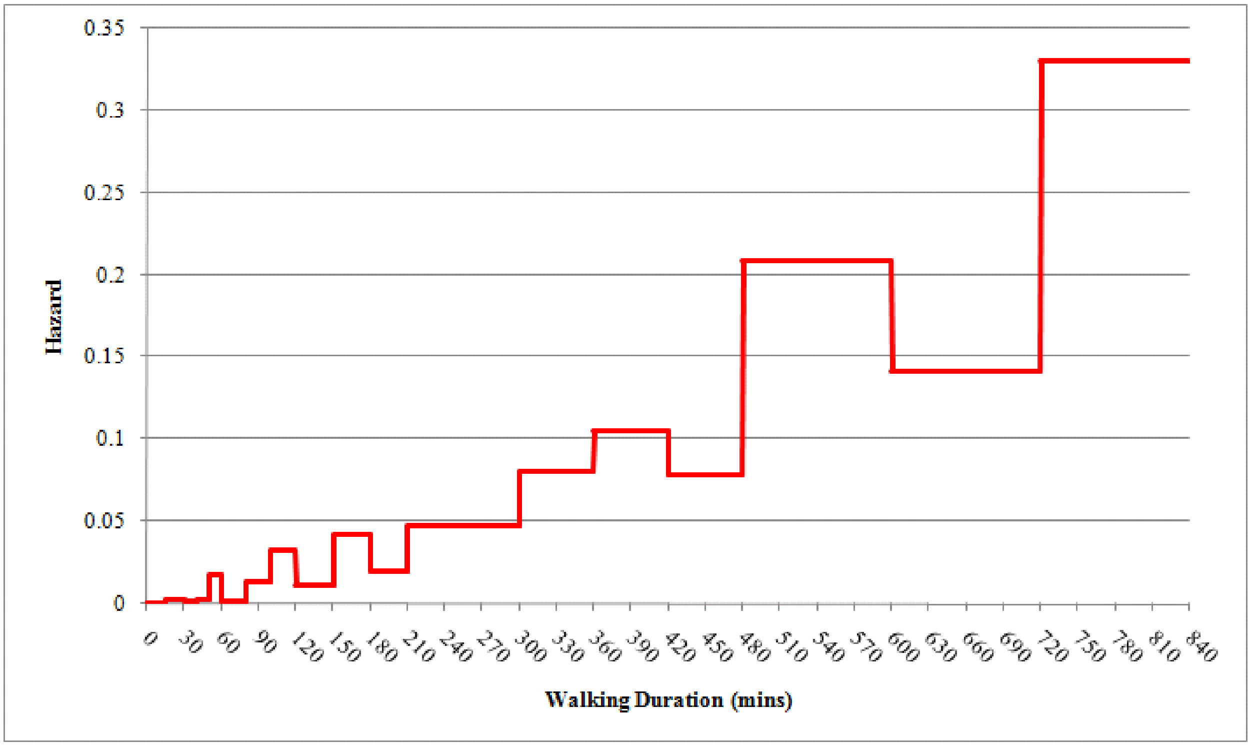
Lack of convenience and the perceived absence of walking conditions due to the motorized vehicles related factors deter individuals from walking as evidenced by the positive coefficients associated with these two attitudinal variables. Similarly, the perceived lack of safety deters people from spending time on walking activities. Likewise, several bicycling related factors deter time allocation to bicycling. Busy lifestyles and the unavailability of bicycle paths/trails, inconvenience in terms of carrying things and lack of paved bicycle facilities, and perceived (lack of) safety are all associated with positive coefficients. These results suggest that there are myriad factors that affect the time allocation to walking and bicycling activities. On the one hand, lack of convenience and busy lifestyles deter individuals from allocating time to walking and bicycling. These factors may not be easily for policymakers to manipulate, but it may be possible to ease lifestyle constraints by providing flexible work schedules and telecommuting options. However, more directly related to transportation planning and design are the findings that poor walking condition and unavailability/(perceived) poor quality of bicycling infrastructure are clearly having an adverse impact on the ability of individuals to spend more time walking and bicycling. It is conceivable that many short trips are taken by the automobile simply because the walking/bicycling infrastructure is perceived as unavailable, insufficient, poor, inadequate, or unsafe. Planners, designers, and policymakers may be able to enhance walking and bicycling use by addressing these issues.

5.4.2.3 Heterogeneity Parameters

The final rows of Table 5.3 present the estimated standard deviation of the heterogeneity parameters. The magnitude of the heterogeneity parameters and their statistical significance highlight the importance of considering common unobserved factors due to individual, social group, and spatial neighborhood effects. The following observations can be made from Table 5.3. First, the heterogeneity effects are, in general, statistically significant at all levels of clustering, except for the household-level clustering effect. This indicates the importance of explicitly incorporating the effects of unobserved factors when analyzing walking and bicycling activity durations. Second, the overall heterogeneity parameters that effect both walking and bicycling activity durations are statistically significant at the individual, social, and spatial level. Among them, the effect of individual-specific factor is the strongest, followed by the effects of spatial clustering and the social grouping. This finding reflects that instead of considering only a single aggregate level, the effect of clustering should be considered at multiple levels (which is the case modeled in the current study). Finally, the differential effects of activity-specific heterogeneity due to individual and spatial factors are more pronounced in walking activity duration compared to bicycling duration. In case of social grouping, it is found that the unobserved factors have significant impact only on bicycling activity duration. This may be attributed to the interactions with and the influence of individuals' peers (social network) group.

5.4.2.4 Baseline Hazard

The baseline hazard plots are shown in Figure 5.2. As were in the case of sample hazard rates, baseline hazards were also calculated under the assumption that the hazard remains constant within each discrete time interval. The baseline hazard functions are found to be non-monotonic and characterized by multiple peaks, similar to the sample hazard functions. This finding clearly indicates that non-parametric hazard functions are preferred over parametric specifications for analyzing walking and bicycling activity durations. Another interesting finding is that there are clear differences between the baseline hazards and the sample hazards. For instance, for walking activity duration, the baseline hazard increases with increase in activity duration, while the sample hazard decreases with increase in activity duration (except for the first 45 minutes). For



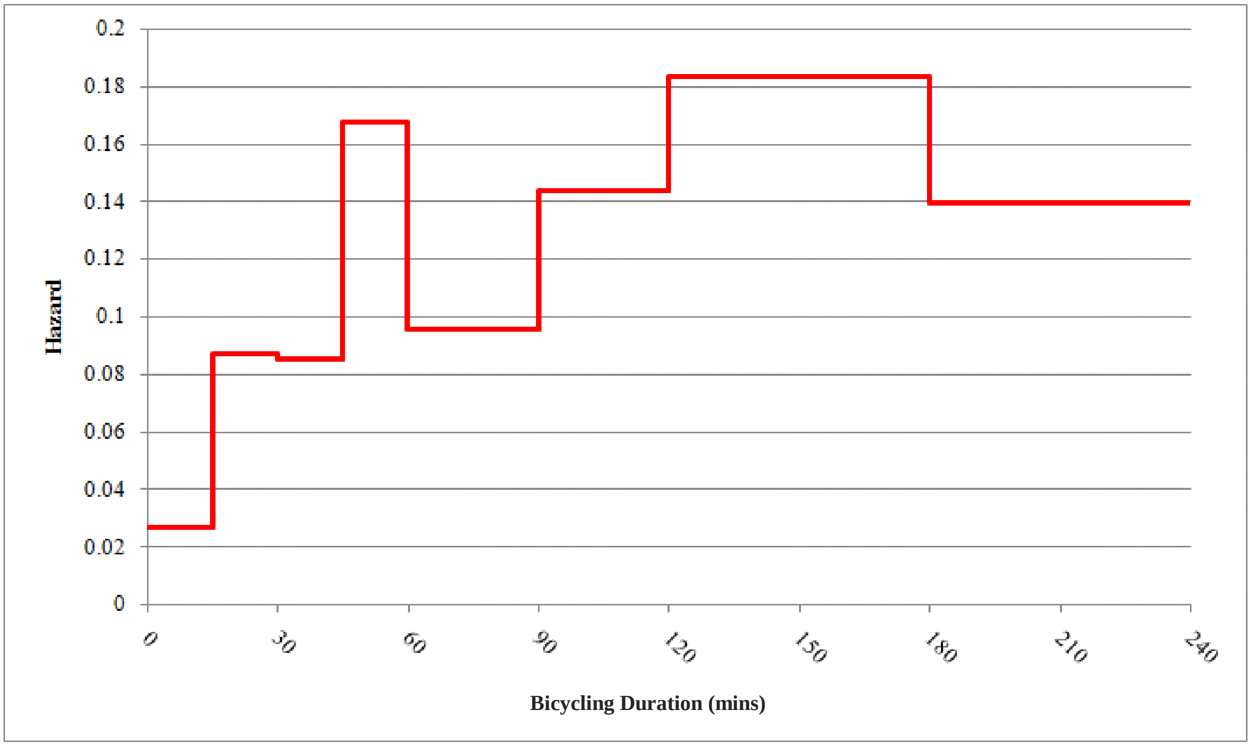


Figure 5.2 Baseline Hazard Functions

bicycling activity duration, the baseline hazard and the sample hazard were found to be more similar in profile; however, the baseline hazard shows more distinct peaks than the sample hazard. These differences between the baseline and sample hazards suggest that it is important to recognize variations in activity durations due to both observed and unobserved factors using approaches such as the one adopted in this study.

5.4.2.5 Threshold Parameters

The threshold parameters are not shown in the Table 5.3, but are available on request from the author. These parameters represent the cut-off points that map the latent propensity (log integrated hazard) of individuals to participate in each activity type to the reported activity duration. As such, they do not have any substantive behavioral interpretations.

5.4.2.6 Overall Measures of Fit

The log-composite likelihood value for the fully specified independent grouped response probit model (IGRP) (that is, independent grouped response probit models for each activity type) at convergence is –6,647,007.8 and that for the fully specified multi-level cross-cluster grouped response probit model (MCGRP) is –4,811,301.2. The composite likelihood ratio test (CLRT) statistic for comparing the MCGRP model with the IGRP model is 3,671,413.2. However, the CLRT statistic does not have the standard chi-squared asymptotic distribution under the null hypothesis, as in the case of the regular maximum likelihood inference procedure. While one can use bootstrapping to obtain the precise distribution of the CLRT statistic or adjust the value of the CLRT statistic using the procedure discussed in Section 2.6 (in Chapter 2), other measures can be used to determine whether the MCGRP model form is statistically superior to the IGRP model form. For instance, the t-statistics on θ, σ_m, η, η_m, δ, and δ_m parameter estimates are statistically significant, indicating that the MCGRP model is likely to be superior to the IGRP model which omits these statistically significant parameters. Further, one may compute an adjusted rho-bar squared value $\bar{\rho}_c^2$ in the composite marginal likelihood approach for the MCGRP model and the IGRP models as $\overline{\rho}_c^2 = 1 - [(\log L_{CML}(\hat{\boldsymbol{\kappa}}) - N) / \log L_{CML}(\mathbf{T})],$ where $\log L_{CML}(\hat{\boldsymbol{\kappa}})$ is the composite marginal log-likelihood at convergence, N is the number of model parameters excluding the thresholds, and $\log L_{CML}(\mathbf{T})$ is the log-likelihood with only thresholds in the model. The value of $\overline{\rho}_c^2$ for the IGRP model and the MCGRP model are 0.17 and 0.40 respectively, once again indicating that the IGRP model may be rejected in favor of the MCGRP model.

5.5 Summary and Conclusions

The widespread interest in sustainable development has transportation and land use professionals and policymakers exploring ways to increase the level of non-motorized mode use. Non-motorized mode use, such as walking and bicycling, not only offer considerable relief from congestion, energy savings, and greenhouse gas (GHG) emission reductions, but also offer health benefits to children and adults alike. Despite the high level of interest in non-motorized modes of transportation, there has been limited progress in the ability to adequately model their use. The aggregate representation of space and time in travel models, the inadequate detail of transportation networks (to include bicycle and pedestrian networks), and the paucity of non-motorized travel survey data have all contributed to this limited progress. More importantly, the profession needs a deeper understanding of the myriad factors and influences that affect non-motorized mode use to make progress on this front.

This study offers a framework and methodology for modeling the time spent walking and bicycling by individuals, while explicitly recognizing heterogeneity arising from individual-specific factors, family or intra-household interactions, social group or peer influences, and spatial clustering effects. In the United States, walking and bicycling activity is often a lifestyle preference that is linked closely to personal and household attitudes, beliefs, values, and perceptions. These attitudes and preferences (inclination or disinclination to the use of non-motorized modes) are likely to be shaped by not only one’s own individual-specific beliefs, but also influences of other household members, social peers, and neighborhood elements.

In this study, the time allocated to walking and bicycling activity over a period of one week is modeled jointly using a hazard model specification, thus providing the ability to examine how effects of various factors differentially impact walking vis-à-vis bicycling. The methodology adopted in this study is capable of accommodating grouped responses that typically are observed in activity-travel survey data sets wherein durations (start and end times) are rounded to the nearest fifth minute. The multilevel cross-cluster model structure is presented in detail in the chapter together with a model estimation approach that overcomes the challenge associated with evaluating a thousand-dimension integral of a multivariate density function. The composite marginal likelihood (CML) approach provides a tractable, easy to implement way to estimate parameters by transforming the large multidimensional integral to a low-dimensional integral.

The model is estimated on a survey sample data set derived from the California add-on of the United States National Household Travel Survey (NHTS) conducted in 2009. The subsample specific to three counties in the San Francisco Bay Area is extracted and analyzed in this study. The continuous time hazard functions suggest that individuals tend to be more uniform in the allocation of time to walking than to bicycling. Higher hazards for bicycling at small duration (up to 45 minutes) suggest that individuals tend to commit a certain minimum amount of time to walking, thus reducing the hazard in those initial periods. The model estimation results show standard individual and household demographic and socio-economic variables impact walking and bicycling activity duration. More importantly, however, there are numerous attitudinal factors and perceptions that affect walking and bicycling activity duration. In addition to busy lifestyles and such constraints, it is found that perceptions of poor walking condition, inadequate bicycling infrastructure, and concerns about safety adversely impact the amount of walking and bicycling undertaken by individuals. These findings are all consistent with expectations and point to the need for professionals and policymakers to consider neighborhood designs, land use configurations, and infrastructure investments that alleviate the concerns and enhance perceptions of walking and bicycling convenience.

Another important finding in this study is the significance of heterogeneity effects at multiple levels in the determination of non-motorized mode use. Travel demand model systems, with virtually no exception, ignore many of the (unobserved) interaction effects, social context, and spatial clustering effects that bring about heterogeneity in behavior. In this study, it is found that unobserved individual specific factors, social/peer group influence, and spatial clustering effects are all significant determinants for walking and bicycling activity duration. The finding that social/peer group influences are important suggests that public education campaigns targeted at specific age group may bring about changes in the non-motorized mode use of children and adults due to "peer" effects. Similarly, effects of spatial clustering should not be ignored in modeling non-motorized mode use as households tend to locate in spatial clusters (zones or neighborhoods) consistent with their lifestyle and travel preferences.

In summary, the results highlight the importance of considering walking and bicycling activity duration in a joint framework that accommodates not only observed variables but also explicitly includes the effects of heterogeneity at multiple levels such as individual, social, and spatial level. Integrated land use-transport model systems able to capture such effects through enhanced model specifications are likely to offer more accurate policy predictions that better inform decision-makers.

Chapter 6

A Spatial Panel Ordered-Response Model With An Application to the Analysis of Urban Land Use Development Intensity Patterns

6.1 Introduction

6.1.1 Background and Motivation

There is increasing interest and attention on recognizing and explicitly accommodating spatial dependence among decision-makers in models of continuous and discrete choices. While specification and modeling considerations related to spatial dependence appear to have originated initially in urban and regional modeling, such considerations have now permeated into economics and mainstream social sciences, including agricultural and natural resource economics, public economics, geography, sociology, political science, and epidemiology. Some recent examples in these fields include assessing harvest level of agricultural products (Ward *et al.*, 2010), determining the siting location for an industry (Alamá-Sabater *et al.*, 2011), and analyzing voter turnout in an election (Facchini and François, 2010). In addition to considering spatial dependence in purely cross-sectional data settings, the field also has expanded to accommodate spatial dependence in the context of panel data. Recent examples of spatial panel econometrics include examining changes in housing prices over time (Holly *et al.*, 2010), analyzing investment treaties between countries (Neumayer and Plümper, 2010), and studying the effects of a municipality's local tax rate structure on the tax rate structures of neighboring municipalities (Gérard *et al.*, 2010). The reader is also referred to a special issue of *Regional Science and Urban Economics*, edited by Arbia and Kelejian (2010), for a collection of recent papers on spatial dependence, and to Elhorst (2009) and Lee and Yu (2010) for good reviews of recent research on spatial panel data models. Anselin (2010) and Anselin *et al.* (2008) are additional resources for overviews of the developments in the spatial econometrics field.

At the same time that spatial considerations are receiving widespread attention, a specific kind of discrete choice structure – the ordered-response multinomial structure –

has also seen a literal explosion in application in many different disciplines, including sociology, biology, political science, marketing, and transportation sciences. Some recent examples of the use of ordered-response structures include examining crash severity (Quddus *et al.*, 2010), analyzing job satisfaction (Luechinger, *et al.*, 2010), assessing stream water quality (Higgs and Hoeting, 2010), studying trip generation (Roorda *et al.*, 2010), and examining monetary policies of a bank (Xiong, 2011). The reader is referred to Greene and Hensher (2010) for a comprehensive history and review of the ordered-response model structure (also, see Section 1.4 for more detail on the ordered-response model structure).

It should be clear from above that both spatial dependencies as well as ordered-response structures are becoming common place in the tool box of researchers in a wide variety of disciplines. However, there has been little research at the interface of spatial dependence and ordered-response structures. In particular, much of the literature on spatial dependency has been confined to the case of continuous dependent variables (and not discrete dependent variables), while much of the ordered-response literature has focused on the case of a (non-spatial) univariate ordered-response system. Of course, in the past decade, spatial dependence structures developed in the context of continuous dependent variables are increasingly being considered for binary discrete choice dependent variables (see Fleming, 2004, Bradlow *et al.*, 2005, Franzese and Hays, 2008, Franzese *et al.*, 2010, Robertson *et al.*, 2009, and LeSage and Pace, 2009; and Bhat and Sener, 2009 provide good reviews). The two dominant techniques, both based on simulation methods, for the estimation of such spatial binary discrete models are the frequentist recursive importance sampling (RIS) estimator (which is a generalization of the more familiar Geweke-Hajivassiliou-Keane or GHK simulator; see Beron *et al.*, 2003 and Beron and Vijverberg, 2004) and the Bayesian Markov Chain Monte Carlo (MCMC)-based estimator (see Kakamu and Wago 2007, LeSage and Pace, 2009). Such methods may be extended to ordered-response structures in a straightforward manner. However, both the RIS and MCMC-based methods are confronted with multi-dimensional normal integration (of the order of the number of observations in the estimation sample when using the general flexible spatial dependence forms adopted for continuous models), and are therefore computationally expensive-to-infeasible to implement (for both binary and ordered-response structures) with the typical computational resources at hand for anything other than small sample sizes (see Bhat, 2011, Smironov, 2010, and Franese *et al.*, 2010). Similar computational considerations have impeded the application of (non-spatial) multivariate ordered-response structures. Specifically, the estimation of models with an arbitrary number of correlated ordered-response outcomes entails, in the usual likelihood function approach, integration of dimensionality equal to the number of outcomes. Again, the norm in such a case has been to apply numerical simulation techniques based on a maximum simulated likelihood (MSL) approach (see Bhat and Srinivasan, 2005 and Balia and Jones, 2008) or a Bayesian inference approach (see Herriges *et al.* , 2008, Jeliazkov *et al.*, 2008, and Hasegawa, 2010), as discussed in Section 1.4. However, these methods become impractical as the number of ordered-response outcomes increases.

In contrast to the extant simulation-based inference procedures discussed above, the CML provides an appealing alternative inference approach. Recent studies that use this approach for non-spatial multivariate binary/ordered-response modeling include Yi *et al.*, 2011, Varin and Czado 2010, Ferdous *et al.*, 2010, and Bhat *et al.*, 2010b. However, there has only been one study so far (by Bhat *et al.*, 2010b) that has employed the CML method in the context of spatial multivariate binary or ordered-response systems (note that spatial dependence immediately leads to a multivariate ordered-response model system because of the dependence generated across the ordered-responses of multiple decision-agents). However, the spatial dependency formulation in Bhat *et al.* (2010b) is based on a spatial error formulation that assumes that the dependency is a "nuisance" issue; it does not consider the structural "spillover" effects caused by exogenous variables that we believe would be an important consideration in land use analysis (as we discuss further later on). Bhat *et al.* (2010b)'s study also employs a cross-sectional model, with no temporal panel element. Further, spatial heterogeneity is not considered and the error correlation is not generated through a flexible autoregressive structure.

6.1.2 The Current Study

The current study develops a formulation for a spatial panel ordered-response model and proposes a composite marginal likelihood (CML) inference approach to obtain model parameter estimates. Spatial dependence is introduced through contemporaneous "spillover" effects in both the exogenous variables as well as the error terms. Such a specification recognizes that spatial dependence is a substantive issue, and is caused by didactic interactions among decision-making agents (as opposed to considering spatial dependence only in the error terms, which is tantamount to viewing spatial dependence as "nuisance" dependence). In the empirical context of the current study, which is on examining the land development intensity levels of spatial units, the implication is that the spatial dependence in the development intensities of proximately located spatial units is a result of interactions between land owners of the corresponding spatial units. Such interactions should naturally arise because land owners of proximately located spatial units (say, parcels), acting as profit-maximizing economic agents, are likely to be influenced by each other's perceptions of net stream of returns from land use development. The peer influences may also be due to strategic or collaborative partnerships between land owners. The net result is that changes in observed variables (such as accessibility to the city-center) and/or unobserved variables (such as neighborhood politics and zoning guidelines) that affect the land use development returns (LUDR) perception of one land owner will also likely lead to a shift in the LUDR perception of land owners of neighboring parcels. We use a spatial lag structure to accommodate these peer interactions, as also suggested by Carrion-Flores *et al.* (2009). Besides, as indicated by Anselin (2003), it behooves the analyst to include spatial "spillover" effects in both the explanatory variables as well as the errors when there are no strong a priori theoretical reasons to restrict global externalities to only the errors or only the explanatory variables.

In addition to spatial dependence, we incorporate (unobserved) spatial heterogeneity by allowing the sensitivity to exogenous variables to vary across land owners. For instance, different land owners may have different intrinsic LUDR perceptions and may also respond differently to the exogenous variables, based on such unobserved factors as individual experiences, risk-taking behavior, and even vegetation conservation values. This would then translate to a land owner-specific random coefficients formulation for the LUDR perceptions, leading to a stationary across-time correlation in land development intensity for the same spatial unit. Such land owner-specific random coefficients and resulting temporal correlations of the land owner's choices across time have been ignored thus far in the literature. In fact, all earlier discrete model spatial dependence studies we are aware of consider a generic time-stationary random effect (that is, a random coefficient only on the intercept) for each spatial unit in their spatial error formulations, but such a formulation is restrictive relative to the more general random-coefficients spatial lag formulation used here. Further, due to computational difficulties with the traditional MSL and Bayesian methods, several earlier studies group spatial units into much fewer regions and consider random effects only at this regional level (and also accommodate spatial dependency effects through a spatial error structure only at this aggregate region level; see, for example, Phaneuf and Palmquist, 2003, Smith and LeSage, 2004, and LeSage and Pace, 2009, Chapter 10). On the other hand, our inference approach allows us to retain spatial dependence effects at the basic disaggregate level of the landowners of the individual parcels, while also allowing spatial heterogeneity (through the random coefficients specification) at this disaggregate level. Such an underlying framework goes beyond data fitting models for the land use development intensity of parcel-level units to more closely linking land use patterns to the decision agents (*i.e.*, the land owners) behind the land use patterns.[37] Finally, we also accommodate time-varying dependency effects across the LUDR perceptions of the same decision agent at different points in time. These time-varying effects may be attributed to the effects of recent experiences and events that may

[37] Of course, one challenge to this notion would be that, over long time periods, parcels may change hands, leading to different land owners at different times. However, we would argue that it is still far more appealing to maintain the linkage between land parcels and land owners (even if not perfect) rather than completely severing this linkage in the modeling process.

influence the risk-taking or risk-averseness or other LUDR-related perceptions of individual land owners. As such, these effects fade over time, with the LUDR perceptions at a particular time being much more affected by perceptions in the recent past than those from sometime back.

The study assesses the ability of the CML inference procedure to recover the underlying parameters of the proposed spatial panel ordered-response structure using simulated experiments. Subsequently, we demonstrate the applicability of the proposed formulation and inference procedure by modeling urban land use development intensity patterns in Austin, Texas, using data from the years 2000, 2003, 2006, and 2008. The land use information used in the current empirical analysis is available at a parcel-level spatial resolution. While various different levels and thresholds may be employed to define the intensity level of land development, we adopt a four category ordinal system: (1) undeveloped land (open space, vacant parcel, *etc.*), (2) less-intensely developed land (residential parcels with single-family detached or two-family attached home), (3) medium-intensely developed land (includes all other types of residential parcels), and (4) most-intensely developed land (includes office, commercial, industrial parcels, *etc.*). The data set comprises 783 parcels from each of the four years.

The rest of the chapter is structured as follows. Section 6.2 discusses the model structure and the estimation approach, Section 6.3 presents a simulation study to evaluate the ability of our proposed approach to recover model parameters and also demonstrates the effects of ignoring spatial dependency and spatial heterogeneity when they are actually present. Section 6.4 describes the data sources and sample formation procedure for the Austin data. Section 6.5 presents the empirical results. The final section summarizes the important findings from the study and concludes the chapter.

6.2 The Model

6.2.1 Basic Formulation

Let q be an index for spatial units (q = 1, 2, …, Q, where Q denotes the total number of spatial units/parcels in the data set), and let t be an index for time period (t = 1, 2, …, T, where T is the number of panel observations for each spatial unit; in the current study, $T = 4$).[38] Let l be an index for the observed land use development category, which may take one of L discrete ordinal values (*i.e.*, $l \in \{1, 2, \ldots, L\}$). Assume that the land use development returns (LUDR) perception of the land owner of the q^{th} parcel at time t is y_{qt}^{*} (in the rest of this section, we will use the term "parcel" to refer to the spatial unit of analysis, though any other spatial unit may be used depending on the nature of the analysis). The LUDR perception is not observed by the analyst. But, in the usual ordered-response framework, we write this latent perception (y_{qt}^{*}) as a function of relevant covariates, and relate this latent propensity to the observed land use l through threshold bounds as follows (see McKelvey and Zavoina, 1975):

$$y_{qt}^{*} = \delta \sum_{q'=1}^{Q} w_{qq'} y_{q't}^{*} + \boldsymbol{\beta}_{\mathbf{q}}^{'} \mathbf{x}_{\mathbf{qt}} + \varepsilon_{qt}, y_{qt} = l \text{ if } \psi_{l-1} < y_{qt}^{*} < \psi_{l}, \ \boldsymbol{\beta}_{\mathbf{q}} = \mathbf{b} + \tilde{\boldsymbol{\beta}}_{\mathbf{q}}, \qquad (6.1)$$

The basic idea of the ordered-response formulation is that land owners with a low LUDR perception will keep their land undeveloped, while land owners with a high LUDR perception will invest their land in intense land use development. In the above equation, the first term reflects the spatial lag structure, where $w_{qq'}$ is the spatial proximity-based weight corresponding to units q and q' (with $w_{qq} = 0$ and $\sum_{q'} w_{qq'} = 1$) for each (and all) q, and δ $(0 < \delta < 1)$ is the spatial autoregressive parameter. $\mathbf{x}_{\mathbf{qt}}$ is a (K×1) vector of exogenous variables corresponding to parcel q and time period t ($\mathbf{x}_{\mathbf{qt}}$ includes a constant), $\boldsymbol{\beta}_{\mathbf{q}}$ is a corresponding (K×1) vector of random coefficients that is K-dimensional multivariate normal (MVN_K). For later use, we will partition $\boldsymbol{\beta}_{\mathbf{q}}$ into a (K×1) mean vector $\mathbf{b}$ and a (K×1) random component $\tilde{\boldsymbol{\beta}}_{\mathbf{q}}$ with mean zero and variance

[38] In the empirical context of the current study, the number of panel observations is the same across spatial units, *i.e.*, the data set is a balanced panel. However, the methodology in this study is generic and equally applicable to unbalanced panels.

$\mathbf{\Omega} = \mathbf{LL}'$ (*i.e.*, $\tilde{\boldsymbol{\beta}}_{\mathbf{q}} \sim \mathrm{MVN_K}[\mathbf{0},\mathbf{\Omega}]$) . It is not necessary that all elements of $\boldsymbol{\beta}_{\mathbf{q}}$ be random; that is, the analyst may specify fixed coefficients on some exogenous variables in the model, though it will be convenient in presentation to assume that all elements of $\boldsymbol{\beta}_{\mathbf{q}}$ are random. Also, note that the element of **b** corresponding to the constant is fixed to zero for identification. The upper bound threshold for ordinal level *l* is represented by ψ_l ($\psi_0 < \psi_1 < \psi_2 ... < \psi_{L-1} < \psi_L$; $\psi_0 = -\infty$ and $\psi_L = +\infty$). The term ε_{qt} in the above equation is a standard normal error term uncorrelated across parcels for a particular time period *t*. However, we allow a first-order autoregressive correlation pattern within each spatial unit-specific series of observations so that $Cov(\varepsilon_{qt}, \varepsilon_{qt'}) = \rho^{|t'-t|}$ $(0 < \rho < 1)$.

The formulation above generates spatial dependence through the spatial lag term, the nature of which is related to the specification of the weight terms $w_{qq'}$. This can take the form of a discrete function such as a contiguity specification ($w_{qq'}$ =1 if the parcels *q* and q' are adjacent and 0 otherwise) or a specification based on a distance threshold ($w_{qq'} = c_{qq'} / \sum_{q'} c_{qq'}$, where $c_{qq'}$ is a dummy variable taking the value 1 if the parcel q' is within the distance threshold and 0 otherwise). It can also take a continuous form such as those based on the inverse of distance $d_{qq'}$ and its power functions $\left(w_{qq'} = (1/d^{\,n}_{qq'}) \left[\sum_{q'} 1/d^{\,n}_{qq'} \right]^{-1} \right)$ $(n > 0)$, the inverse of exponential distance, and the shared border length $\tilde{d}_{qq'}$ between parcels $w_{qq'} = \tilde{c}_{qq'} \tilde{d}_{qq'} / \left(\sum_{q'} \tilde{c}_{qq'} \tilde{d}_{qq'}, \right)$ (where $\tilde{c}_{qq'}$ is a dummy variable taking the value 1 if the parcels *q* and q' are adjoining, and 0 otherwise). All of these functional forms for the weight matrix may be tested empirically. In addition to spatial dependence, the random coefficient vector $\boldsymbol{\beta}_{\mathbf{q}}$ accommodates spatial heterogeneity as well as implicitly generates spatial heteroscedasticity. Note that we are able to disentangle spatial dependence and spatial heterogeneity because of the availability of panel data. Further, the vector $\boldsymbol{\beta}_{\mathbf{q}}$ generates time-invariant temporal dependence effects in the LUDR perceptions of the same land owner.

Several restrictive models are obtained from the model developed here. If $\rho = 0,$ this indicates lack of time-varying temporal correlation. If $\delta = 0,$ the result is a non-spatial model. If the elements of $\boldsymbol{\Omega}$ are zero, the indication is the lack of time-invariant temporal effects as well as unobserved spatial heterogeneity. If the elements of $\boldsymbol{\Omega}$ corresponding to the non-diagonal elements of $\boldsymbol{\Omega}$ are zero, but not the diagonal elements, it represents the case of the presence of time-invariant and unobserved heterogeneity effects, but without correlation between these effects. If the elements of $\boldsymbol{\Omega}$ except for that corresponding to the constant are collectively zero, the model collapses to a random-effects structure. If $\rho = 0,\ \delta = 0,$ and all elements of $\boldsymbol{\Omega}$ are identically zero, the result is a standard ordered-response model.

6.2.2 Matrix Formulation

The model proposed above may be written in a more compact form to facilitate the discussion of the estimation technique. To do so, we define the following vectors and matrices:

$\mathbf{y}_{\mathbf{t}}^{*} = (y_{1t}^{*}, y_{2t}^{*}, y_{3t}^{*}, ..., y_{Qt}^{*})'$ $(Q \times 1$ matrix),

$\mathbf{y}^{*} = [(\mathbf{y}_{\mathbf{1}}^{*})', (\mathbf{y}_{\mathbf{2}}^{*})', (\mathbf{y}_{\mathbf{3}}^{*})', ..., (\mathbf{y}_{\mathbf{T}}^{*})']'$ $(QT \times 1$ matrix),

$\boldsymbol{\varepsilon}_{\mathbf{t}} = (\varepsilon_{1t}, \varepsilon_{2t}, \varepsilon_{3t}, ..., \varepsilon_{Qt})'$ $(Q \times 1$ matrix),

$\boldsymbol{\varepsilon} = (\boldsymbol{\varepsilon}_{\mathbf{1}}', \boldsymbol{\varepsilon}_{\mathbf{2}}', \boldsymbol{\varepsilon}_{\mathbf{3}}', ..., \boldsymbol{\varepsilon}_{\mathbf{T}}')'$ $(QT \times 1$ matrix),

$\mathbf{x}_{\mathbf{qt}} = (x_{qt1}, x_{qt2}, x_{qt3}, ..., x_{qtK})'$ $(K \times 1$ matrix),

$\mathbf{x}_{\mathbf{t}} = (\mathbf{x}_{\mathbf{1t}}, \mathbf{x}_{\mathbf{2t}}, \mathbf{x}_{\mathbf{3t}}, ..., \mathbf{x}_{\mathbf{Qt}})'$ $(Q \times K$ matrix),

$\mathbf{x} = (\mathbf{x}_{\mathbf{1}}', \mathbf{x}_{\mathbf{2}}', \mathbf{x}_{\mathbf{3}}', ..., \mathbf{x}_{\mathbf{T}}')'$ $(QT \times K$ matrix),

$$\widetilde{\mathbf{x}}_{\mathbf{t}} = \begin{bmatrix} \mathbf{x}'_{\mathbf{1t}} & 0 & 0 & \cdots & 0 \\ 0 & \mathbf{x}'_{\mathbf{2t}} & 0 & \cdots & 0 \\ 0 & 0 & \mathbf{x}'_{\mathbf{3t}} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \mathbf{x}'_{\mathbf{Qt}} \end{bmatrix} \text{ ($Q \times KQ$ block diagonal matrix),}$$

$$\widetilde{\mathbf{x}} = (\widetilde{\mathbf{x}}'_{\mathbf{1}}, \widetilde{\mathbf{x}}'_{\mathbf{2}}, \widetilde{\mathbf{x}}'_{\mathbf{3}}, ..., \widetilde{\mathbf{x}}'_{\mathbf{T}})' \quad \text{($QT \times KQ$ matrix),}$$

$$\widetilde{\boldsymbol{\beta}} = (\widetilde{\boldsymbol{\beta}}'_{\mathbf{1}}, \widetilde{\boldsymbol{\beta}}'_{\mathbf{2}}, \widetilde{\boldsymbol{\beta}}'_{\mathbf{3}}, ..., \widetilde{\boldsymbol{\beta}}'_{\mathbf{Q}})' \quad \text{($KQ \times 1$ matrix).}$$

Also, collect all the weights $w_{qq'}$ into a spatial weight matrix **W**. The vector $\widetilde{\boldsymbol{\beta}}$ above has a mean vector of zero and a variance matrix $\mathbf{I_Q} \otimes \boldsymbol{\Omega}$ (of size *QT*×*QT*), where $\mathbf{I_Q}$ is an identity matrix of size *Q*. Note also that the error vector $\boldsymbol{\varepsilon}_{\mathbf{t}}$ is distributed multivariate normal with a mean vector of zero and a temporal autoregressive covariance matrix $\boldsymbol{\Lambda}$ (of size *T*×*T*) given below:

$$\boldsymbol{\Lambda} = \begin{bmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{T-1} \\ \rho & 1 & \rho & \cdots & \rho^{T-2} \\ \rho^2 & \rho & 1 & \cdots & \rho^{T-3} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ \rho^{T-1} & \rho^{T-2} & \rho^{T-3} & \cdots & 1 \end{bmatrix} \tag{6.2}$$

Then, the error vector $\boldsymbol{\varepsilon}$ is distributed multivariate normal with a mean vector of zero and a covariance matrix $\boldsymbol{\Lambda} \otimes \mathbf{I_Q}$ (of size *QT*×*QT*).

Using the vector and the matrix notations defined above, Equation (6.1) may be re-written compactly as:

$$\mathbf{y}^* = \delta(\mathbf{I_T} \otimes \mathbf{W})\mathbf{y}^* + \mathbf{xb} + \widetilde{\mathbf{x}}\widetilde{\boldsymbol{\beta}} + \boldsymbol{\varepsilon},$$

where $\mathbf{I_T}$ is an identity matrix of size *T*. After further matrix manipulation to write $\mathbf{y}^*$ in reduced form, we obtain:

$$\mathbf{y}^* = \mathbf{S}\mathbf{x}\mathbf{b} + \mathbf{S}\widetilde{\mathbf{x}}\widetilde{\boldsymbol{\beta}} + \mathbf{S}\boldsymbol{\varepsilon},\ \mathbf{S} = \left[\mathbf{I_{QT}} - \delta(\mathbf{I_T} \otimes \mathbf{W})\right]^{-1} = \mathbf{I_T} \otimes \left[(\mathbf{I_Q} - \delta\mathbf{W})^{-1}\right] \tag{6.3}$$

The expected value and the variance of $\mathbf{y}^*$ are then as follows:

$E(\mathbf{y}^*) = \mathbf{S}\mathbf{x}\mathbf{b} = \mathbf{B}$, and

$$Var(\mathbf{y}^*) = \mathbf{S}\tilde{\mathbf{x}}(\mathbf{I_Q} \otimes \mathbf{\Omega})\tilde{\mathbf{x}}'\mathbf{S}' + \mathbf{S}(\mathbf{\Lambda} \otimes \mathbf{I_Q})\mathbf{S}' = \mathbf{\Sigma} \qquad (6.4)$$

An important point from the reduced form in Equation (6.3) is that our contemporaneous spatial lag formulation specifies a spatial externality effect due to the time-invariant random coefficients too (see the $\mathbf{S}\tilde{\mathbf{x}}\tilde{\boldsymbol{\beta}}$ component on the right side of Equation (6.3)). That is, spatial dependence is implicitly generated in the observation-unit specific (time-invariant) coefficients. For instance, the preference and responsiveness to signals relevant to decision-making (such as how land owners respond to market place proximity or to proximity to lakes and other recreation centers) may themselves be correlated based on proximity of landowners' parcels. This is in addition to the usual "spillover" effects (or spatial externality effects) originating from the exogenous variables ($\mathbf{x}$) and the error terms ($\boldsymbol{\varepsilon}$).[39]

6.2.3 Estimation Approach

The parameter vector to be estimated is $\boldsymbol{\theta} = (\mathbf{b}', \psi_1, \psi_2, \psi_3, \cdots, \psi_{L-1}, \boldsymbol{\omega}, \delta, \rho)'$, where $\boldsymbol{\omega}$ is a column vector obtained by vertically stacking the lower triangle elements of the matrix $\mathbf{L}$ (recall that $\mathbf{\Omega} = \mathbf{LL}'$). Let the actual observed land development intensity level of spatial unit *q* at time period *t* be m_{qt} ($m_{qt} \in \{1, 2, \ldots, L\}$). Then, the likelihood function for the model is:

[39] Note that the spatially structured effects probit model used in earlier studies that accommodates random effects at an aggregate regional level (see Smith and LeSage, 2004, and LeSage and Pace, 2009) is a restrictive spatial dependency specification compared to the one adopted here. In particular, if the only random coefficient was on the constant term, and this randomness was at an aggregate region level rather than a disaggregate parcel level, and if there are no additional spatial externality effects due to exogenous variables and the error term $\boldsymbol{\varepsilon}$, then the spatial dependency in the reduced form of Equation (6.4) is similar to that in the spatially structured effects probit model.

$$L(\boldsymbol{\theta}) = \Pr(\mathbf{y} = \mathbf{m}) = \int\limits_{D_{y^*}} \phi_{QT}(\mathbf{y}^* \mid \mathbf{b}, \boldsymbol{\Sigma}) d\mathbf{y}^*, \tag{6.5}$$

where $D_{y^*} = \{\boldsymbol{y}^* : \psi_{(q,m_{qt}-1,t)} < y^*_{qt} < \psi_{q,m_{qt},t}, \forall\, q = 1,2,...,Q,\ t = 1,2,...,T\}$ and $\phi_{QT}(.)$ is the multivariate normal density function of dimension *QT*. **m** is a *QT*×1-vector of observed ordinal outcomes as follows: $\mathbf{m} = (m_{11}, m_{21}, m_{31}, ..., m_{Q1}, m_{12}, m_{22}, m_{32}, ..., m_{Q2}, ..., m_{1T}, m_{2T}, m_{3T}, ..., m_{QT})'$. The integration domain D_{y^*} is simply the multivariate region of the elements of the $\mathbf{y}^*$ vector determined by the observed vector of ordinal outcomes.

The dimensionality of the rectangular integral in the likelihood function is *QT*. As discussed earlier, the use of numerical simulation techniques based on a maximum simulated likelihood (MSL) approach or a Bayesian inference approach, even if feasible, can lead to convergence problems during estimation (Bhat *et al.*, 2010a; Müller and Czado, 2005). The alternative is to use the composite marginal likelihood (CML) approach, as discussed in Section 6.1.1. In the current study we use the pairwise composite marginal likelihood method based on the product of the likelihood contributions from pairs of observation units across time periods. To write this function, define two threshold vectors of size *QT*×1 as follows:

$$\boldsymbol{\tau} = (\psi_{1,m_{11}-1,1}, \psi_{2,m_{21}-1,1}, ..., \psi_{Q,m_{Q1}-1,1}, \psi_{1,m_{12}-1,2}, \psi_{2,m_{22}-1,2}, ..., \psi_{Q,m_{Q2}-1,2}, ... \psi_{1,m_{1T}-1,T}, \psi_{2,m_{2T}-1,T}, ..., \psi_{Q,m_{QT}-1,T})',$$
$$\vartheta = (\psi_{1,m_{11},1}, \psi_{2,m_{21},1}, ..., \psi_{Q,m_{Q1},1}, \psi_{1,m_{12},2}, \psi_{2,m_{22},2}, ..., \psi_{Q,m_{Q2},2}, ... \psi_{1,m_{1T},T}, \psi_{2,m_{2T},T}, ..., \psi_{Q,m_{QT},T})'.$$

Let *g* be an index that can takes the values from 1 to *QT*. Then,

$$\begin{aligned} L_{CML}(\boldsymbol{\theta}) &= \left(\prod_{g=1}^{QT-1} \prod_{g'=g+1}^{QT} \Pr\left([\mathbf{y}]_g = [\mathbf{m}]_g, [\mathbf{y}]_{g'} = [\mathbf{m}]_{g'}\right) \right) \\ &= \left(\prod_{g=1}^{QT-1} \prod_{g'=g+1}^{QT} \begin{bmatrix} \Phi_2(\varphi_g, \varphi_{g'}, \nu_{gg'}) - \Phi_2(\varphi_g, \mu_{g'}, \nu_{gg'}) \\ -\Phi_2(\mu_g, \varphi_{g'}, \nu_{gg'}) + \Phi_2(\mu_g, \mu_{g'}, \nu_{gg'}) \end{bmatrix} \right), \end{aligned} \tag{6.6}$$

where $\varphi_{\mathbf{g}} = \dfrac{[\vartheta]_g - [\mathbf{B}]_g}{\sqrt{[\boldsymbol{\Sigma}]_{gg}}}$, $\mu_g = \dfrac{[\boldsymbol{\tau}]_g - [\mathbf{B}]_g}{\sqrt{[\boldsymbol{\Sigma}]_{gg}}}$, $\nu_{gg'} = \dfrac{[\boldsymbol{\Sigma}]_{gg'}}{\sqrt{[\boldsymbol{\Sigma}]_{gg}}\sqrt{[\boldsymbol{\Sigma}]_{g'g'}}}$.

In the above expression, $[\vartheta]_g$ represents the g^{th} element of the column vector ϑ, and similarly for other vectors. $[\boldsymbol{\Sigma}]_{gg}$ represents the gg^{th} element of the matrix $\boldsymbol{\Sigma}$. The CML estimator is obtained by maximizing the logarithm of the function in Equation (6.6).

The pairwise marginal likelihood function of Equation (6.6) comprises $QT(QT-1)/2$ pairs of bivariate probability computations, which can itself become quite time consuming. Fortunately, in a spatial-temporal case where spatial dependency drops quickly with inter-observation distance, the pairs formed from the closest spatial observation units provide much more information than pairs from spatial units that are far away. In fact, as demonstrated by Varin and Vidoni (2009), Bhat *et al.* (2010a), and Varin and Czado (2010) in different empirical contexts, retaining all pairs not only increases computational costs, but may also reduce estimator efficiency. We examine this issue by creating different distance bands and, for each specified distance band, we consider only those pairings in the CML function that are within the spatial distance band. Then, we develop the asymptotic variance matrix $\mathbf{V}_{CML}(\hat{\boldsymbol{\theta}})$ for each distance band and select the threshold distance value that minimizes the total variance across all parameters as given by $tr[\mathbf{V}_{CML}(\hat{\boldsymbol{\theta}})]$ (*i.e.*, the trace of the matrix $[\mathbf{V}_{CML}(\hat{\boldsymbol{\theta}})]$).

The asymptotic covariance matrix $\mathbf{V}_{CML}(\hat{\boldsymbol{\theta}})$ may be computed from the Godambe sandwich information matrix ($\mathbf{G}(\boldsymbol{\theta})$) as follows:

$$\mathbf{V}_{CML}(\hat{\boldsymbol{\theta}}) = [\mathbf{G}(\boldsymbol{\theta})]^{-1} = [\mathbf{H}(\boldsymbol{\theta})]^{-1}\mathbf{J}(\boldsymbol{\theta})[\mathbf{H}(\boldsymbol{\theta})]^{-1}, \qquad (6.7)$$

where $\mathbf{H}(\boldsymbol{\theta}) = E\left[-\frac{\partial^2 \log L_{CML}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right]$ and $\mathbf{J}(\boldsymbol{\theta}) = E\left[\left(\frac{\partial \log L_{CML}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right)\left(\frac{\partial \log L_{CML}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}'}\right)\right]$.

The matrix $\mathbf{H}(\boldsymbol{\theta})$ of Equation (6.7) can be estimated in a straightforward manner using the Hessian of the negative of $\log L_{CML}(\boldsymbol{\theta})$, evaluated at the CML estimate $\hat{\boldsymbol{\theta}}$ (as discussed in chapters 4 and 5). However, the estimation of the $\mathbf{J}(\boldsymbol{\theta})$ matrix is not straightforward because of the underlying spatial and temporal dependence. But, because the spatial dependence pattern implied by the spatial lag structure fades with distance, one can use the windows sampling method of Heagerty and Lumley (2000) to estimate $\mathbf{J}(\boldsymbol{\theta})$. Here we use the windows sampling method proposed by Bhat (2011). Bhat's approach is as follows:

- Overlay the spatial region under consideration with a square grid providing a total of $\tilde{D}$ internal and external nodes. Then, select the observational unit closest to each of the $\tilde{D}$ grid nodes to obtain D observational units from the original Q observational units (d = 1, 2, 3, …, D; $\tilde{D} \geq D$).
- Let $\tilde{\mathbf{C}}$ be a $Q \times D$ matrix with its d^{th} column filled with a Q×1 vector of zeros and ones, with a zero value in the q'^{th} row (q'=1,2,…Q) if the observational unit q' is not within the specified threshold distance of unit d, and a one otherwise. Also, let $\mathbf{C} = \mathbf{1}_T \otimes \tilde{\mathbf{C}}$, where $\mathbf{1}_T$ is a T×1-matrix of ones. Then, the columns of $\mathbf{C}$ provide pseudo-independent sets of observational units.[40]
- Let the score matrix corresponding to the pairings in column d of matrix $\mathbf{C}$ be $\mathbf{s}_{CML,d}(\boldsymbol{\theta})$. Also, let N_d be the sum of the d^{th} column of $\mathbf{C}$, and let $\tilde{W} = \sum_{g=1}^{QT-1} \sum_{g'=g+1}^{QT} [\mathbf{R}]_{gg'}$, where $\mathbf{R} = \mathbf{1}_{T \times T} \otimes \tilde{\mathbf{R}}$. $\tilde{\mathbf{R}}$ is a Q×Q matrix with its q^{th} column filled with a Q×1 vector of zeros and ones, with a zero value in the q'^{th} row (q' =1,2,…Q) if the observational unit q' is not within the specified threshold distance of unit q, and a one otherwise (by construction, $\tilde{R}_{q'q} = 1$ if $q' = q$).
- Then, the $\mathbf{J}(\boldsymbol{\theta})$ matrix may be empirically estimated as:

[40] As indicated by Bhat (2011), there needs to be a balance here between the number of sets of pairings *D* and the proximity of points. The smaller the value of *D*, the less proximal are the sets of observation units and more likely that the sets of observational pairings will be independent. However, at the same time, the value of *D* needs to be reasonable to obtain a good empirical estimate of **J**, since this empirical estimate is based on averaging the cross-product of the score functions (computed at the convergent parameter values) across the *D* sets of observations.

$$\mathbf{J}\left(\hat{\boldsymbol{\theta}}\right)=\frac{\tilde{W}}{D}\left[\sum_{d=1}^{D}\left[\frac{1}{N_d}\left([\mathbf{s}_{CML,d}(\boldsymbol{\theta})][\mathbf{s}_{CML,d}(\boldsymbol{\theta})]'\right)_{\hat{\boldsymbol{\theta}}}\right]\right]. \tag{6.8}$$

A final issue regarding estimation. The positive definiteness of $\boldsymbol{\Sigma}$ is ensured as long as $0<\delta<1, 0<\rho<1$ and the matrix $\boldsymbol{\Omega}$ is positive-definite. To ensure the constraints on the autoregressive terms δ and ρ**,** we parameterize these terms as $\delta=1/[1+\exp(\tilde{\delta})]$ and $\rho=1/[1+\exp(\tilde{\rho})]$, respectively. Once estimated, the $\tilde{\delta}$ and $\tilde{\rho}$ estimates can be translated back to estimates of δ and ρ. The matrix $\boldsymbol{\Omega}$ can be guaranteed to be positive definite by writing the logarithm of the pairwise-likelihood in terms of the Cholesky-decomposed elements of $\boldsymbol{\Omega}$ and maximizing with respect to these elements of the Cholesky factor. That is, we write $\boldsymbol{\Omega}$ as $\mathbf{LL}'$ (where **L** is the lower triangular Cholesky factor of $\boldsymbol{\Omega}$), and estimate the elements of the matrix **L**.

6.3 Simulation Study

In this section, we undertake a simulation experiment with two objectives in mind. The first objective is to examine the ability of the proposed CML inference approach to recover the parameters of the spatial panel ordered-response model in this study. The second is to examine the effects of ignoring spatial dependence and spatial heterogeneity (when both are actually present).

6.3.1 Experimental Design

To set up the experiment, we generate 400 observations (*i.e.*, *QT* = 400) using prespecified values for the $\boldsymbol{\theta}$ vector. We assume that the generated observations correspond to 100 parcels (*i.e.*, *Q* = 100) and 4 time periods (*i.e.*, *T* = 4). We further assume that there are three ordered categories of the observed land use development intensity level and the corresponding threshold values are set to –1 (ψ_1) and 1 (ψ_2). We also consider three independent variables (**x**) in the analysis, all of which are drawn from standard univariate normal distributions. We consider the coefficient on the first variable to be fixed, but allow randomness in the next two elements of the coefficient vector. Specifically, the covariance matrix of $\boldsymbol{\beta}_{\mathbf{q}}$ is specified to be as follows:

$$\boldsymbol{\Omega} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & {\sigma_{22}}^{2} & 0 \\ 0 & 0 & \sigma_{33}^{2} \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

The mean vector for $\boldsymbol{\beta}_{\mathbf{q}}$ is set to **b** = (0.5, 0.8, 1). Next, we generate the weight matrix (**W**) by borrowing the spatial locations of 100 parcels in Austin, Texas, based on the 2008 land use survey data that is used in the empirical analysis of this study (see Section 6.4). While several different functional forms may be used to generate the weights from the spatial configuration of the 100 parcels, we use a continuous inverse of distance specification in this simulation analysis. We also consider all the $QT(QT-1)/2$ pairs of bivariate probability computations in the composite marginal likelihood function for the simulation. To examine the potential impact of different levels of spatial and temporal dependence on the performance of the CML approach, we consider two values of the spatial autoregressive coefficient δ corresponding to low dependence (δ = 0.25) and high dependence (δ = 0.75), as well as two values of the temporal autoregressive coefficient ρ corresponding to low dependence (ρ = 0.25) and high dependence (ρ = 0.75). Thus, in total, there are four possible combinations of the spatial and temporal autoregressive coefficients considered in the simulations.

The set-up above is used to develop the **B** matrix and the $\boldsymbol{\Sigma}$ matrix (see Equation (6.4)) for each of the four combinations just discussed. A $(QT \times 1)$ vector of the latent variable $\mathbf{y}^{*}$ (in Equation (6.3)) is drawn from the multivariate normal distribution with mean **B** and covariance structure $\boldsymbol{\Sigma}$. The generated latent variables are then translated into the "observed" vector **y** using the specified threshold values. For each of the four combinations, the data generation process is undertaken 20 times with different realizations of the latent variable $\mathbf{y}^{*}$ from the values of **B** and $\boldsymbol{\Sigma}$.

The CML estimation procedure is applied to each data set to estimate data-specific values of the vector $\boldsymbol{\theta} = (\mathbf{b}', \psi_1, \psi_2, \sigma_{22,}, \sigma_{33}, \delta, \rho)'$. The Godambe information-based covariance matrix and the corresponding standard errors are also computed. Finally, for each of the four combinations of the spatial and temporal dependency coefficients, the mean estimate for each model parameter across the twenty data sets is obtained and a parameter-specific mean absolute percentage bias or APB value (relative to the "true" value of the parameter) is computed. Similarly, the mean standard error for each model parameter is computed across the twenty data sets and is labeled as the asymptotic standard error (ASE) for the parameter.

The main purpose of the methodology proposed here is to accommodate spatial dynamics and spatial heterogeneity in the context of panel data. Therefore, to examine the potential problems that could arise from ignoring spatial dynamics and spatial heterogeneity, we estimate two additional models on the twenty data sets generated for each combination of spatial and temporal dependence levels. The first model ignores the spatial autocorrelation coefficient δ (that is, assumes $\delta = 0$), while the second model assumes away any spatial heterogeneity (that is, assumes that all elements of the covariance matrix $\boldsymbol{\Omega}$ are identically zero).[41] For ease in presentation, we will refer to the first model as the ordered-response model with spatial heterogeneity (or the ORH model), and the second model as the ordered-response model with spatial dependence (or the ORS model). We compare these two restrictive formulations with the general ordered-response model with spatial dependence and heterogeneity (or the ORSH model), based on the mean APB measure across all parameters and the adjusted composite log-likelihood ratio test (ADCLRT) value (see Pace *et al*., 2011 and Bhat, 2011 for more details on the ADCLRT statistic, which is the equivalent of the log-likelihood ratio test statistic when a composite marginal likelihood inference approach is used; this statistic has an approximate chi-squared asymptotic distribution (also see Section 2.6)). The

[41] Of course, as indicated earlier, setting all elements of $\boldsymbol{\Omega}$ to zero also implies the absence of time-stationary temporal dependence across observations for the same parcel, as well as leads to a reduction in spatial dependence (see Section 6.2.2).

ADCLRT statistic needs to be computed for each data set separately, and compared with the chi-squared table value with the appropriate degrees of freedom. Here we identify the number of times (out of the 20 model runs corresponding to the 20 data sets) that the ADCLRT value rejects the ORH and ORS models in favor of the ORSH model.

6.3.2 Simulation Results

Tables 6.1a and 6.1b provide the results for the ability of the CML approach to recover the parameters of the spatial panel ordered-response model, while Table 6.2 provides the results showing the implications of ignoring spatial dynamics and spatial heterogeneity when present. We discuss these results in the subsequent two sections, each section focusing on a specific objective of the simulation exercise.

6.3.2.1 Ability of CML to recover model parameters

In the low spatial autoregressive coefficient (δ) case in Table 6.1a, the absolute percentage bias (APB) ranges from 0.03% to 6.22% for the low temporal autoregressive coefficient (ρ) case (overall mean value of 2.28% - see last row of table under the sub-column titled “absolute percentage bias”), and from 0.09% to 7.67% for the high temporal autoregressive coefficient case (overall mean value of 3.06%). In the high spatial autoregressive coefficient case (see Table 6.1b), the APB ranges from 2.50% to 7.62% for the low ρ case (mean of 5.05%), and from 0.55% to 13.74% for the high ρ case (mean of 6.88%). Overall, these are very good measures for the ability to recover parameter estimates, and indicate that the CML is able to recover parameters well. Of course, the results indicate that the recovery of parameters is particularly good for the mean of the coefficients on the exogenous variables (the APB values for the **b** vector elements are, in general, less than 5%; see the first numeric row panel of Tables 6.1a and 6.1b). On the other hand, the standard deviations of the coefficients on the exogenous variables (*i.e.*, the σ_{22} and σ_{33} parameters that correspond to the square root of the elements of the $\mathbf{\Omega}$ matrix) are better recovered for the case of low spatial dependence than for the case of high spatial dependence (see the higher APBs corresponding to these

Table 6.1a Ability of the CML Approach to Recover the Parameters of the Spatial Panel Ordered-Response Model - The Low Spatial Autoregressive Coefficient Case

Parameter	Low temporal autoregressive coefficient (ρ=0.25)				High temporal autoregressive coefficient (ρ=0.75)			
	True Value	Parameter Estimates		Asymptotic Standard Error (ASE)	True Value	Parameter Estimates		Asymptotic Standard Error (ASE)
		Mean Estimate	Absolute Percentage Bias (APB)			Mean Estimate	Absolute Percentage Bias (APB)	
b_1	0.5000	0.4986	0.28	0.0056	0.5000	0.5075	1.49	0.0055
b_2	0.8000	0.7942	0.73	0.0100	0.8000	0.8124	1.55	0.0103
b_3	1.0000	1.0161	1.61	0.0113	1.0000	1.0767	7.67	0.0119
ψ_1	-1.0000	-1.0622	6.22	0.0104	-1.0000	-1.0217	2.17	0.0100
ψ_2	1.0000	1.0116	1.16	0.0110	1.0000	1.0320	3.20	0.0117
σ_{22}	1.0000	1.0397	3.97	0.0183	1.0000	0.9734	2.66	0.0180
σ_{33}	1.0000	0.9406	5.94	0.0182	1.0000	0.9479	5.21	0.0180
δ	0.2500	0.2514	0.58	0.0200	0.2500	0.2586	3.45	0.0212
ρ	0.2500	0.2501	0.03	0.0222	0.7500	0.7507	0.09	0.0053
Overall mean value across parameters			2.28	0.0141	-	-	3.06	0.0124

Table 6.1b Ability of the CML Approach to Recover the Parameters of the Spatial Panel Ordered-Response Model - The High Spatial Autoregressive Coefficient Case

Parameter	Low temporal autoregressive coefficient (ρ=0.25)				High temporal autoregressive coefficient (ρ=0.75)			
	True Value	Parameter Estimates		Asymptotic Standard Error (ASE)	True Value	Parameter Estimates		Asymptotic Standard Error (ASE)
		Mean Estimate	Absolute Percentage Bias (APB)			Mean Estimate	Absolute Percentage Bias (APB)	
b_1	0.5000	0.4780	4.40	0.0058	0.5000	0.4978	0.43	0.0065
b_2	0.8000	0.8354	4.43	0.0103	0.8000	0.8270	3.37	0.0117
b_3	1.0000	1.0528	5.28	0.0121	1.0000	1.0975	9.75	0.0143
ψ_1	-1.0000	-1.0757	7.57	0.0123	-1.0000	-1.1374	13.74	0.0142
ψ_2	1.0000	1.0250	2.50	0.0119	1.0000	0.9945	0.55	0.0125
σ_{22}	1.0000	0.9499	5.01	0.0179	1.0000	0.8710	12.90	0.0326
σ_{33}	1.0000	0.9444	5.56	0.0168	1.0000	0.9115	8.85	0.0202
δ	0.7500	0.6929	7.62	0.0034	0.7500	0.6739	10.14	0.0034
ρ	0.2500	0.2422	3.12	0.0087	0.7500	0.7339	2.15	0.0103
Overall mean value across parameters			5.05	0.0110	-	-	6.88	0.0140

parameters in the third numeric row panel of Table 6.1b compared to Table 6.1a). This is not surprising, since these covariance parameters enter the likelihood function in a more complex non-linear fashion in general than the mean parameters of the coefficients. This is particularly so in the presence of high spatial dependence, since the **S** matrix gets applied in a non-linear fashion to the $\mathbf{\Omega}$ matrix during estimation (see Equation (6.4)). But when the spatial dependence is low, the non-linear effect is not as high as in the case of the high spatial dependence case, leading to the better recovery ability of the standard deviation parameters. The results also indicate that the ability to recover the threshold parameters (*i.e.*, ψ_1 and ψ_2) is, in general, better and more stable in the case of low temporal dependence than in the case of high temporal dependence (see the lower APBs corresponding to these threshold parameters in Tables 6.1a and 6.1b). This is an issue that needs further exploration in future studies.

Finally, there are also patterns in the ability to recover the spatial and temporal autoregressive parameters. For the low spatial autoregressive parameter (δ = 0.25), the APB values are 0.58% and 3.45% for the low and high temporal autoregressive coefficient cases, respectively. For the high spatial autoregressive parameter (δ = 0.75), the corresponding APB values are 7.62% and 10.14%, respectively. The implication is that the spatial dependency parameter may be relatively easy to recover when the magnitudes of the spatial and temporal dependency autoregressive coefficients are both small. However, for the temporal dependency parameter ρ, the results indicate very good recovery and stability for all different combinations of the δ and ρ parameters. This is because the parameter ρ is directly associated with the magnitude of correlation across observations on the same spatial unit, and changes in this parameter will have immediate and substantial impacts on the log-likelihood function (regardless of the magnitude of the spatial dependency effect or the magnitude of ρ itself).

The asymptotic standard error (ASE) values of the parameters indicate that the CML estimator appears to be quite efficient. In particular, the ASE values of all the parameters, except δ and ρ, range from 1-4% of the mean estimates. For δ and ρ, the ASE values range from 0.5-8.2% and 0.7-8.9% of the mean estimates, respectively.

6.3.2.2 Effects of ignoring spatial effects

This section focuses on the implications of ignoring each of spatial dynamics and spatial heterogeneity when both are present. To examine the effect of ignoring spatial dynamics when present, the results of the ORH model may be compared with those from the ORSH model. On the other hand, to assess the impact of ignoring spatial heterogeneity when present, the results of the ORS model may be compared with those from the ORSH model. Table 6.2 provides the results. As may be observed, two sets of mean APB values are computed for the ORSH model, one for comparison with the ORH model and another for comparison with the ORS model. For comparison with the ORH model, the mean APB values for the ORSH model are computed without considering the APB values for the δ parameter, because the δ parameter is implicitly fixed at zero in the ORH model. For comparison with the ORS model, the mean APB values for the ORSH model are computed without considering the APB values for the σ_{22} and σ_{33} parameters (that correspond to the square root of the elements of the $\mathbf{\Omega}$ matrix characterizing spatial heterogeneity). Note again that the σ_{22} and σ_{33} parameters are implicitly fixed to zero in the ORS model.

The results indicate that the mean APB values are higher for the ORH and ORS models than for the ORSH model. Not surprisingly, the ORH model performs better in the two low spatial dependence cases than in the two high spatial dependence cases, since ignoring spatial dependence when such dependence is low should be of less consequence than ignoring such dependence when high. However, even in the two low spatial dependence cases, the ORH model may be rejected compared to the "correct" ORSH specification based on the adjusted composite likelihood ratio test (ADCLRT) statistic (note that the ORSH specification rejects the simpler ORH and ORS specifications for each of the twenty data sets generated). The results also indicate that the ORS model (which ignores spatial heterogeneity) performs very poorly across the board. In this regard, we should also point out that the ORSH and ORH models always converged, while the ORS model experienced occasional convergence-related problems in the high spatial dependence case. In particular, because of convergence problems, the results in

Table 6.2 Effects of Ignoring Spatial Effects When Present

Evaluation Metric	***δ* = 0.25, *ρ* = 0.25**			***δ* = 0.25, *ρ* = 0.75**			***δ* = 0.75, *ρ* = 0.25**			***δ* = 0.75, *ρ* = 0.75**		
	ORSH Model	**ORH Model**	**ORS Model**	**ORSH Model**	**ORH Model**	**ORS Model**	**ORSH Model**	**ORH Model**	**ORS Model**	**ORSH Model**	**ORH Model**	**ORS Model**
Mean APB												
For comparison of ORSH model with ORH model	2.49	3.07	-	3.01	3.07	-	4.73	16.14	-	6.47	17.53	-
For comparison of ORSH model with ORS model	1.51	-	35.14	2.80	-	29.09	4.99	-	27.61	5.73	-	29.14
Mean composite log-likelihood value at convergence	-135,448	-135,522	-139,956	-133,954	-134,050	-138,155	-133,275	-134,792	-136,781	-132,667	-134,143	-136,948
Number of times the adjusted composite likelihood ratio test (ADCLRT) statistic favors the ORSH model	-	All twenty times when compared with $\chi_1^2 = 3.84$ value (mean ADCLRT statistic is 97.80)	All twenty times when compared with $\chi_2^2 = 5.99$ value (mean ADCLRT statistic is 6,693.97)	-	All twenty times when compared with $\chi_1^2 = 3.84$ value (mean ADCLRT statistic is 139.38)	All twenty times when compared with $\chi_2^2 = 5.99$ value (mean ADCLRT statistic is 5,834.31)	-	All twenty times when compared with $\chi_1^2 = 3.84$ value (mean ADCLRT statistic is 2,173.70)	All fifteen times when compared with $\chi_2^2 = 5.99$ value (mean ADCLRT statistic is 6,395.06)	-	All twenty times when compared with $\chi_1^2 = 3.84$ value (mean ADCLRT statistic is 2,073.70)	All eighteen times when compared with $\chi_2^2 = 5.99$ value (mean ADCLRT statistic is 4,862.69)

Table 6.2 for the ORS model are based on estimations on fifteen data sets for the (δ = 0.75 , ρ = 0.25) case and on eighteen data sets for the (δ = 0.75 , ρ = 0.75) case. Also, the ORS model is clearly outperformed by the ORSH model.

Overall, the simulation results show that the CML estimator recovers the parameters of the spatial panel ordered-response model very well. The CML estimator also seems to be quite efficient based on the low asymptotic standard error estimates of the parameters compared to the mean estimates of the parameters. In addition, the results clearly highlight the bias in estimates if spatial dependence and/or spatial heterogeneity is ignored when both are actually present. An interesting suggestion from our simulation study is that ignoring spatial heterogeneity is of much more serious consequence than ignoring spatial lag dynamics. Further theoretical and empirical exploration of this finding is left for future work.

6.4 Data

6.4.1 Data Sources

The primary data used in the empirical exercise of this study is drawn from the land use data sets collected by the City of Austin Watershed Protection and Development Review Department for the years 2000, 2003, 2006, and 2008 (City of Austin, 2011).[42] For each analysis year, the land use information considered in the empirical analysis represents the ground land use condition at that time.[43] The City of Austin uses a 3-digit land use code that classifies the collected information into different land use types such as single-family, multi-family, mobile homes, apartment/condo, group quarters, office, industrial, and open space/vacant land (see City of Austin, 2011 for a complete list of land use classifications). This land use information is maintained at a parcel-level spatial resolution and made available to the public in Geographic Information System (GIS) format (shape file format).

[42] 2008 is the latest year for which land use information for the City of Austin is available.

[43] Specifically, the data sets describe ground conditions in October 2000, June 2003, June 2006, and October 2008, which are about equally spaced in time (the time period between successive data collection efforts spans between 2 years 4 months and 3 years).

In addition to the land use information, several other secondary GIS data sets are used to obtain supplementary information. These include:

1) A GIS transportation network layer for the study area, obtained from the City of Austin. The transportation network is represented as street centerlines and includes information such as street name, roadway functional class, and speed limit.

2) A GIS school location layer for the Austin area, obtained from the Texas Education Agency (School data, 2010). This layer includes information such as school name, location, grade (elementary, middle school, or high school), and teaching institution type (regular/alternative).

3) A GIS layer with information on parks in the Austin area, including park name, park type (neighborhood, greenbelt, or nature preserve), and park location. This GIS layer was obtained from the City of Austin.

4) A GIS layer with information on water bodies in the Austin area, as obtained from the City of Austin. This layer includes the locations of Lake Travis, Lake Austin, Lady Bird Lake, Walter E. Long Lake, and Colorado River.

5) A GIS layer on city boundaries for Austin and other neighboring cities, obtained from the Capital Area Council of Governments (CACOG, 2010).

6) A GIS layer on aircraft landing facilities, such as airports and airfields in the Austin area. This GIS layer was obtained from the Capital Area Council of Governments (CACOG, 2010).

7) A GIS contour layer with information on average elevation at different points in the study area. This GIS layer was obtained from the Capital Area Council of Governments (CACOG, 2010).

6.4.2 Sample Formation and Description

The land use data (and the data from the secondary sources) were processed in several steps to obtain the sample for the current analysis. First, the land use GIS layers (created by the City of Austin) for the years 2000, 2003, 2006, and 2008 were spatially merged. Second, a 1.75 square miles (4.53 square kilometers) area near the western boundary of the City of Austin was selected for this study. This area was selected because the land use pattern here has undergone substantial changes between 2000 and 2008. Third, information on the land use of each parcel in each year was translated into four mutually exclusive ordinal land development intensity categories for this study: (1) undeveloped land (includes open space, rural area, agricultural land, and vacant parcels), (2) land developed with low level of intensity (includes residential parcels with single-family detached and two-family attached homes, (3) land developed with medium level of intensity, including all other types of residential parcels such as apartment, condo, three/fourplex, group quarters, and retirement homes), and (4) land developed with high level of intensity, including parcels developed for office, commercial, and industrial use). Note, however, that the development intensity classification used in the current study is simply one of many that may be used by the analyst. Specifically, the intensity classification may be customized to the planning purpose at hand. Fourth, variables derived from the secondary data sources were appended to the parcel-level data. The final sample for analysis includes land use information for 783 parcels.

Table 6.3 presents the number (and the percentage) of parcels by land use development intensity (LUDI) and year of observation. The table clearly indicates the rapid pace of development between 2003 and 2006, which is consistent with the general ground reality in the Austin area (see http://www.ci.austin.tx.us/landuse/tabular.htm and http://www.ci.austin.tx.us/growth/). While 36-37% of the land parcels were undeveloped in 2000 and 2003, this percentage drops to 10-13% by 2006 and beyond. During the analysis time period, the shares of medium-intensely and most-intensely developed parcels remained somewhat constant, indicating that the land owners found converting undeveloped parcels to less-intensely developed parcels to be the most profit maximizing investment.

Table 6.3 Number (Percentage) of Parcels by Land Use Development Intensity (LUDI) Level and Year of Observation

Land Use Development Intensity (LUDI) Level	**Year of Observation**			
	2000	**2003**	**2006**	**2008**
Undeveloped land (includes open space, rural area, agricultural land, and vacant parcels)	285 (36.4)	290 (37.0)	98 (12.5)	80 (10.2)
Less-intensely developed land (includes residential parcels with single-family detached and two-family attached homes)	469 (59.9)	450 (57.5)	642 (82.0)	660 (84.3)
Medium-intensely developed land (includes all other residential parcels such as apartment, condo, three/fourplex, group quarters, and retirement homes)	14 (1.8)	26 (3.3)	22 (2.8)	22 (2.8)
Most-intensely developed land (includes parcels developed for office, commercial, or industrial use)	15 (1.9)	17 (2.2)	21 (2.7)	21 (2.7)
Total number of parcels	**783**	**783**	**783**	**783**

6.5 Empirical Analysis

6.5.1 Model Selection and Variable Specification

Several weight matrix specifications were considered in our empirical analysis to characterize the nature of the dynamics of the spatial lag dependence. These included (1) a contiguity specification that generates spatial dependence based on whether or not two parcels are contiguous, (2) another contiguity specification but based on shared boundary length, (3) the inverse of a continuous distance specification where the distance is measured as the Euclidean distance (crow fly distance) from the centroids of each parcel, (4) the inverse of the square of the continuous distance specification, and (5) the inverse of the exponential of the continuous distance specification. For the last three continuous distance-based specifications, we also explored alternative distance bands to select the pairs of observations for inclusion in the composite marginal likelihood (CML) estimation. As indicated earlier, this distance band determination may be based on minimizing the trace of the variance matrix of parameters given by $tr[\mathbf{V}_{CML}(\hat{\boldsymbol{\theta}})]$. Our results did not show substantial variations in the trace value for different distance bands (regardless of the specific continuous functional form used to represent the distance separation and the variable specification used), though the best estimator efficiency was obtained at about 0.25 miles for all the three continuous distance specifications formulations and all variable specifications we attempted. Further, the results indicated that for all variable specifications, the best spatial weight matrix specification was consistently the inverse of the continuous distance specification with the 0.25 mile distance band. This determination was based on the composite likelihood information criterion (CLIC) statistic, which may be used to compare the data fit of non-nested formulations, as discussed in Section 2.6. This CLIC statistic takes the form shown below (see Varin and Vidoni, 2005):

$$\text{CLIC} = \log L_{CML}(\hat{\boldsymbol{\theta}}) - tr\left[\hat{\mathbf{J}}(\hat{\boldsymbol{\theta}})\hat{\mathbf{H}}(\hat{\boldsymbol{\theta}})^{-1}\right]$$

where $\hat{\boldsymbol{\theta}}$ is the estimated model parameter vector, and $\hat{\mathbf{J}}(\hat{\boldsymbol{\theta}})$ and $\hat{\mathbf{H}}(\hat{\boldsymbol{\theta}})$ are the “vegetable” and “bread” matrices used in the estimation of the asymptotic variance matrix $\mathbf{V}_{CML}(\hat{\boldsymbol{\theta}})$ (see Section 6.2.3). In the current context, the weight specification that provides the highest value of the CLIC statistic is preferred over the other competing weight specifications. Of all the weight matrix specifications that were considered here, the best three specifications and the corresponding CLIC statistics are presented in Table 6.4. These statistics correspond to the best variable specification that emerged from our empirical analysis (see the next paragraph for more on this) and for the optimal distance band of 0.25 miles for the continuous distance weight specifications. The results in the table clearly show the superiority of the inverse of the continuous distance specification over other weight matrix specifications. Thus, all subsequent results in this study correspond to the inverse distance weight specification with a 0.25 mile distance band.

Table 6.4 Model Selection Based on the Weight Matrix Specification

	Weight Matrix Specification		
	Contiguity	**Inverse of continuous distance (0.25 mile distance band)**	**Inverse of continuous distance square (0.25 mile distance band)**
Log-composite likelihood at convergence	-724619.52	-718753.28	-720435.17
Trace value	1780.35	1343.63	2338.49
CLIC statistic	-726399.87	-720096.92	-722773.66

Concurrent with the weight matrix specification, we also explored several different variable specifications and functional forms of the variables. The final specification included the following three sets of variables: (1) proximity (in the form of distance) to natural amenities (such as parks and lakes), schools, and the central business district (CBD) area of Austin,[44] (2) ease of access to the transportation system (distance to Interstate IH-35 and distance to a public airfield), and (3) year-specific dummy variables (for the years 2006 and 2008) and geographic location/contour variables (whether or not the parcel is located within the Austin City limit and the average elevation of a parcel above the sea level). For the first two sets of variables, several linear

[44] Parks as used here refers to such natural outdoor recreations areas as parks, greenbelts, and nature preserves. Similarly, a lake as used here refers to either Lake Travis, Lake Austin, Lady Bird Lake, Walter E. Long Lake, or Colorado River.

and non-linear functional forms were considered (such as the logarithm of distance, the square of distance, and spline variables that allow piece-wise linear effects of distance on the utilities). In addition, we also considered dummy variables for different ranges of distance for these variables (for instance, parcel is within 2 miles of a park and parcel is within 5 miles of a park). Further, various interactions of the many variables were also considered whenever adequate observations were available to test such interaction effects. The final specification was based on intuitive, data fit, and statistical significance considerations. Interestingly, all the distance variables were best reflected as dummy variables in this final specification, though the threshold value for translation of the distance variables to the dummy variables varied across the variables. The final specification includes some variables that are not statistically significant at the usual 5% level of significance. These are retained because the effects of these variables are intuitive and may provide guidance in future research efforts. The results of the final specification are discussed in the next section.

6.5.2 Model Estimation Results

Table 6.5 presents the model estimation results. The column titled "Parameter - Mean Estimate" provides the mean estimate of each parameter and the corresponding t-statistic of the mean estimate. Each of these estimates provides the mean effect of the corresponding row variable on the land use development returns (LUDR) perception of land owners. Since all the variables in the final specification appear as dummy variables, the relative magnitudes of the mean effects provide an estimate of the importance of the variable in affecting the LUDR perception of land owners. Note also that we attempted a (normally distributed) random coefficients specification for the variables through a general specification of the $\mathbf{\Omega}$ matrix. However, only the variance parameters corresponding to the constant, "distance to a lake", and "distance to an airfield" variables turned out to be statistically significant. Further, we could not reject the null hypothesis that the off-diagonal (covariance) elements of the $\mathbf{\Omega}$ matrix corresponding to these random coefficients were all zero. The column titled "Parameter - Standard Deviation

Table 6.5 Model Estimation Results (Weight Matrix: inverse of distance, Distance Band: 0.25 miles)

	Parameter - Mean Estimate		Parameter - Standard Deviation Estimate	
	Estimate	t-stat	Estimate	t-stat
Constant	0.000	-	0.006	4.25
Closeness to natural amenities, school, and the CBD				
Distance to a park ≤ 2 miles (base: park > 2 miles)	0.112	1.21	-	-
Distance to a lake ≤ 5 miles (base: lake > 5 miles)	0.623	5.38	1.301	8.38
Distance to a school ≤ 2 miles (base: school > 2 miles)	0.044	1.19	-	-
Distance to the downtown area ≤ 9 miles (base: downtown > 9 miles)	-0.203	-1.56	-	-
Ease of access to the transportation system				
Distance to IH-35 ≤ 9 miles (base: IH-35 > 9 miles)	0.322	5.15	-	-
Distance to a public airfield ≤ 1 miles (base: airfield > 1 miles)	-0.224	-2.44	0.355	1.91
Year-specific dummy variables and other variables				
Year 2006 (base: Years 2000/2003)	0.136	4.08	-	-
Year 2008	0.147	4.36	-	-
Parcel is located in Austin city (base: parcel is located outside Austin city)	-0.807	-4.88	-	-
Average elevation of parcel ≤ 1000 feet above mean sea level (base: average elevation > 1000 feet)	-0.242	-3.39	-	-
Auto-regressive parameters[45]				
Spatial auto-regressive co-efficient (δ)	0.905	50.49	-	-
Temporal auto-regressive co-efficient (ρ)	0.344	1.59	-	-
Thresholds				
ψ_1	-5.438	-6.66	-	-
ψ_2	-1.850	-6.77	-	-
ψ_3	-1.267	-6.14	-	-

[45] Standard errors of the auto-regressive parameters are estimated using the delta method.

Estimate" provides the standard deviation estimates of the random coefficients and their corresponding t-statistics.

The first variable in Table 6.5 corresponds to the constant, whose mean estimate is fixed at zero for identification. However, the statistically significant estimate of the standard deviation on the constant indicates that there is unobserved heterogeneity in the LUDR perception across land owners, attributable to such unobserved factors as individual experiences, risk-taking behavior, and vegetation conservation values. In the following sections, we discuss the effects of the non-constant variables on the latent LUDR perception by variable category.

6.5.2.1 Proximity to Natural Amenities, School, and the CBD

The effects of this set of variables suggests that parcels located within close proximity of a park (distance ≤ 2 miles) and/or a lake (distance to a lake ≤ 5 miles distance) are perceived by land owners as providing high returns to development relative to parcels located farther away from such natural amenities. These effects are to be expected, since areas with good access to natural recreation are prime profitable locations for residential land use (see Espey and Owusu-Edusei, 2001, and Geoghegan 2002). Interestingly, however, the results show substantial variation in the LUDR perceptions of land owners of parcels within 5 miles of a lake, with 32% of landowners having a negative LUDR perception and 68% having a positive LUDR perception. This may suggest variations in nature conservation values across land owners, so that some land owners of parcels close to lakes may place a high premium on keeping their land undeveloped and "pristine".

Proximity to a school also affects land development intensity level. As expected, owners of parcels close to a school (school ≤ 2 miles) are likely to perceive their parcels as having high development value (see Li and Liu, 2007). The final variable in this category indicates a lower LUDR perception for parcels located in close proximity (≤ 9 miles) of the Austin CBD relative to those located farther away (> 9 miles). This is interesting, and suggests the tension between the urban amenities (access to retail places and public services such as hospitals) on the one hand that may increase the demand for development in already densely developed areas, and the urban "disamenities" (such as traffic congestion effects and air quality problems) on the other hand that may decrease demand for development in already dense neighborhoods (see Anas *et al.*, 1998, Irwin and Bockstael, 2002, and Carrión-Flores and Irwin, 2004). According to our results, the "disamenities" effect exceeds the "amenities" effect offered by parcels located in close proximity to the Austin CBD area, leading to an overall negative LUDR perception for these parcels.

6.5.2.2 Ease of Access to the Transportation System

Several earlier studies (for instance, see Carrión-Flores and Irwin, 2004 and Chakir and Parent, 2009) have found that proximity and access to major roadways generally has a positive impact on development intensity (even if certain kinds of developments such as industrial facilities are precluded by zoning regulations to be located very close to major roadways). The result on the "distance to IH-35" variable in Table 6.5 is consistent with these earlier studies, and indicates that parcels in the analysis area within 9 miles of IH-35 are less likely to be in an undeveloped state than parcels farther away from IH-35.

The second variable in the "access to transportation system" category shows that land owners of parcels that are proximal to a public airfield (distance to an airfield ≤ 1 mile) are, on average, likely to have a negative perception of the profitability of development of their land; that is, these land owners are more likely to keep their land undeveloped than invest money in development. This is perhaps because of noise pollution and air space invasiveness effects of aircrafts landing or taking off from airfields. However, it is important to note that there is heterogeneity in the LUDR perception of land owners of parcels close to airfields, with 25% of land owners perceiving a positive LUDR (see the standard deviation estimate of the "distance to airfield ≤ 1 mile" variable in Table 6.5). Such LUDR heterogeneity is not surprising, since some parcels close to airfields may not be that impacted by aircraft noise and space invasiveness because of the alignment of runways vis-à-vis the parcel location. For these parcels, the close proximity to air transport may be more of a “pull” effect than a “push” effect.

6.5.2.3 Year-Specific Dummy Variables and Other Variables

The dummy variables for 2006 and 2008 essentially reflect the higher propensity of parcels to be developed in some form or the other relative to 2000 and 2003. This trend of a higher development intensity pattern after 2005 (relative to before 2005) is consistent with the actual trend observed in land development intensity in the Austin area (see, for example, http://austin.housealmanac.com). The final two variables suggest that land owners of parcels located within Austin city limits and located at a lower elevation (less than or equal to 1000 ft above sea level) have a lower LUDR perception than land owners of parcels located outside Austin city limits and at a higher elevation (more than 1000 ft above sea level), respectively.

6.5.2.4 Autoregressive Parameters and Thresholds

The results indicate the presence of spatial dependence in land use development decisions. Specifically, the estimated spatial autoregressive coefficient (δ) is 0.905 and highly statistically significant, strongly supporting the hypothesis of the presence of spatial spillover effects in the LUDR perceptions of land owners of proximally located spatial units. That is, there is strong evidence of didactic interactions between land owners of proximally located parcels.

The temporal autoregressive coefficient (ρ) is also moderately statistically significant with a magnitude of 0.344. This is evidence of the presence of land owner-specific unobserved effects that fade over time. Of course, this temporal fading effect is in addition to the time-invariant unobserved effects that influence the LUDR perception of a land owner at all time points (as captured by the random coefficients on the constant, the “distance to a lake” variable, and the “distance to a public airfield” variable).

Finally, the thresholds values serve to translate the latent propensity into the observed ordered categories of the land use type.

6.5.2.5 Overall Measures of Fit

The results of the spatial panel ordered-response model estimated in the current study show clear evidence of spatial heterogeneity, spatial lag dynamics due to didactic interactions between land owners, as well as time-variant temporal correlation in the LUDR perceptions of the same individual. Thus, the model estimated here is superior to a model that ignores these spatial and temporal effects. One can also assess the data fit degradation from ignoring spatial and temporal effects by estimating a simple ordered-response (OR) model that assumes away the presence of these spatial-temporal effects. An adjusted composite likelihood ratio test (ADCLRT) statistic can then be computed from the composite marginal likelihood values at convergence of the model estimated here and the simple OR model. This statistic has a chi-square asymptotic distribution with 5 degree of freedom. The statistic has a value of 11,874, which is higher than the corresponding critical chi-squared value with five degree of freedom and soundly rejects the OR model at any reasonable level of significance. This again demonstrates very strong evidence of spatial dynamics and temporal dependence at play in land-use development intensity decisions.

6.6 Summary and Conclusions

This study proposes and estimates a spatial panel ordered-response probit model with temporal autoregressive error terms to analyze changes in urban land development intensity level over time. Such a model structure maintains a close linkage between the land owner's decision (unobserved to the analyst) and the land development intensity level (observed by the analyst), and accommodates proximity-based spatial didactic interactions among the land owners that causes "spillover" effects. In addition, temporal dependency (due to unobserved factors) is generated across the LUDR perceptions of the same land owner over time – the effects of some of these factors may fade away over time, while the effects of other factors may remain time-invariant. The model structure also incorporates (unobserved) spatial heterogeneity by allowing the sensitivity to exogenous variables to vary across land owners.

The study addresses the well recognized econometric challenge of estimating spatial discrete choice models with medium-to-large sized sample by using a composite marginal likelihood (CML) inference approach in estimation. The CML approach can be applied to data sets of any size and does not require any simulation machinery. To evaluate the ability of the CML approach to recover model parameters in a spatial-temporal context, we undertake a simulation exercise. The results indicate that the CML approach recovers the parameters reasonably well. In addition, the simulation study demonstrates that ignoring spatial dependency and spatial heterogeneity when both are actually present will introduce substantial bias. Further, there is a suggestion in the result that ignoring spatial heterogeneity is of much more serious consequence than ignoring spatial lag dynamics.

The model system proposed in the current study is applied to examine urban land development intensity levels using parcel-level data from Austin, Texas. The results suggest that closeness to natural and other amenities (such as park, lake, school, and urban center), distance to major roadways, average elevation of the parcel, and whether or not the parcel is located in Austin city have significant effect on the LUDR perceptions of the land owners. The results also indicate the presence of spatial “spillover” effects (caused by didactic interactions among the land owners), spatial heterogeneity, and time-varying temporal effect in the LUDR perceptions of the same land owner. The findings from this analysis underscore the importance of considering such effects in the study of land development intensity level to obtain consistent parameter estimates.

Part III

Chapter 7

Applications of the Models

7.1 Introduction

The behavior-oriented models estimated in chapters 4 through 6 have many statistically significant parameters that indicate the superiority of each of these models over their corresponding restricted version (or naïve model). However, the purpose here is not just to estimate a series of models that offer better data fit but also to propose models that have practical applications and can be used to undertake policy analyses. The models estimated in the current dissertation can be used to perform such analyses. For example, the multivariate ordered-response model developed in Chapter 4 can be used to determine the change in the number of out-of-home episodes for each activity purpose-accompaniment type combination due to changes in individual- and/or household-level socio-demographic characteristics over time. This type of analysis provides important insight on how a change in one explanatory variable can affect activity behavior of individuals. Also, the model systems proposed here can be used to examine differential impacts of changing trends in policy variables on different demographic segments of population.

In addition to demonstrating practical applications of the models proposed in this dissertation, another objective of the current chapter is to answer questions such as "is there any tangible benefit of adopting the behavior-oriented model over the naïve model?" and "how much better off one would be if the behavior-oriented model is used instead of the naïve model?". We answer these questions by assessing the predictive capability of the two models in a comparative framework. The rest of the chapter is structured as follows. Section 7.2 discusses an application of the multivariate ordered-response model developed in Chapter 4. Section 7.3 demonstrates an application of the joint model of walking and bicycling activity duration estimated in Chapter 5. Section 7.4 illustrates an application of the spatial panel ordered-response model proposed in Chapter 6. The last section provides a brief summary and concludes the chapter.

7.2 Application of the Multivariate Ordered-Response Model With Flexible Error Structure

This model is used to examine the change in the adults' out-of-home episode participation behavior by activity purpose-accompaniment type combination due to changes in the independent variables over time. This is particularly important because of changing employment-related and demographic trends. For instance, the number of employed individuals is projected to continue to rise (albeit at a slower rate than in the past), despite the short-term slump due to the economy (see the latest national employment projections to 2016 by the Bureau of Labor Statistics, 2007). Also, according to the US Census Bureau estimates from the Current Population Survey (CPS) (see US Census Bureau, 2009a), the structure of the household is changing with a decrease in nuclear family households and an increase in single individual households. Such socio-demographic changes will have an effect on weekday episode participation, and the model developed in Chapter 4 can be used to assess these impacts and provide reliable information that can be used for activity-based travel demand forecasting and air quality analysis.

In this section, we demonstrate the application of the model by studying the effect of two socio-demographic changes. The first is an increase in the number of full-time employed adults and the second is a decrease in nuclear family households along with a concomitant increase in single individual households. The increase in the number of full-time employed adults is reflected by randomly selecting current non-employed adults in the sample and designating them as full-time employees so that the number of full-time employees increases by 20% over the current full-time employment level. As indicated earlier, such a change mirrors the projected increase in employment levels in the U.S. population. The change in nuclear family households is similarly “implemented” by randomly selecting 20% of individuals who belong to nuclear families and placing them in single individual households. The impact of the two changes discussed above is evaluated by modifying exogenous variables to reflect the change, computing revised expected aggregate values for number of episodes in each combination category, and then obtaining a percentage change from the baseline estimates.

The effects of the changes in variables can be evaluated on each combination level of number of episodes across all the 30 episode categories. But there are about 80 trillion such combination levels. So, in this section, we present the results only for the episode level combinations for two categories: meals with friends and physically inactive recreation with friends. These are two of the most common episode categories participated in during weekdays, as observed earlier in Section 4.3.2. Besides, the estimation results indicate that employment status and household structure, the two variables being examined here, have a direct influence on the “meals with friends” and “physically inactive recreation with friends” categories.

Table 7.1 presents the results from both the multivariate ordered-response probit model (MORP) (plain font) and independent ordered-response probit model (IORP) (italicized font) models. For each model, the predicted change in the number of individuals participating in each combination level of “meals with friends” and “physically inactive recreation with friends” is computed as a percentage of the baseline (actual) numbers of individuals in each combination level. For ease in presentation, and also because the share of individuals participating in three or more episodes of physically active recreation with friends is very small, we have consolidated the 2 and 3 episode levels into a single 2+ episode level in Table 7.1. The results show a decrease in the (0,0) combination level due to an increase in full-time employed adults and decrease (increase) in nuclear family (single individual) households. This is, of course, because of the positive effect of full-time employed status on both the episode categories under consideration, and the negative (positive) effect of nuclear family household (single individual households) on both the episode categories (see Table 4.2). However, the percentage reduction in the number of individuals in the (0,0) cell is lower in the MORP

Table 7.1 Impact of Changes on the Percentage of Individuals Choosing Each Combination Level of "Meals with Friends" and "Physically Inactive Recreation with Friends" Episodes

Change	**Number of "meals with friends" episodes**	**Number of "physically inactive recreation with friends" episodes**		
		0	1	2+
Increase in full-time employed adults by 20% (and corresponding decrease in the number of non-employed adults)	0	-3.99[46] *-4.38*[47]	-3.79 *-1.70*	-4.14 *1.24*
	1	10.52 *10.92*	8.24 *7.33*	6.54 *6.62*
	2	13.67 *20.12*	15.56 *9.48*	34.07 *13.46*
Decrease in nuclear family households by 20% (and corresponding increase in the number of single individual households)	0	-1.57 *-1.80*	2.07 *2.66*	4.08 *9.03*
	1	0.82 *0.94*	4.12 *4.18*	5.84 *5.46*
	2	2.80 *5.47*	6.86 *5.83*	22.66 *10.87*

[46] Percentage change in the number of individuals from the MORP model participating in each combination level of episode category.

[47] Percentage change in the number of individuals from the IORP model participating in each combination level of episode category.

case because of the positive correlation in the propensities of participation in the two episode categories. At the other extreme, both models show, as expected, an increase in the (2,2+) combination level. However, the MORP model indicates a substantially higher increase because of the complementary effect (positive correlation) in the unobserved propensities. The changes in the other cells, in general, also show a shift toward combinations of higher levels of episode participation in the two episode categories due to changes in the socio-demographic variables.

Overall, the exercise above demonstrates the application of the MORP model to predict the shifts in number of episodes of different activity purposes and accompaniment types due to changing socio-demographic characteristics of the population. In addition, the results also point to the biased results that can be obtained by ignoring the jointness in the propensity to participate in different episode categories.

7.3 Application of the Joint Model of Walking and Bicycling Activity Duration

The model estimated in Chapter 5 can be employed to predict individuals' walking and bicycling activity participation durations. However, in this section we demonstrate usefulness of the model beyond quantifying the use of non-motorized transport mode by analyzing physical activity participation level of individuals (*i.e.*, total time spent by individuals in walking and bicycling activities together). Specifically, the model is applied to predict changes in physical activity participation duration over a period of one week due to changes in two socio-demographic characteristics: age and presence of children in the household. We choose these two variables because of the projected demographic trends. For instance, the US Census Bureau predicts that the senior population in the USA is likely to increase by 40% over the current level in the next five years as 14.93 million baby boomers become senior citizen in that time period. The Bureau also projects that the senior population in the USA will more than double by 2050. At the same time period, the share of population age 15 or less is expected to be less than the senior population. According to the US Census Bureau estimates from the

Current Population Survey (CPS), the household size is changing with an increase in the number of households with no/fewer children (US Census Bureau, 2009b).

In this section, the socio-demographic change corresponding to an increase in the senior citizens was "implemented" by randomly selecting a sample of individuals in the age groups 5 to 10 years and 11 to 15 years (the impacts of these two age groups are statistically significant on the dependent variables) and removing them from these age groups so that the number of senior individuals (age over 65) increased by 40% over the current level. Similarly, the change in the number of households with children was achieved by randomly selecting 20% households with children aged 5 to 10 years (the effect of this variable is statistically significant in the current model) and recoding these records as households with no children. To predict physical activity participation duration due to these socio-demographic changes, we adjusted the relevant independent variables (as just discussed), estimated the discrete durations of walking and bicycling activity, transformed the discrete activity durations to continuous time durations, and combined the continuous time durations of walking and bicycling into a single physical activity participation duration.

Table 7.2 presents the results from the multi-level cross-cluster grouped response probit model (MCGRP) and the independent grouped response probit model (IGRP) models. For each model, the predicted physical activity participation duration was grouped into three categories: low physical activity (duration < 150 minutes/week), medium physical activity (150 ≤ duration ≤ 300 minutes/week), and high physical activity (duration > 150 minutes/week).[48] To examine the effects of the socio-demographic changes, for each model, the predicted change in the number of individuals participating in each level of physical activity is computed as a percentage of the actual numbers of individuals participating in each level of physical activity. The results suggest

[48] In 2008, US Department of Health and Human Services published "2008 Physical Activity Guidelines for Americans", which is designed to provide policymakers, health professionals, and general public with information on the type and the amount of physical activity required to maintain good health and reduce the risk of chronic diseases. The categories of physical activity participation levels used here are compatible with the classifications provided in the guidelines.

Table 7.2 Impact of Changes on the Percentage of Individuals Choosing to Participate in Different Levels of Physical Activity

Change	Level of physical activity	MCGRP model	IGRP model
Increase in senior population (age > 65 years) by 40% (and corresponding decrease in non-senior population)	Low activity (activity duration < 150 minutes/week)	-2.14	-1.19
	Medium activity (150 ≤ activity duration ≤ 300 minutes/week)	0.36	-0.36
	High activity (activity duration > 300 minutes/week)	4.26	3.19
Decrease in the number of households with children by 20% (and corresponding increase in the number of households without children)	Low activity (activity duration < 150 minutes/week)	0.00	-0.71
	Medium activity (150 ≤ activity duration ≤ 300 minutes/week)	-1.46	-1.09
	High activity (activity duration > 300 minutes/week)	2.13	3.19

that both models predict a decrease in the low physical activity participation level and an increase in the high physical activity participation level due to an increase in the share of senior population. This shift is due to the positive effects of the non-senior age group variables on the hazard rates (and negative effects on the activity durations). A decrease in the number of households with children scenario also shows a similar pattern because of the positive co-efficient associated with the presence of children (aged 5 to 10 years) in the household variable. However, the percentage shifts in the number of individuals participating in different levels of physical activity predicted by the MCGRP and the IGRP models are different. The difference between the two model predictions will become even wider if the models results' are applied to a large segment of population to predict the number of individuals likely to change their physical activity participation level due to implementation of some policy action. In general, failing to consider walking and bicycling activity durations jointly and ignoring the unobserved heterogeneity due to individual-, social-, and spatial-specific factors can result in inaccurate and biased forecasting.

7.4 Application of the Spatial Panel Ordered-Response Model

In this section, the model estimated in Chapter 6 is applied to predict the effects of a change in the independent variables on the percentage change in the aggregate share of each ordinal land use intensity category for the year 2008, while accommodating the spatial and temporal dependency effects. This application is motivated by the realization that the parameter estimates presented in Table 6.5 do not directly provide the marginal effects of variables on the probability of the ordinal land use development intensity categories (as observed by Franzese and Hays, 2008, this is an issue seldom considered in the spatial choice literature, with many papers simply presenting the parameter results and stopping there). To obtain a sense of the marginal effects, we compute a "pseudo-elasticity effect" for each variable. In addition, bootstrapping is used to obtain the standard error estimates of the "pseudo-elasticity" effects.

All the exogenous variables in the model estimated in Chapter 6 were introduced as dummy variables. To compute the pseudo-elasticity effects for each of these variables, the value of each variable is changed to one for the subsample of parcels for which the variable takes a value of zero, and to zero for the subsample of parcels for which the variable takes a value of one. The shifts in expected aggregate shares for each ordinal land development intensity (LDI) category in the two subsamples is then added after reversing the sign of the shift in the second subsample. Next, the effective percentage change in the expected share of each ordinal LDI category is computed due to a change in the dummy variable from 0 to 1.

The elasticity effects and their standard errors (in parenthesis) for the ordered-response model with spatial dependence and heterogeneity (the ORSH model) and the simple ordered-response (OR) model are presented in Table 7.3, along with the p-value for the difference in elasticity estimates from the two models. The first entry under the "ORSH model" sub-column in the table indicates that, on average, parcels located within a 2-mile radius of a park are 20.96% less likely to be undeveloped relative to parcels located more than 2 miles away from a park. The other entries under the "ORSH model" sub-columns (and the "OR model" sub-columns) may be similarly interpreted.

Several observations may be made from the results in Table 7.3. First, the numbers in the table indicate the relative importance of each exogenous variable in influencing the ordinal land use development intensity category. For instance, the ORSH model (and the OR model) results indicate that proximity to a lake is the most important determinant of intense land development, with parcels located closer to a lake (≤ 5 miles) being about 150% (2.5 times) more likely to be intensely developed compared to parcels located far away (> 5 miles) from a lake (see the "ORSH model" and "OR model" sub-columns of the last column of Table 7.3 under the row "Distance to a lake ≤ 5 miles"). On the other hand, parcels located near an airfield and within Austin city (at least in the context of the area used in the current demonstration exercise) are the least likely to be intensely developed. Similarly, parcels located far away from IH-35 (> 9 miles from IH-35) and parcels within Austin city limits are the most likely to be in an undeveloped state

Table 7.3 Elasticity Effects of Variables on the Land Use Development Intensity Level (Standard error)[49]

Variable	Undeveloped land			Less-intensely developed land			Medium-intensely developed land			Most-intensely developed land		
	ORSH model	OR model	p-value for difference	ORSH model	OR model	p-value for difference	ORSH model	OR model	p-value for difference	ORSH model	OR model	p-value for difference
Distance to a park ≤ 2 miles (base: park > 2 miles)	-20.96 (4.96)	-4.89 (12.55)	-	-4.78 (4.44)	-0.17 (0.61)	-	30.64 (16.85)	3.61 (8.92)	0.157	56.65 (19.40)	3.97 (14.55)	0.030
Distance to a lake ≤ 5 miles (base: lake > 5 miles)	-79.10 (6.12)	-67.37 (6.12)	0.176	-79.25 (3.86)	-7.87 (3.15)	0.000	-25.47 (68.61)	80.66 (10.49)	0.127	167.64 (15.73)	148.36 (21.05)	-
Distance to a school ≤ 2 miles (base: school > 2 miles)	-56.05 (26.43)	-10.44 (10.54)	0.110	3.29 (3.90)	0.00 (0.46)	-	10.38 (2.41)	6.25 (7.01)	-	8.05 (3.17)	6.9 (11.33)	-
Downtown ≤ 9 miles (base: Downtown > 9 miles)	68.54 (18.17)	41.32 (10.72)	0.198	3.10 (11.67)	0.29 (1.39)	-	-58.03 (13.38)	-29.85 (8.14)	0.073	-67.17 (7.57)	-42.39 (9.70)	0.045
IH-35 ≤ 9 miles (base: IH-35 > 9 miles)	-173.26 (97.81)	-4.62 (9.40)	0.087	7.96 (7.72)	-0.07 (0.4)	-	31.52 (5.47)	3.31 (7.08)	0.002	40.62 (3.62)	3.86 (10.72)	0.001
Airfield ≤ 1 miles (base: Airfield > 1 miles)	57.73 (7.42)	30.95 (6.50)	0.007	8.06 (10.65)	1.11 (1.05)	-	-77.99 (38.30)	-29.17 (7.14)	-	-108.92 (40.98)	-38.32 (9.83)	0.094
Parcel is located in Austin city (base: parcel is located outside Austin city)	216.30 (41.12)	129.93 (9.54)	0.041	-4.25 (26.77)	0.69 (4.50)	-	-105.35 (12.63)	-96.79 (6.57)	-	-108.09 (6.76)	-121.62 (6.46)	0.148
Elevation ≤ 1000 feet above mean sea level (base: elevation > 1000 feet)	120.97 (49.74)	32.52 (6.62)	0.079	-2.06 (13.14)	0.53 (1.09)	-	-54.95 (9.15)	-24.19 (4.35)	0.003	-58.86 (4.41)	-36.00 (5.92)	0.002

[49] The standard errors of the elasticity effects are computed using 100 bootstrap draws. A “-” entry in the table indicates that the difference is not statistically significant even at the 0.20 level of significance.

(see the first two numeric sub-columns in Table 7.3). Second, the elasticity effects of both the ORSH and the OR models are in the same direction. However, a visual comparison of the results indicates that the elasticity effects predicted by the ORSH model are higher than the OR model prediction (the only exception is the effect of "Parcel is located in Austin" variable on the most-intensely developed land use category). The higher magnitudes from the ORSH model reflect the spatial multiplier effect caused by spatial dependence. Specifically, a change in a variable relevant to one land owner (that has an impact on the LUDR perception of the land owner) also affects the LUDR perceptions of land owners of proximally located parcels, which then have a "circular" and reinforcing influence back on the LUDR perception of the land owner (this spatial multiplier effect is captured by the $\mathbf{S}$ matrix in Equation (6.3)). In contrast, the OR model ignores the presence of the "spillover" phenomenon and assumes away any spatial interaction effects among land owners. Finally, the entries in the p-value columns for each ordinal land use intensity category indicate that many of the differences in elasticity effects between the ORSH and OR models are statistically significant at the 0.1 level or lower, clearly underscoring the importance of accommodating spatial dynamics and spatial heterogeneity in the current empirical context.

Overall, these results reinforce the findings from the simulation exercise in Section 6.3 (Chapter 6) and indicate the potentially substantial biases in elasticity effects if spatial dependence and/or heterogeneity are ignored.

7.5 Summary

In the previous chapters (Chapter 4 through Chapter 6) we have demonstrated that the CML approach can be used to develop behaviorally rich models that are also statistically superior. In this chapter we applied the models in a number of demonstration exercises to evaluate the effect of changes in a number of explanatory variables. For each exercise, the model predictions were also compared with the naïve model predictions. The results suggest that ignoring the multidimensional nature of the models developed here can result in inaccurate and bias prediction/policy evaluation.

Chapter 8

Synopsis and Directions for Future Research

8.1 Introduction

The research in the field of travel demand modeling is driven by the need to understand individuals' behavior in the context of travel-related decisions as accurately as possible. In this context, the activity-based approach to modeling travel demand has received substantial attention in the past decade, both in the research arena as well as in practice. At the same time, recent efforts have been focused on more fully realizing the potential of activity-based models by explicitly recognizing the multi-dimensional nature of activity-travel decisions. For instance, while some earlier activity-based models assumed that individuals' non-mandatory activity participation decisions (such as eating out, going to theater) are made in isolation, more recent activity-based models recognize that, in general, individuals' non-mandatory activity participation decisions are inter-related (within the individual) and also based on group decisions made at the household-level (across individuals in the household). However, as more behavioral elements/dimensions are added, the dimensionality of the model systems tends to explode, making the estimation of such models all but infeasible using traditional inference methods. As a result, analysts and practitioners often trade-off between recognizing attributes that will make a model behaviorally more representative (from a theoretical viewpoint) and being able to estimate/implement a model (from a practical viewpoint).

An alternative approach to deal with the estimation complications arising from multi-dimensional choice situations is the technique of composite marginal likelihood (CML). This is an estimation technique that is gaining substantial attention in the statistics field, though there has been relatively little coverage of this method in transportation and other fields. The CML method, which belongs to the more general class of composite likelihood function approaches, is based on forming a surrogate likelihood function that compounds much easier-to-compute, lower-dimensional, marginal likelihoods. The CML approach has the advantage of reproducibility of results and can be easily implemented using simple optimization software for likelihood estimation. Under the usual regularity assumptions, the CML estimator is consistent, unbiased, and asymptotically normally distributed.

The discussion above provides a brief overview of the background that motivated the research undertaken in the current dissertation. Specifically, the overarching goal of the current research work was to demonstrate applicability of the CML approach in the area of activity-travel demand modeling and to highlight the benefits of behaviorally rich choice structures that can be estimated using the CML approach. The goal of the dissertation is achieved in three steps. Each of these steps makes a distinct research contribution, as discussed in the next section (Section 8.2). Then, Section 8.3 concludes the dissertation by identifying limitations of the current research and highlighting areas for future research.

8.2 Research Contributions

8.2.1 Evaluating Performance of the CML Approach

As indicated earlier, the CML approach is a relatively new estimation technique. Accordingly, before adopting the approach to model individuals' activity-travel behavior, we sought to first assess the effectiveness of the CML approach. Specifically, we evaluated the performance of the CML approach in terms of its ability to recover the parameters of an ordered-response model system. The evaluation exercises were undertaken using two types of simulated data: aspatial cross-sectional data and spatial panel data. For cross-sectional data, both low and high error correlation structures were considered. For panel data, low and high spatial and temporal autoregressive parameters and their combinations were considered. Overall, the simulation results demonstrate the ability of the CML approach to recover the parameters very well in an ordered-response choice model context.

In this dissertation, we also empirically examined the efficiency of the CML estimator. Specifically, the CML estimator (theoretically speaking) loses some efficiency relative to traditional maximum likelihood estimation, though some earlier empirical investigations suggest that such efficiency loss is negligible. In the current research, this issue was investigated further by comparing the performance of the CML approach with the maximum-simulated likelihood (MSL) approach in multivariate ordered-response situations. The ability of the two approaches to recover model parameters in simulated cross-sectional data sets was examined, as was the efficiency of estimated parameters and computational cost. The results indicate that the CML recovers parameters as well as the MSL estimation approach in the simulation contexts used in the current analysis, while also doing so at a substantially reduced computational cost. Further, any reduction in the efficiency of the CML approach relative to the MSL approach is in the range of non-existent to small.

In summary, when taken together with its conceptual and implementation simplicity, the CML approach appears to be a promising approach for the estimation of not only the multivariate ordered-response model considered here, but also for other otherwise analytically-intractable econometric models.

8.2.2 Developing Multidimensional Choice Models Using the CML Approach

In the dissertation, a series of econometric models was developed that are behaviorally rich but have a complex dependence structure, and are generally considered impractical and/or infeasible to be estimated by traditional estimation approaches. This dissertation demonstrates that such models can indeed be estimated using the CML technique. The salient features of each of these models and important empirical findings from the studies are discussed in turn in the next three sections (Section 8.2.2.1 to Section 8.2.2.3).

8.2.2.1 A Multivariate Ordered-Response Model With Flexible Error Structure

A multivariate ordered-response model was developed to examine the interactions in non-work activity episode decisions across household and non-household members at the level of activity generation. The six activity purpose categories considered in the study were: (1) family care, (2) maintenance shopping, (3) non-maintenance shopping, (4) meals, (5) physically active recreation, and (6) physically inactive recreation. The companionship arrangement for episodes was considered in five categories: (1) alone, (2) only family, (3) only relatives, (4) only friends, and (5) mixed company. The total number of activity purpose-companionship type categories is 30, and the model system developed here jointly considers the number of episodes in each of these 30 categories.

A salient feature of this model system is that the dependence between the number of episodes of different purpose-companionship types due to both observed exogenous variables as well as unobserved factors can be accommodated without any difficulty. The empirical analysis in the study used data drawn from the 2007 American Time Use Survey (ATUS) and provided important insights into the determinants of adults' weekday activity episode generation behavior. For instance, the results indicate the presence of distinct gender effects in activity type participation and accompaniment, with women being more responsible for family care and shopping activities, and men being more likely to undertake active and inactive leisure activities either alone or with friends. Further, there are also clear age-related effects. Individuals below the age of 40 years are the least likely to participate in activity episodes alone and most likely to participate in episodes with mixed company, suggesting a combination of the family orientation and larger social networks of younger individuals. Race, education level, employment and student status, household structure and presence of children, household income, the day of week, and season of the year also have important effects on adults' weekday activity episodes by purpose and the social context of participation.

Overall, the results from this model underscored the substantial linkages in the activity episode generation of adults based on activity purpose and accompaniment type. The extent of this linkage varies by individual demographics, household demographics, day of the week, and season of the year. The results also highlighted the need to accommodate complementarity and substitution effects in inter-individual interactions and in activity episode participation decision.

8.2.2.2 A Joint Model of Walking and Bicycling Activity Duration

In this study, the time allocated by individuals in walking and bicycling activity over a period of one week was analyzed jointly using a proportional hazard model specification. An important aspect of this joint model system is that the model is capable of incorporating grouped duration responses (which is commonly observed in activity-travel surveys and is a result of individuals rounding off activity durations when reporting their time-use patterns). Another key feature of the model structure developed here is that it recognizes the presence of unobserved heterogeneity in walking and bicycling activity participation. Specifically, the model structure accommodates variations in the activity durations for different activity types based on unobserved factors that are specific to the individual, the household, the social cluster/peer group to which the individual is part of, and the spatial cluster to which the individual belongs. For accurate prediction of activity duration and evaluation of policy actions, it is important to consider the effects of unobserved factors that contribute to heterogeneity in walking and bicycling activity durations (or non-motorized transport mode use behavior) at multiple levels. The model specification also generates a rich covariance pattern structure among the hazard functions for participation in different activities for the same individual as well as between different individuals. Such a model specification would require the evaluation of a 1764-dimensional integral in the traditional maximum likelihood inference approach, which is next to infeasible. Also, the specification, because of the hazard duration structure, leads to the mixing of normal and extreme-value error terms. We dealt with the first complication (1764 dimensions) by resorting to the estimation technique of composite marginal likelihood. We took care of the second complication (mixing of error terms) by removing the non-normality of the type I extreme value error term and replacing it with a weighted mixture of normally distributed variables (*i.e.*, we used the normal scale mixture (NSM) representation of the extreme value distribution).

The model system was applied to a survey sample drawn from the California add-on of the United States National Household Travel Survey (NHTS) conducted in 2009. In addition to individual- and household-level socio-demographic information, the

California-specific NHTS 2009 data set contained detailed attitudinal information on walking and cycling activities, including factors that were likely to influence individual's walking and bicycling duration. This made the NHTS 2009 data set particularly appropriate for the current study. The model results show that individual- and household-demographic and socio-economic variables impact individuals' walking and bicycling activity durations. Also, there are numerous attitudinal factors and perceptions that affect these durations. For example, busy lifestyles, perceptions of poor walking environment and inadequate bicycling infrastructure, and concerns about safety adversely impact the amount of walking and bicycling undertaken by individuals. These findings are consistent with expectations and point to the need for professionals and policymakers to consider neighborhood designs, land use configurations, and infrastructure investments that alleviate the concerns and enhance perceptions of bicycling and walking convenience. In addition, the model results suggest that there are significant unobserved individual-level, social group, and spatial proximity effects that contribute to heterogeneity in walking and bicycling activity duration. These effects were significant even after controlling for observed variables. The unobserved effects were found to have a differential impact on walking and bicycling activity durations, thus suggesting the need to treat walking and bicycling separately and to model them in a joint framework.

8.2.2.3 A Spatial Panel Ordered-Response Probit Model With Temporal Autoregressive Error Terms

This study proposed and estimated a spatial panel ordered-response probit model with temporal autoregressive error terms to analyze changes in urban land development intensity levels over time. Such a model structure offers several salient features. First, the model maintains a close linkage between the land owner's decision (represented by the land use development return (LUDR) perceptions of the land owners, this is a latent variable that cannot be observed by the analyst) and the land development intensity level (observed by the analyst). It is important to maintain such a linkage since the decision to change (or to maintain) the current land development intensity level is actually made by the land owners. Second, the model specification accommodates spatial interactions between land owners that leads to spatial spillover effects. Such a specification recognizes that spatial dependence is caused by didactic interactions between decision-making agents (as opposed to considering spatial dependence only in the error terms, which is tantamount to viewing spatial dependence as “nuisance” dependence). Third, the model structure incorporates spatial heterogeneity as well as spatial heteroscedasticity by allowing the sensitivity to exogenous variables to vary across land owners. Finally, the model accommodates time-invariant and time-varying temporal dependence. Time-invariant temporal dependence represents the effects of landowner-specific unobserved factors that do not change over time (such as individual experiences, risk-taking behavior, and vegetation conservation values). Time-varying temporal dependence captures the effects of landowner-specific unobserved factors that fade away over time (such as the effects of recent experiences and events).

Before undertaking an empirical analysis, we evaluated the model in a simulation design to examine the effects of ignoring spatial dependence and spatial heterogeneity when both are actually present (this is in addition to the simulation exercise that was undertaken to evaluate the ability of the CML approach to recover model parameters in the spatial panel data context, as discussed in Section 8.2.1). The results demonstrate that ignoring spatial dependency and spatial heterogeneity when both are actually present will lead to bias in parameter estimation. An interesting observation from our simulation study is that ignoring spatial heterogeneity is of much more serious consequence than ignoring spatial lag dynamics.

The proposed model system was applied to examine urban land development intensity level using parcel-level data from Austin, Texas area for the years 2000, 2003, 2006, and 2008. In the current analysis, a four category ordinal system was used to define the intensity level of land development: (1) undeveloped land (open space, vacant parcel, *etc.*), (2) less-intensely developed land (residential parcels with single-family detached or two-family attached home), (3) medium-intensely developed land (includes all other types of residential parcels), and (4) most-intensely developed land (includes office, commercial, industrial parcels, *etc.*). The final data set comprised of 783 parcels from each time period. The model results suggest that closeness to natural and other amenities (such as park, lake, school, and urban center), distance to major roadways, average elevation of the parcel, and whether or not the parcel is located in Austin city have significant effect on the LUDR perceptions of the land owners. The results also highlight the importance to consider spatial "spillover" effects, spatial heterogeneity, and temporal effects in the study of land development intensity level to obtain consistent parameter estimates and policy evaluation.

8.2.3 Demonstrating Applications of the Multidimensional Choice Models

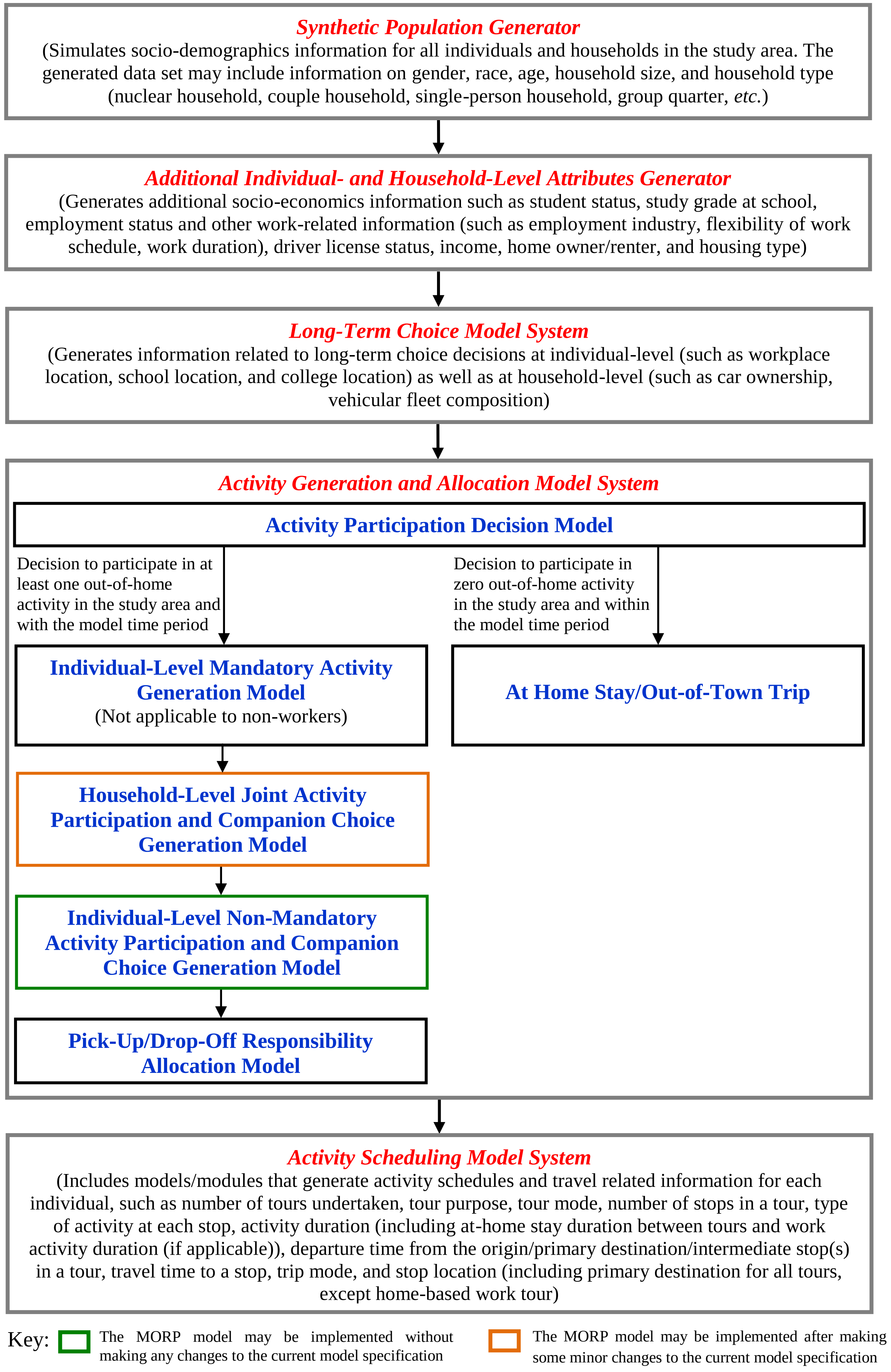
The multidimensional econometric models discussed in the previous sections were applied to examine their benefits vis-à-vis extant and more naïve methods. These exercises:

- Highlighted practical/real life application(s) of the models developed in the current dissertation,
- Underscored the biased results that could be obtained if the multidimensional nature of the models developed here are ignored,
- Provided a comparison between the performance of the estimated multidimensional choice models and the naïve models, and
- Quantified the effects of accommodating behavioral elements in the model specification.

8.3 Limitations of the Current Research and Directions for Future Work

The current dissertation makes several research contributions, as discussed in the previous section. However, there are, of course, limitations of the current research work that need to be explored in the future. In addition, there are research areas which may not necessarily fall under the category of limitations of the current research effort, but may be viewed as expanding the scope of the current work. A few of these research ideas/thoughts are discussed below.

1) The exercises undertaken here to evaluate the ability of the CML approach to recover model parameters are far from exhaustive. Future research efforts in this direction may include examining the ability of the CML approach to recover parameters in the context of additional types of data such as cross-sectional data with heteroscedasticity, (continuous and grouped) duration data, time series data with different levels of dependence. Also, we leave additional comparisons of the CML approach with the MSL approach for high dimensional model contexts and alternative covariance patterns as directions for further research.
2) The multidimensional choice models developed here were employed to undertake empirical analyses using data that were area-specific. Hence, transferring the current model results to other geographic areas should be done with extreme caution. Alternatively, the model results may be re-calibrated to custom fit new study area(s). Such exercises are likely to provide important insights on decision agents' behaviors.
3) The multivariate ordered-response probit model (the MORP model) developed in Chapter 4 may be embedded in an activity-based modeling framework to generate individuals' non-mandatory activity episodes by purpose and companion choice jointly (See Figure 8.1). The model system can be extended along several dimensions. For instance, the model can be used to generate all activity episodes (*i.e.*, both mandatory and non-mandatory activity episodes) of all individuals in a household (*i.e.*, activity episodes of both children and adults) without making any substantial changes to the current model specification. Also, currently the model accommodates three activity-related decisions that individuals are likely to make jointly: "what" (*i.e.*, the type of activity to participate in), "how many times" (*i.e.*, the frequency of participation), and "with whom" (*i.e.*, who to participate the activity with). In reality though, individuals are likely to make a number of other activity-travel-related decisions also jointly, such as "when", "where", and "what mode". The current model framework may be extended/revised to incorporate these additional dimensions.



Synthetic Population Generator

(Simulates socio-demographics information for all individuals and households in the study area. The generated data set may include information on gender, race, age, household size, and household type (nuclear household, couple household, single-person household, group quarter, *etc.*)

Additional Individual- and Household-Level Attributes Generator

(Generates additional socio-economics information such as student status, study grade at school, employment status and other work-related information (such as employment industry, flexibility of work schedule, work duration), driver license status, income, home owner/renter, and housing type)

(Generates information related to long-term choice decisions at individual-level (such as workplace location, school location, and college location) as well as at household-level (such as car ownership, vehicular fleet composition)

Activity Generation and Allocation Model System
Activity Participation Decision Model
Decision to participate in at least one out-of-home activity in the study area and with the model time period
Decision to participate in zero out-of-home activity in the study area and within the model time period
Individual-Level Mandatory Activity Generation Model
(Not applicable to non-workers)
At Home Stay/Out-of-Town Trip
Household-Level Joint Activity Participation and Companion Choice Generation Model
Individual-Level Non-Mandatory Activity Participation and Companion Choice Generation Model
Pick-Up/Drop-Off Responsibility Allocation Model
Activity Scheduling Model System

(Includes models/modules that generate activity schedules and travel related information for each individual, such as number of tours undertaken, tour purpose, tour mode, number of stops in a tour, type of activity at each stop, activity duration (including at-home stay duration between tours and work activity duration (if applicable)), departure time from the origin/primary destination/intermediate stop(s) in a tour, travel time to a stop, trip mode, and stop location (including primary destination for all tours, except home-based work tour)

Key: 🟩 The MORP model may be implemented without making any changes to the current model specification 🟧 The MORP model may be implemented after making some minor changes to the current model specification

Figure 8.1: Schematic Representation of an Activity-Based Modeling Framework

4) In the 2009 NHTS data, residential location information was available only at the Census tract level. As a result, the spatial unit of analysis used to define spatial clustering was the traffic analysis zone (TAZ). This is a rather aggregate spatial representation of clustering, and a finer resolution for spatial clustering needs to be considered. Also, due to data limitations, we were unable to estimate the joint model of walking and bicycling activity durations by purpose.

5) The spatial panel ordered-response probit model with temporal autoregressive error terms (proposed in Chapter 6) maintains a close link between the landowner and land development intensity level. However, we were not able to incorporate land owners' information in our model since the data did not provide such information. Also, we studied didactic interactions between land owners in the context of a "continuous space" study area (i.e., the study area was not sever by any natural or man-made barrier). It will be interesting to analyze interactions between land owners when the study area is segmented.

REFERENCES

Agyemang-Duah, K., and F.L. Hall (1997) Spatial transferability of an ordered response model of trip generation. *Transportation Research Part A*, 31(5), 389-402.

Agyemang-Duah, K., F.L. Hall, and W.P. Anderson (1995) Trip generation for shopping travel. *Transportation Research Record*, 1493, 12-20.

Agrawal, A.W., and P. Schimek (2007) Extent and correlates of walking in the USA. *Transportation Research Part D*, 12(8), 548-563.

Alamá-Sabater, L., A. Artal-Tur, and J.M. Navarro-Azorín (2011) Industrial location, spatial discrete choice models and the need to account for neighbourhood effects. *The Annals of Regional Science*, Special Issue. Available at: http://www.springerlink.com/content/84976h7q10t6h566/fulltext.pdf.

Anas, A., R. Arnott, and K.A. Small (1998) Urban spatial structure. *Journal of Economic Literature*, 34(3), 1426-1464.

Andersen, L.B., P. Schnohr, M. Schroll, and H.O. Hein (2000) All-cause mortality associated with physical activity during leisure time, work, sports, and cycling to work. *Archives of Internal Medicine*, 160(11), 1621-1628.

Anselin, L. (2003) Spatial externalities, spatial multipliers and spatial econometrics. *International Regional Science Review*, 26(2), 153-166.

Anselin, L. (2010) Thirty years of spatial econometrics. *Papers in Regional Science*, 89(1), 3-25.

Anselin, L., J.L. Gallo, and H. Jayet (2008) Spatial panel econometrics. In *The Econometrics of Panel Data: Fundamentals and Recent Developments in Theory and Practice*, L. Mátyás, and P. Sevestre (Eds.), Springer, New York, 46, 625-660.

Apanasovich, T.V., D. Ruppert, J.R. Lupton, N. Popovic, N.D. Turner, R.S. Chapkin, and R.J. Carroll (2008) Aberrant crypt foci and semiparametric modelling of correlated binary data. *Biometrics*, 64(2), 490-500.

Arbia, G., and H. Kelejian (2010) Advances in spatial econometrics. *Regional Science and Urban Economics*, 40(5), 253-254.

Arentze, T.A., and H.J.P. Timmermans (2004) A learning-based transportation oriented simulation system. *Transportation Research Part B*, 38(7), 613-633.

Arentze, T. A., and H.J.P. Timmermans (2008) Social networks, social interactions, and activity-travel behavior: A framework for microsimulation. *Environment and Planning B*, 35(6), 1012-1027.

ATUS (2008) American time use survey user’s guide understanding ATUS 2003 to 2007. Available at: http://www.bls.gov/tus/atususersguide.pdf.

Axhausen, K.W. (2005) Social networks and travel: Some hypotheses. In *Social Aspects of Sustainable Transport: Transatlantic Perspectives*, K. Donaghy, S. Poppelreuter, and G. Rudinger (Eds.), Ashgate Publishing, Aldershot, England, Chapter 7, 90-108.

Badland, H., and G. Schofield (2005) Transport, urban design, and physical activity: An evidence-based update. *Transportation Research Part D*, 10(3), 177-196.

Balia, S., and A.M. Jones (2008) Mortality, lifestyle and socio-economic status. *Journal of Health Economics*, 27(1), 1-26.

Bassett, D.R. Jr., J. Pucher, R. Buehler, D.L. Thompson, and S.E. Crouter (2008) Walking, cycling, and obesity rates in Europe, North America, and Australia. *Journal of Physical Activity and Health*, 5, 795-814.

Bellio, R., and C. Varin (2005) A pairwise likelihood approach to generalized linear models with crossed random effects. *Statistical Modelling*, 5(3), 217-227.

Beron, K.J., J.C. Murdoch, and W.P.M. Vijverberg (2003) Why cooperate? Public goods, economic power, and the Montreal Protocol. *Review of Economics and Statistics*, 85(2), 86-97.

Beron, K. J., and W.P.M. Vijverberg (2004) Probit in a spatial context: A Monte Carlo analysis. In *Advances in Spatial Econometrics: Methodology, Tools and Applications*, L. Anselin, R.J.G.M. Florax, and S.J. Rey (Eds.), Springer, New York, 169-195.

Bhat, C.R. (1996) A generalized multiple durations proportional hazard model with an application to activity behavior during the work-to-home commute. *Transportation Research Part B*, 30(6), 465-480.

Bhat, C.R. (1999) An analysis of evening commute stop-making behavior using repeated choice observations from a multi-day survey. *Transportation Research Part B*, 33(7), 495-510.

Bhat, C. R. (2000) A multi-level cross-classified model for discrete response variables. *Transportation Research Part B*, 34(7), 567-582.

Bhat, C.R. (2001) Quasi-random maximum simulated likelihood estimation of the mixed multinomial logit model. *Transportation Research Part B*, 35(7), 677-693.

Bhat, C.R. (2003) Simulation estimation of mixed discrete choice models using randomized and scrambled Halton sequences. *Transportation Research Part B*, 37(9), 837-855.

Bhat, C.R. (2008) The multiple discrete-continuous extreme value (MDCEV) model: Role of utility function parameters, identification considerations, and model extensions. *Transportation Research Part B*, 42(3), 274-303.

Bhat, C.R. (2011) The maximum approximate composite marginal likelihood (MACML) estimation of multinomial probit-based unordered response choice models. *Transportation Research Part B*, forthcoming.

Bhat, C.R., and N. Eluru (2009) A copula-based approach to accommodate residential self-selection effects in travel behavior modeling. *Transportation Research Part B*, 43(7), 749-765.

Bhat, C.R., and R. Gossen (2004) A mixed multinomial logit model analysis of weekend recreational episode type choice. *Transportation Research Part B*, 38(9), 767-787.

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

Bhat, C.R., and F.S. Koppelman (1999) Activity-based modeling of travel demand. In *The Handbook of Transportation Science*, Hall, R.W. (Ed.), Kluwer Academic Publishers, Norwell, Massachusetts, 35-61.

Bhat, C.R., and R.M. Pendyala (2005) Modeling intra-household interactions and group decision-making. *Transportation*, 32(5), 443-448.

Bhat, C.R., and A.R. Pinjari (2008) Duration modeling. In *Handbook of Transport Modelling*, 2nd edition, D.A. Hensher and K.J. Button (Eds.), Elsevier Science, 105-132.

Bhat, C.R., and I.N. Sener (2009) A copula-based closed-form binary logit choice model for accommodating spatial correlation across observational units. *Journal of Geographical Systems*, 11(3), 243-272.

Bhat, C.R., I.N. Sener, and N. Eluru (2010b) A flexible spatially dependent discrete choice model: Formulation and application to teenagers' weekday recreational activity participation. *Transportation Research Part B*, 44(8-9), 903-921.

Bhat, C.R., and S. Srinivasan (2005) A multidimensional mixed ordered-response model for analyzing weekend activity participation. *Transportation Research Part B*, 39(3), 255-278.

Bhat, C.R., C. Varin, and N. Ferdous (2010a) A comparison of the maximum simulated likelihood and composite marginal likelihood estimation approaches in the context of the multivariate ordered response model. In *Advances in Econometrics: Maximum Simulated Likelihood Methods and Applications*, W. Greene, and R.C. Hill (Eds.), Emerald Group Publishing Limited, 26, 65-106.

Böhning, D., and W. Seidel (2003) Recent developments in mixture models. *Computational Statistics & Data Analysis*, 41(3-4), 349-357.

Bradley, M., and P. Vovsha (2005) A model for joint choice of daily activity pattern types of household members. *Transportation*, 32(5), 545-571.

Bradlow, E.T., B. Bronnenberg, G.J. Russell, N. Arora, D.R. Bell, S.D. Duvvuri, F.T. Hofstede, C. Sismeiro, R. Thomadsen, and S. Yang (2005) Spatial models in marketing. *Marketing Letters*, 16(3-4), 267-278.

Bricka, S., and C.R. Bhat (2006) A comparative analysis of GPS-based and travel survey-based data. *Transportation Research Record*, 1972, 9-20.

Bureau of Labor Statistics (2007) Employment projections 2006-2016. Available at: http://www.bls.gov/news.release/ecopro.nr0.htm.

CACOG (2010) Geospatial data. Available at: http://www.capcog.org/information-clearinghouse/geospatial-data/#county-boundaries.

Cao, X., P.L. Mokhtarian, and S.L. Handy (2009) The relationship between the built environment and nonwork travel: A case study of Northern California. *Transportation Research Part A*, 43(5), 548-559.

Caragea, P.C., and R.L. Smith (2007) Asymptotic properties of computationally efficient alternative estimators for a class of multivariate normal models. *Journal of Multivariate Analysis*, 98(7), 1417- 1440.

Carrasco, J-A., B. Hogan, B. Wellman, and E. J. Miller (2008) Collecting social network data to study social activity-travel behavior: An egocentric approach. *Environment and Planning B*, 35, 961-980.

Carrasco, J-A., and E.J. Miller (2009) The social dimension in action: A multilevel, personal networks model of social activity frequency between individuals. *Transportation Research Part A*, 43(1), 90-104.

Carrión-Flores, C.E., A. Flores-Lagunes, and L. Guci (2009) Land use change: A spatial multinomial choice analysis. Paper prepared for presentation at the III World Conference of Spatial Econometrics, Barcelona, Spain.

Carrión-Flores, C., and E.G. Irwin (2004) Determinants of residential land-use conversion and sprawl at the rural-urban fringe. *American Journal of Agricultural Economics*, 86(4), 889-904.

Chakir, R., and O. Parent (2009) Determinants of land use changes: A spatial multinomial probit approach. *Papers in Regional Science*, 88(2), 327-344.

Chen, M.-H., and D.K. Dey (2000) Bayesian analysis for correlated ordinal data models. In *Generalized Linear Models: A Bayesian Perspective*, D.K. Dey, S.K. Gosh, and B.K. Mallick (Eds.), Marcel Dekker, New York.

Chib, S., and R. Winkelmann (2001) Markov Chain Monte Carlo analysis of correlated count data. *Journal of Business and Economic Statistics*, 19(4), 428-435.

Choy, S.T.B., and J.S.K. Chan (2008) Scale mixtures distributions in statistical modeling. *Australian & New Zealand Journal of Statistics*, 50(2), 135-146.

City of Austin (2011) City of Austin GIS data sets. Available at: ftp://ftp.ci.austin.tx.us/GIS-Data/Regional/coa_gis.html#base_map.

Cooper, A.R., N. Wedderkopp, H. Wang, L.B. Andersen, K. Froberg, and A.S. Page (2006) Active travel to school and cardiovascular fitness in Danish children and adolescents. *Medicine & Science in Sports & Exercise*, 38(10), 1724-1731.

Cox, D., and N. Reid (2004) A note on pseudolikelihood constructed from marginal densities. *Biometrika*, 91(3), 729-737.

Craig, P. (2008) A new reconstruction of multivariate normal orthant probabilities. *Journal of the Royal Statistical Society: Series B*, 70(1), 227-243.

de Leon, A.R. (2005) Pairwise likelihood approach to grouped continuous model and its extension. *Statistics & Probability Letters*, 75(1), 49-57.

de Nazelle, A., and D.A. Rodríguez (2009) Tradeoffs in incremental changes towards pedestrian-friendly environments: Physical activity and pollution exposure. *Transportation Research Part D*, 14(4), 255-263.

Doherty, S.T., and K.W. Axhausen (1999) The development of a unified modeling framework for the household activity-travel scheduling process. In *Traffic and Mobility: Simulation-Economics-Environment*, W. Brilon, F. Huber, M. Schreckengerg, and H. Wallentowitz (Eds.), 35-56.

Dugundji, E., and J. Walker (2005) Discrete choice with social and spatial network interdependencies: An empirical example using mixed generalized extreme value models with field and panel effects. *Transportation Research Record*, 1921, 70-78.

Egan, K., and J. Herriges (2006) Multivariate count data regression models with individual panel data from an on-site sample. *Journal of Environmental Economics and Management*, 52(2), 567-581.

Elhorst, J.P. (2009) Spatial panel data models. In *Handbook of Applied Spatial Analysis*, M.M. Fischer, and A. Getis (Eds.), Springer, Berlin.

Engle, R.F., N. Shephard, and K. Sheppard (2007) Fitting and testing vast dimensional time-varying covariance models. Finance Working Papers, FIN-07-046, Stern School of Business, New York University.

Engler, D.A., M. Mohapatra, D.N. Louis, and R.A. Betensky (2006) A pseudolikelihood approach for simultaneous analysis of array comparative genomic hybridizations. *Biostatistics*, 7(3), 399-421.

Espey, M., and K. Owusu-Edusei (2001) Neighborhood parks and residential property values in Greenville, South Carolina. *Journal of Agricultural and Applied Economics*, 33(3), 487-492.

Facchini, F., and A. François (2010) A Border effect in political mobilization? Territorial dependence and electoral turnout in national election. Available at: http://ses.telecom-paristech.fr/francois/publications/WP/Territorial%20dependence.pdf.

Ferdous, N., N. Eluru, C.R. Bhat, and I. Meloni (2010) A multivariate ordered-response model system for adults' weekday activity episode generation by activity purpose and social context. *Transportation Research Part B*, 44(8-9), 922-943.

Flegal, K.M., M.D. Carroll, C.L. Ogden, and L.R. Curtin (2010) Prevalence and trends in obesity among US adults, 1999-2008. *The Journal of the American Medical Association*, 303(3), 235-241.

Fleming, M. (2004) Techniques for estimating spatially dependent discrete choice models. In *Advances in Spatial Econometrics: Methodology, Tools and Applications*, L. Anselin, R.J.G.M. Florax, and S.J. Rey (Eds.), Springer, New York, 145-168.

Forsyth, A., J.M. Oakes, B. Lee, and K.H. Schmitz (2009) The built environment, walking, and physical activity: Is the environment more important to some people than others? *Transportation Research Part D*, 14(1), 42-49.

Frank, L.D., T.L. Schmid, J.F. Sallis, J. Chapman, and B.E. Saelens (2005) Linking objectively measured physical activity with objectively measured urban form: Findings from SMARTRAQ. *American Journal of Preventive Medicine,* 28(2), 117-125.

Franzese, R.J., and J.C. Hays (2008) Empirical models of spatial interdependence. In *Oxford Handbook of Political Methodology*, J. Box-Steffensmeier, H. Brady, and D. Collier (Eds.), Oxford University Press, 570-604.

Franzese, R.J., J.C. Hays, and L.M. Schaffer (2010) Spatial, temporal, and spatiotemporal autoregressive probit models of binary outcomes: Estimation, interpretation, and presentation. APSA 2010 Annual Meeting Paper. Available at SSRN: http://ssrn.com/abstract=1643867.

Frühwirth-Schnatter, S., and H. Wagner (2005) Data augmentation and Gibbs sampling for regression models of small counts. Research Report IFAS 2004-04, available at: http://www.ifas.jku.at/.

Frusti, T., C.R. Bhat, and K.W. Axhausen (2003) An exploratory analysis of fixed commitments in individual activity-travel patterns. *Transportation Research Record*, 1807, 101-108.

Garrard, J., S. Crawford, and N. Hakman (2006) Revolutions for women: Increasing women's participation in cycling for recreation and transport, summary of key findings. Available at: http://www.australianwomensport.com.au/images/Reports/victorian%20research/Revolutions%20key%20findings%20A4.pdf.

Genius, M., and E. Strazzera (2008) Applying the copula approach to sample selection modeling. *Applied Economics*, 40(11), 1443-1455.

Genz, A. (1992) Numerical computation of multivariate normal probabilities. *Journal of Computational and Graphical Statistics*, 1(2), 141-149.

Genz, A. (2003) Fully symmetric interpolatory rules for multiple integrals over hyper-spherical surfaces. *Journal of Computational and Applied Mathematics*, 157(1), 187-195.

Genz, A., and F. Bretz (1999) Numerical computation of multivariate t-probabilities with application to power calculation of multiple contrasts. *Journal of Statistical Computation and Simulation*, 63(4), 361-378.

Genz, A., and F. Bretz (2002) Comparison of methods for the computation of multivariate t probabilities. *Journal of Computational and Graphical Statistics*, 11(4), 950-971.

Geoghegan, J. (2002) The value of open spaces in residential land use. *Land Use Policy*, 19(1), 91-98.

Gérard, M., H. Jayet, and S. Paty (2010) Tax interactions among Belgian municipalities: Do interregional differences matter? *Regional Science and Urban Economics*, 40(5), 336-342.

Geweke, J. (1991) Efficient simulation from the multivariate normal and student-t distributions subject to linear constraints. *Computer Science and Statistics: Proceedings of the Twenty Third Symposium on the Interface*, 571-578, Foundation of North America Inc., Fairfax.

Gliebe, J.P., and F.S. Koppelman (2002) A model of joint activity participation between household members. *Transportation*, 29(1), 49-72.

Godambe, V. (1960) An optimum property of regular maximum likelihood equation. *The Annals of Mathematical Statistics*, 31(4), 1208-1211.

Golob, T.F. (1990) The dynamics of household travel time expenditures and car ownership decisions. *Transportation Research Part A*, 24(6), 443-463.

Golob, T.F., and L. van Wissen (1989) A joint household travel distance generation and car ownership model. *Transportation Research Part B*, 23(6), 471-491.

Gorsuch R.L. (1983) *Factor analysis*. Second edition, Lawrence Erlbaum associates publishers.

Goulias, K.G., and T.G. Kim (2001) Multi-level analysis of activity and travel patterns: Accounting for person- and household- specific observed and unobserved effects simultaneously. *Transportation Research Record*, 1752, 23-31.

Goulias, K.G., and T.G. Kim (2005) An analysis of activity type classification and issues related to the with whom and for whom questions of an activity diary. In *Progress in Activity-Based Analysis*, Timmermans, H.J.P. (Ed.), Elsevier, Oxford, England.

Greene, W.H., and D.A. Hensher (2010) *Modeling Ordered Choices: A Primer*. Cambridge University Press, Cambridge.

Hajivassiliou, V., D. McFadden, and P. Ruud (1996) Simulation of multivariate normal rectangle probabilities and their derivatives theoretical and computational results. *Journal of Econometrics*, 72(1-2), 85-134.

Hajivassiliou, V., and D. McFadden (1998) The method of simulated scores for the estimation of LDV models. *Econometrica*, 66(4), 863-896.

Hasegawa, H. (2010) Analyzing tourists' satisfaction: A multivariate ordered probit approach. *Tourism Management*, 31(1), 86-97.

Haskell, W.L., I-M Lee, R.R. Pate, K.E. Powell, S.N. Blair, B.A. Franklin, C.A. Macera, G.W. Heath, P.D. Thompson, and A. Bauman (2007) Physical activity and public health: Updated recommendations for adults from the ACSM and the AHA. *Circulation*, 116, 1081-1093.

Hautsch, N. (1999) Analyzing the time between trades with a gamma compounded hazard model. An application to LIFFE bund future transactions. Working paper, Center of Finance and Econometrics, University of Konstanz.

Hensher, D.A., and F.L. Mannering (1994) Hazard-based duration models and their application to transport analysis. *Transport Reviews*, 14(1), 63-82.

Herriges, J.A., D.J. Phaneuf, and J.L. Tobias (2008) Estimating demand systems when outcomes are correlated counts. *Journal of Econometrics*, 147(2), 282-298.

Higham, N.J. (2002) Computing the nearest correlation matrix – A problem from finance. *IMA Journal of Numerical Analysis*, 22(3), 329-343.

Higgs, M.D., and J.A. Hoeting (2010) A clipped latent variable model for spatially correlated ordered categorical data. *Computational Statistics & Data Analysis*, 54(8), 1999-2011.

Hjort, N.L., and C. Varin (2008) ML, PL, QL in Markov Chain Models. *Scandinavian Journal of Statistics*, 35(1), 64-82.

Hjorthol, R., and A. Fyhri (2009) Do organized leisure activities for children encourage car-use? *Transportation Research Part A*, 43(2), 209-218.

Holly, S., M.H. Pesaran, and T. Yamagata (2010) A spatio-temporal model of house prices in the USA. *Journal of Econometrics*, 158(1), 160-173.

Hothorn, T., F. Bretz, and A. Genz (2001) On multivariate *t* and gauss probabilities in *R*. *R News*, 1(2), 27-29.

Hunt, J.D., and J. E. Abraham (2007) Influences on bicycle use. *Transportation*, 34(4), 453-470.

Irwin, E.G., and N.E. Bockstael (2002) Interacting agents, spatial externalities and the evolution of residential land use patterns. *Journal of Economic Geography*, 2(1), 31-54.

Jacobsen P.L. (2003) Safety in numbers: More walkers and bicyclists, safer walking and bicycling. *Injury Prevention*, 9(3), 205-9.

Jacobson, S.H., and D.M. King (2009) Measuring the potential for automobile fuel savings in the US: The impact of obesity. *Transportation Research Part D*, 14(1), 6-13.

Jeliazkov, I., J. Graves, and M. Kutzbach (2008) Fitting and comparison of models for multivariate ordinal outcomes. *Advances in Econometrics*, 23, 115-156.

Joe, H., and Y. Lee (2009) On weighting of bivariate margins in pairwise likelihood. *Journal of Multivariate Analysis*, 100(4), 670-685.

Jones, P. (1979) New approaches to understanding travel behaviour: The human activity approach. In *Behavioral Travel Modeling*, D.A. Hensher, and P.R. Stopher (Eds.), Redwood Burn Ltd., London, 55-80.

Jones, K., and C. Duncan (1996) People and places: The multilevel model as a general framework for the quantitative analysis of geographical data. In *Spatial Analysis: Modelling in a GIS Environment*, P. Longley, and M. Batty (Eds.), GeoInformation International, Cambridge, 79-104.

Jones, P., F.S. Koppelman, and J. Orfeuil (1990) Activity analysis: State-of-the-art and future directions. In *Developments in Dynamic and Activity-based Approaches to Travel Analysis*. A compendium of papers from the 1989 Oxford Conference, P. Jones (Ed.), Avebury, U.K., 34-35.

Kakamu, K., and H. Wago (2007) Bayesian spatial panel probit model with an application to business cycle in Japan. Working paper. Available at: http://www.mssanz.org.au/modsim05/proceedings/papers/kakamu_2.pdf.

Kapur, A., and C.R. Bhat (2007) On modeling adults' daily time use by activity purpose and accompaniment arrangement. *Transportation Research Record*, 2021, 18-27.

Kato, H., and M. Matsumoto (2009) Intra-household interaction in a nuclear family: A utility-maximizing approach. *Transportation Research Part B*, 43(2), 191-203.

Keane, M. (1990) Four essays in empirical macro and labor economics. PhD Thesis, Brown University.

Keane, M. (1994) A computationally practical simulation estimator for panel data. *Econometrica*, 62(1), 95-116.

Kiefer, N.M. (1988) Economic duration data and hazard functions. *Journal of Economic Literature*, 26(2), 646-679.

Kim, S., and G.F. Ulfarsson (2008) Curbing automobile use for sustainable transportation: Analysis of mode choice on short home-based trips. *Transportation*, 35(6), 723-737.

Kingham, S., and S. Ussher (2007) An assessment of the benefits of the walking school bus in Christchurch, New Zealand. *Transportation Research Part A*, 41(6), 502-510.

Kitamura, R. (1987) A panel analysis of household car ownership and mobility, infrastructure planning and management. *Proceedings of the Japan Society of Civil Engineers*, 383/IV-7, 13-27.

Kitamura, R. (1988) A dynamic model system of household car ownership, trip generation, and modal split: Model development and simulation experiment. *Proceedings of the 14th Australian Road Research Board Conference*, Part 3, 96-111.

Kline, P. (1994) *An easy guide to factor analysis*. First edition, Routledge.

Kuk, A.Y.C., and D.J. Nott (2000) A pairwise likelihood approach to analyzing correlated binary data. *Statistics & Probability Letters*, 47(4), 329-335.

LaMondia, J., and C.R. Bhat (2011) A conceptual and methodological framework of leisure activity loyalty accommodating the travel context. *Transportation*, forthcoming.

Lee, L-F., and J. Yu (2010) Estimation of spatial autoregressive panel data models with fixed effects. *Journal of Econometrics*, 154(2), 165-185.

Lele, S.R. (2006) Sampling variability and estimates of density dependence: A composite-likelihood approach. *Ecology*, 87(1), 189-202.

LeSage, J.P., and R.K. Pace (2009) *Introduction to spatial econometrics*. Chapman & Hall/CRC, Taylor & Francis Group, Boca Raton.

Li, X., and X.P. Liu (2007) Defining agents' behaviors to simulate complex residential development using multicriteria evaluation. *Journal of Environmental Management*, 85(4), 1063-1075.

Lindsay, B.G. (1988) Composite likelihood methods. *Contemporary Mathematics*, 80, 221-239.

Liu, I., and A. Agresti (2005) The analysis of ordered categorical data: An overview and a survey of recent developments. *TEST: An Official Journal of the Spanish Society of Statistics and Operations Research*, 14(1), 1-73.

Luechinger, S., A. Stutzer, and R. Winkelmann (2010) Self-selection models for public and private sector job satisfaction. In *Research in Labor Economics*, S.W. Polachek, and K. Tatsiramos (Eds.), Emerald Group Publishing Limited, 30, 233-251.

Lumeng, J.C., P. Forrest, D.P. Appugliese, N. Kaciroti, R.F. Corwyn, and R.H. Bradley (2010) Weight status as a predictor of being bullied in third through sixth grades. *The Journal of Pediatrics*, 125(6), 1301-1307.

Mallett, W.J., and N. McGuckin (2000) Driving to distractions: Recreational trips in private vehicles. *Transportation Research Record*, 1719, 267-272.

Mardia, K., J.T. Kent, G. Hughes, and C.C. Taylor (2009) Maximum likelihood estimation using composite likelihoods for closed exponential families. *Biometrika*, 96(4), 975-982.

McFadden, D., and K. Train (2000) Mixed MNL models for discrete response. *Journal of Applied Econometrics*, 15(5), 447-470.

McGinn, A.P., K.R. Evenson, A.H. Herring, S.L. Huston, and D.A. Rodriguez (2007) Exploring associations between physical activity and perceived and objective measures of the built environment. *Journal of Urban Health*, 84(2), 162-184.

McKelvey, R.D., and W. Zavoina (1971) An IBM Fortran IV program to perform n-chotomous multivariate probit analysis. *Behavioral Science*, 16, 186-187.

McKelvey, R.D., and W. Zavoina (1975) A statistical model for the analysis of ordinal-level dependent variables. *Journal of Mathematical Sociology*, 4, 103-120.

Meka, S., R. Pendyala, and M. Kumara (2002) A structural equations analysis of within-household activity and time allocation between two adults. Presented at the 81st Annual Meeting of the Transportation Research Board, Washington, D.C.

Meurs, H. (1989) Dynamic analysis of trip generation. Presented at the International Conference on dynamic behavior analysis, Kyoto, Japan, July 21-23, 1989.

Meyer, B.D. (1990) Unemployment insurance and unemployment spells. *Econometrica*, 58(4), 757-782.

Mi, X., T. Miwa, and T. Hothorn (2009) mvtnorm: New numerical algorithm for multivariate normal probabilities. *The R Journal*, 1, 37-39.

Mitchell, J., and M. Weale (2007) The reliability of expectations reported by British households: Micro evidence from the BHPS. National Institute of Economic and Social Research discussion paper.

Molenberghs, G., and G. Verbeke (2005) *Models for Discrete Longitudinal Data*. Springer Series in Statistics, Springer Science + Business Media, Inc., New York.

Muller, G., and C. Czado (2005) An autoregressive ordered probit model with application to high frequency financial data. *Journal of Computational and Graphical Statistics*, 14(2), 320-338.

Neumayer, E., and T. Plümper (2010) Spatial effects in dyadic data. *International Organization*, 64(1), 145-166.

NHTS (2009) Telephone (CATI) questionnaire extended interview. Available at: http://nhts.ornl.gov/2009/pub/ExtendedInterview.pdf.

Ogden, C.L., M.D. Carroll, L.R. Curtin, M.M. Lamb, and K.M. Flegal (2010) Prevalence of high body mass index in US children and adolescents, 2007-2008. *The Journal of the American Medical Association*, 303(3), 242-249.

Ogilvie, D., M. Egan, V. Hamilton, and M. Petticrew (2004) Promoting walking and cycling as an alternative to using cars: Systematic review. *BMJ*, 329, 763-766.

Pace, L., A. Salvan, and N. Sartori (2011) Adjusting composite likelihood ratio statistics. *Statistica Sinica*, 21(1), 129-148.

Parsons Brinckerhoff Quade and Douglas, Inc. (2000) Comparative analysis weekday and weekend travel with NPTS integration for the RT-HIS: Regional travel-household interview survey. Prepared for the New York Metropolitan Council and the North Jersey Transportation Planning Authority, February 2000.

Pendyala, R.M., and K.G. Goulias (2002) Time use and activity perspectives in travel behavior research. *Transportation*, 29(1), 1-4.

Phaneuf, D.J., and R.B. Palmquist (2003) Estimating spatially and temporally explicit land conversion models using discrete duration. Prepared for the 2003 AERE Workshop in Madison, WI. Unpublished, available at: http://www.aere.org/meetings/0306workshop_Phaneuf.pdf.

Pikora, T., B. Giles-Corti, F. Bull, K. Jamrozik, and R. Donovan (2003) Developing a framework for assessment of the environmental determinants of walking and cycling. *Social Science & Medicine,* 56(8), 1693-1703.

Pinjari, A.R., and C.R. Bhat (2011) Activity-based travel demand analysis. In *Handbook in Transport Economics*, A. de Palma, R. Lindsey, E. Quinet, and R. Vickerman (Eds.), Edward Elgar Publishing, forthcoming.

Pinjari, A.R., B.S. Rajagopalan, N. Ferdous, and C.R. Bhat (2008) Activity-based model development: A review of current activity-based travel demand models. Report prepared for the North Central Texas Council of Governments (NCTCOG), March 2008.

Plaut, P.O. (2005) Non-motorized commuting in the US. *Transportation Research Part D,* 10(5), 347-356.

Prentice, R.L., and L.A. Gloeckler (1978) Regression analysis of grouped survival data with application to breast cancer data. *Biometrics*, 34, 57-67.

Quddus, M.A., C. Wang, and S.G. Ison (2010) Road traffic congestion and crash severity: Econometric analysis using ordered response models. *Journal of Transportation Engineering*, 136(5), 424-435.

R Development Core Team (2009) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Rebonato, R., and P. Jaeckel (1999) The most general methodology for creating a valid correlation matrix for risk management and option pricing purposes. *The Journal of Risk*, 2(2), 17-28.

Renard, D., G. Molenberghs, and H. Geys (2004) A pairwise likelihood approach to estimation in multilevel probit models. *Computational Statistics & Data Analysis*, 44(4), 649-667.

Rietveld, P., and V. Daniel (2004) Determinants of bicycle use: Do municipal policies matter? *Transportation Research Part A,* 38(7), 531-550.

Robertson, R.D., G.C. Nelson, and A. De Pinto (2009) Investigating the predictive capabilities of discrete choice models in the presence of spatial effects. *Papers in Regional Science,* 88(2), 367-388.

Robinson, D.L. (2005) Safety in numbers in Australia: More walkers and bicyclists, safer walking and bicycling. *Health Promotion Journal of Australia*, 16 (1), 47-51.

Roorda, M.J., A. Paez, C. Morency, R. Mercado, and S. Farber (2010) Trip generation of vulnerable populations in three Canadian cities: A spatial ordered probit approach. *Transportation*, 37(3), 525-548.

Saelens, B.E., J.F. Sallis, and L.D. Frank (2003) Environmental correlates of walking and cycling:Findings from the transportation, urban design, and planning literatures. *Annals of Behavioral Medicine*, 25(2), 80-91.

Sallis, J.F., J.J. Prochaska, and W.C. Taylor (2000) A review of correlates of physical activity of children and adolescents. *Medicine & Science in Sports & Exercise*, 32(5), 963-975.

Schoettle., K., and R. Werner (2004) Improving "the most general methodology to create a valid correlation matrix". In *Risk Analysis IV, Management Information Systems*, C.A. Brebbia (Ed.), WIT Press, Southampton, U.K., 701-712.

School data (2010) School district locator - data download. Available at: http://ritter.tea.state.tx.us/SDL/sdldownload.html.

Scott, D.M., and K.W. Axhausen (2006) Household mobility tool ownership: Modeling interactions between cars and season tickets. *Transportation*, 33(4), 311-328.

Scott, D.M., and P.S. Kanaroglou (2002) An activity-episode generation model that captures interactions between household heads: Development and empirical analysis. *Transportation Research Part B*, 36(10), 875-896.

Scotti, C. (2006) A bivariate model of Fed and ECB main policy rates. International Finance Discussion Papers 875, Board of Governors of the Federal Reserve System (U.S.).

Sener, I.N., and C.R. Bhat (2007) An analysis of the social context of children's weekend discretionary activity participation. *Transportation*, 34(6), 697-721.

Sener, I.N., N. Eluru, and C.R. Bhat (2009) On jointly analyzing the physical activity participation levels of individuals in a family unit using a multivariate copula framework. Technical paper, Department of Civil, Architectural & Environmental Engineering, The University of Texas at Austin.

Smirnov, O.A. (2010) Spatial econometrics approach to integration of behavioral biases in travel demand analysis. *Transportation Research Record*, 2157, 1-10.

Smith, T., and J.P. LeSage (2004) A Bayesian probit model with spatial dependencies. In *Advances in Econometrics: Volume 18: Spatial and Spatiotemporal Econometrics*, J.P. LeSage, and R.K. Pace (Eds.), Elsevier Ltd, Oxford, 127-160.

Srinivasan, K.K., and S.R. Athuru (2005) Analysis of within-household effects and between household differences in maintenance activity allocation. *Transportation*, 32(5), 495-521.

Srinivasan, S., and C.R. Bhat (2005) Modeling household interactions in daily in-home and out-of-home maintenance activity participation. *Transportation*, 32(5), 523-544.

Srinivasan, S., and C.R. Bhat (2006) Companionship for leisure activities: An empirical analysis using the American Time Use Survey. Presented at the Innovations in Travel Demand Modeling Conference, Transportation Research Board Conference Proceedings, 42(2), 129-136.

Srinivasan, S., and C.R. Bhat (2008) An exploratory analysis of joint-activity participation characteristics using the American time use survey. *Transportation*, 35(3), 301-328.

Stein, M., Z. Chi, and L. Welty (2004) Approximating likelihoods for large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(2), 275-296.

Strong, W.B., R.M. Malina, C.J.R. Blimkie, S.R. Daniels, R.K. Dishman, B. Gutin, A.C. Hergenroeder, A. Must, P.A. Nixon, J.M. Pivarnik, T. Rowland, S. Trost, and F. Trudeau (2005) Evidence based physical activity for school-age youth. *The Journal of Pediatrics*, 146(6), 732-737.

Swallen, K.C, E.N. Reither, S.A. Haas, and A.M. Meier (2005) Overweight, obesity and health-related quality of life among adolescents: The national longitudinal study of adolescent health. *The Journal of Pediatrics*, 115(2), 340-347.

Thornton, A., T.L. Orbuch, and W.G. Axinn (1995) Parent-child relationships during the transition to adulthood. *Journal of Family Issues*, 16(5), 538-564.

Thorpe, K.E., C.S. Florence, D.H. Howard, and P. Joski (2004) The impact of obesity on rising medical spending. *Health Affairs*, 23, 480-486.

Timmermans, H.J.P., and J. Zhang (2009) Modeling household activity travel behavior: Examples of state of the art modeling approaches and research agenda. *Transportation Research Part B*, 43(2), 187-190.

Tonne, C., S. Melly, M. Mittleman, B. Coull, R. Goldberg, and J. Schwartz (2007) A case–control analysis of exposure to traffic and acute myocardial infarction. *Environmental Health Perspective*, 115(1), 53-57.

Train, K. (2003) *Discrete Choice Methods with Simulation*. First edition, Cambridge University Press, Cambridge.

US Census Bureau (2009a) FM-1. Families, by presence of own children under 18: 1950 to present. Available at: http://www.census.gov/population/socdemo/hh-fam/fm1.xls.

US Census Bureau (2009b) FM-3. Average number of own children under 18 per family, by type of family: 1955 to present, http://www.census.gov/population/www/socdemo/hh-fam.html, accessed October 29, 2010.

Varin, C. (2008) On composite marginal likelihoods. *Advances in Statistical Analysis*, 92(1), 1-28.

Varin, C., and C. Czado (2008) A mixed probit model for the analysis of pain severity diaries. Available at: http://www-m4.ma.tum.de/Papers/Czado/varin_czado_pain_diaries.pdf

Varin, C., and C. Czado (2010) A mixed autoregressive probit model for ordinal longitudinal data. *Biostatistics*, 11(1), 127-138.

Varin, C., G. Host, and O. Skare (2005) Pairwise likelihood inference in spatial generalized linear mixed models. *Computational Statistics & Data Analysis*, 49(4), 1173-1191.

Varin, C., N. Reid, and D. Firth (2011) An overview of composite likelihood methods. *Statistica Sinica*, 21(1), 5-42.

Varin, C., and P. Vidoni (2005) A note on composite likelihood inference and model selection. *Biometrika*, 92(3), 519-528.

Varin, C., and P. Vidoni (2006) Pairwise likelihood inference for ordinal categorical time series. *Computational Statistics & Data Analysis*, 51(4), 2365-2373.

Varin, C., and P. Vidoni (2009) Pairwise likelihood inference for general state space models. *Econometric Reviews*, 28(1-3), 170-185.

Vovsha, P., and M. Bradley (2006) Advanced activity-based models in context of planning decisions. *Transportation Research Record*, 1981, 34-41.

Vovsha, P., E. Peterson, and R. Donnelly (2003) Explicit modeling of joint travel by household members: Statistical evidence and applied approach. *Transportation Research Record*, 1831, 1-10.

Wang, Y., M.A. Beydoun, L. Liang, B. Caballero, and S.K. Kumanyika (2008) Will all Americans become overweight or obese? Estimating the progression and cost of the US obesity epidemic. *Obesity*, 16(10), 2323-2330.

Wang, D., and J. Li (2009) A model of household time allocation taking into consideration of hiring domestic helpers. *Transportation Research Part B*, 43(2), 204-216.

Wang, M., and J. M. Williamson (2005) Generalization of the Mantel-Haenszel estimating function for sparse clustered binary data. *Biometrics*, 61(4), 973-981.

Ward, P.S., R.J.G.M. Florax, A. Flores-Lagunes (2010) Agricultural productivity and anticipated climate change in Sub-Saharan Africa: A spatial sample selection model. Available at: http://ageconsearch.umn.edu/bitstream/61635/2/Ward%20et%20al%202010%20AAEA%20Paper.pdf.

Wen, C-H., and F.S. Koppelman (1999) An integrated system of stop generation and tour formation for the analysis of activity and travel patterns. *Transportation Research Record*, 1676, 136-144.

Williams, A. (2003) Adolescents' relationships with parents. *Journal of Language and Social Psychology*, 22(1), 58-65.

Winship, C., and R.D. Mare (1984) Regression models with ordinal variables. *American Sociological Review*, 49(4), 512-525.

World Health Organization (WHO) (2006) Obesity and owerweight. Available at: http://www.who.int/dietphysicalactivity/media/en/gsfs_obesity.pdf.

Xing, Y., S.L. Handy, and P.L. Mokhtarian (2010) Factors associated with proportions and miles of bicycling for transportation and recreation in six small US cities. *Transportation Research Part D*, 15(2), 73-81.

Xiong, W. (2011) Measuring the monetary policy stance of the people's bank of China: An ordered probit analysis. *China Economic Review*, forthcoming.

Yamamato, T., and R. Kitamura (1999) An analysis of time allocation to in-home and out-of-home discretionary activities across working days and non-working days. *Transportation*, 26(2), 211-230.

Yamamoto, T., R. Kitamura, and R.M. Pendyala (2004) Comparative analysis of time-space prism vertices for out-of-home activity engagement on working and non-working days. *Environment and Planning B*, 31(2), 235-250.

Yi, G.Y., L. Zeng, and R.J. Cook (2011) A robust pairwise likelihood method for incomplete longitudinal binary data arising in clusters. *Canadian Journal of Statistics*, 39(1), 34-51.

Zhao, Y., and H. Joe (2005) Composite likelihood estimation in multivariate data analysis. *The Canadian Journal of Statistics*, 33(3), 335-356.