ACTIVITY-BASED TRAVEL DEMAND MODEL: APPLICATION AND INNOVATION

LI SIYU

(B. Eng. Tsinghua University)

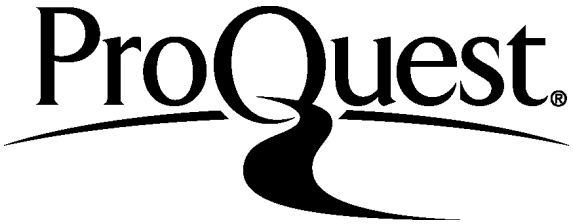
A THESIS SUBMITTED

FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

DEPARTMENT OF CIVIL AND ENVIRONMENTAL ENGINEERING

NATIONAL UNIVERSITY OF SINGAPORE

2015



Declaration

I hereby declare that this thesis is my original work and it has been written by me in its entirety. I have duly acknowledged all the sources of information which have been used in the thesis.

This thesis has also not been submitted for any degree in any university previously.

Li Siyu

June 2015

Acknowledgments

I am indebted to a great number of people who generously offered friendship, inspiration, advise, and encouragement throughout my PhD study. First of all, I would like to express my special appreciation and thanks to my supervisor Professor Lee Der-Horng. Academically, he has been a tremendous mentor and the discussions with him have always been enlightening. In daily life, his unique characteristics and sense of humor are influential, cheering me up at those dark moments. I truly feel that his immense knowledge, sharp views and concise advice are valuable for not only this study, but also my career development.

I am deeply grateful to my co-supervisor and one of the PIs of Future (Urban) Mobility IRG, Professor Moshe Ben-Akiva, who guided me throughout the SimMobilty project and my research, and gave me the opportunity to work with other researchers in FM IRG of Singapore-MIT Alliance for Research and Technology. His eternal enthusiasm and extensive experience in research have gradually shaped the way I do research.

The gratitude also goes to SMART and NUS, who provides me with financial supports, a decent work place and occasional free food.

Thanks to the members of my doctoral committee: Professor Diao Mi and Professor Meng Qiang, for their encouragement and helping me to improve the thesis. Professor Meng Qiang is also the lecturer of several modules I took during the PhD study. His challenging questions in exams are impressive and I learned a lot from him and his modules.

I would like to express my appreciation to all the colleagues in FM and those with whom I ever worked with, Maya Abou Zeid, Xiang Yunke, Kakali Basak, Francisco Pereira, Harish Loganathan, Milan Lovric, Sebastian Raveau Feliu.

Moreover, I would like to thank those fellow modelers that I worked together for developing the activity-based framework: Xiao Yu, Huang He, Carlos Carrion, Tan Rui and Muhammad Adnan.

I also would like to thank all the friends I encountered in NUS for the support and companionship, Sun Lijun, Lu Yang, Jin Jiangang, Wu Xian, Qin Han, Lu Zhaoyang, He Nanxi, Zhao Kangjia, Ye Qing, Katarzyna Anna Marczuk, Cheng Zhiru, etc.

Finally, my deepest gratitude goes to my parents Li Guangsheng and Si Fengping, my girlfriend Jiang Yudan and all my family members for their continuous love, support and understanding. I made the decision to start the journey myself but manage to finish it because of you.

Table of Contents

Executive Summary

The primary and most fundamental purpose of transportation planning is to support decision-makings for transportation systems with data and information. Models are extensively used in the transportation planning process in order to forecast future travel patterns and develop efficient future transportation systems. Travel demand modeling is at the heart of the transportation planning process. It is used to estimate the distribution of travel demand or traffic flows made on alternative transportation systems in the future. The last century witnessed the transition in the focus of the transportation planning process: from capital-intensive investment to policy domains and inventory-based planning that focus on operation and management. Throughout the years, travel demand modeling has been evolving to respond to the emerging planning contexts. Among all these evolutions, the most significant one is from the aggregate four-step method to the disaggregate and activity-based perspective.

This thesis is dedicated to the design and implementation of an activity-based modeling framework for Singapore, specifically focusing on the comprehensive development of the framework and moving the state-of-the-art into empirical and innovative implementations. This core objective is served in the thesis with several specific issues to address. The first part of the thesis focuses on the data needed for the development of operational activity-based models and two specific topics are investigated: to generate time-dependent travel time with fine resolution and to utilize existing trip-based travel surveys in Singapore for the development of activity-based models. The second part of the thesis is devoted to the design and implementation of an operational full-fledged activity-based modeling framework, the pre-day activity-based model, and its simulator for Singapore. The system design, estimation of individual components, key features

of the simulator, interface between the model and the simulator and model calibration/validation process are introduced with details. The third part of the thesis intends to advance the studies on human daily activity patterns by providing new perspective and methodology in the modeling and learning of daily activity patterns using probabilistic context-free grammars. Practically, the proposed methodology sheds light on the issue of generating customized choice sets and provides an alternative daily activity pattern model implementation—a two-stage choice model for daily activity patterns based on the grammar-based representation of daily activity patterns—for the pre-day modeling framework.

In summary, the original contribution of the thesis is two-fold: Empirically, an operational activity-based modeling framework is implemented, which serves as a benchmark and more advanced models of various activity-travel decision-making facets can be incorporated and tested. The framework also represents one of the core functionalities in an integrated travel demand and supply simulation platform with a consistent individual-based representation of travelers. In the future, the framework may be used to test the efficiency of a series of travel demand management (TDM) policies and strategies. Methodologically, the grammar-based representation of daily activity patterns is unique and innovative. Its applications in the activity-based modeling framework, such as representing daily activity patterns as activity sequences and generating customized choice sets for each individual are explored for the first time in this thesis.

List of Tables

List of Tables

List of Figures

List of Figures

List of Figures

Chapter 1

Introduction

1.1 Overview

1.1.1 Transportation planning and travel demand modeling

A transportation system, with its infrastructure, level of service and maintenance, as well as social and environmental benefits, has been one of the major factors in determining the economic prospect of interested regions (e.g., cities, regional areas, and countries). According to Lefevre et al. (2014), the global transport investment is between US$1.4 and US$2.1 trillion each year. With the increasing urban dwellers worldwide especially in developing countries[1] and the closer connection among cities and regions, the annual investment is expected to keep rising gradually. The growing urban population has brought uncertainty into the future directions of urbanization, which influences the geographical and functional layout of transportation systems and the flow of transportation-related investment. Decision-makings are therefore necessary in order to neutralize the potential risks embedded in the uncertainty.

Decision-makings for transportation systems are essential and valuable as the investment in transportation is particularly influential: the roads, rails, and stations will last for centuries as well as dictate long-term urban development and resource consumption patterns. Before putting projects and policies into practice, it is very important that the expenditure to be properly justified, the permission to be granted, the social and environmental impacts to be fully evaluated. The primary and most fundamental purpose of transportation planning is to

[1] According to the estimates from the United Nations, the urban population is expected to be doubled in the next 30 to 40 years, where 96 percent of the growth will be taken place in developing countries.

support those decision-makings for transportation systems with data and information such that transportation facilities and services are created at a reasonable cost with a minimal environmental impact and to enhance economic activities.

Typical transportation planning involves forecasting future travel patterns to develop efficient future transportation systems. To serve the purpose, transportation planning uses models extensively, where behavioral realism, mathematical relationships and simulation techniques are blended together to represent human behavior in decision-makings. Forecasting models are used in sequence for population forecasts, economic forecasts, land use forecasts and travel forecasts. Travel forecasts are usually carried out with travel demand modeling.

Travel demand modeling is at the heart of the transportation planning process. It is used to estimate the distribution of travel demand or traffic flows made on alternative transportation systems in the future. The prediction of future travel demand is an essential task of the long-range transportation planning process for determining strategies (for example but not limited to, expansion of transportation supply and service, pricing skims and land use policies) that can accommodate future needs under alternative socio-economic scenarios, and for alternative transport services and land-use configurations. Historically, travel demand models were developed in 1950s originally for the purpose of forecasting highway demand (City of Chicago, 1959). Throughout the years, travel demand models have been evolving and we shall see subsequently in the introduction that how the needs to apply the models to modern transportation planning contexts as well as the needs for more realistic representation of behavior in travel demand modeling have stimulated the process.

1.1.2 Emerging planning contexts

As mentioned above, the predecessor of modern transportation planning process started with highway planning in the United States in 1950s, when urban areas were being connected by interstate highways. Over the half century, although the basic mission of transportation planning has remained the same over the periods, that is to support decision-makings of providing mobility in as safe and cost-effective manner as possible, the core of it has been stretched and augmented (as shown in Figure 1.1) to reflect changing issues of concern to

policy makers and to respond to the much expanded contexts (Meyer, 2000).

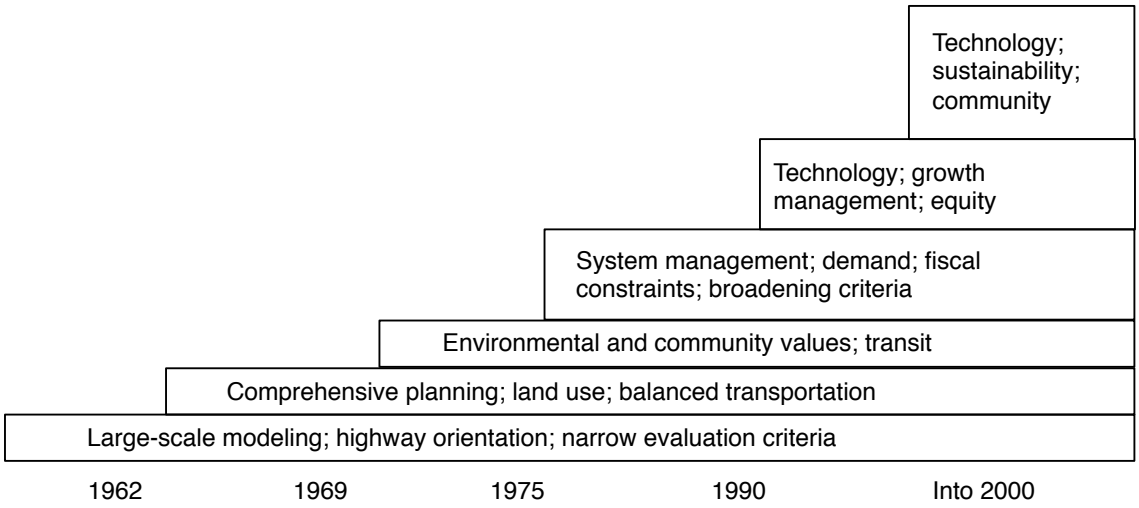


Figure 1.1: Key issues in the evolution of transportation planning (adapted from Meyer, 2000)

According to Meyer (2000), the emerging planning contexts from decade to decade are summarized as follows.

In 1950s, the focus of transportation planning was to accommodate the increased traffic loads (resulted from the booming auto ownership) using improved highway systems. It was clear that the extension of highway networks involved large capital investment, which was then linked to economic development.

Into 1960s, the uncoordinated planning practice had caused conflicts between highways, housing and land use goals. A formalized transportation planning process with long-term goals, broader perspective, coordination with land use plans, and analytical analysis began to be valued.

The next decade labeled the end of post-war economic boom, inflation, fuel shortages and increasing concerns on environmental and equity issues. Under such circumstances, the opposition and criticism to the transport development with large capital investment started to emerge and put considerable pressure on the planning process and its ability to adapt to the change. In this period, new focus of transportation planning emerged, such as environmental impact and air conformity analysis for transportation projects, multimodal transportation planning process with explicit and equal consideration of public transit, and

short-range capital investment and transportation planning to reduce traffic congestion and other immediate problems.

Into 1980s, there had been a growing concern on the increased range and complexity of issues to be addressed in the transportation planning process since the last decade. The need for systematic transportation planning with flexibility to take care local concerns was confirmed along with the attention to transportation system management. At the end of the decade, congestion had reached severe level in many suburban and downtown areas as a result of urban sprawling. Investment into new highways was difficult with limited budget, strengthened environmental awareness and the risk of inducing more traffic. As a result, transportation demand management (TDM) strategies that aim to mitigate congestion by modifying demand patterns (e.g., reducing trips, temporal shift of trips and promotion of high occupancy modes) started to emerge and to be evaluated in the transportation planning process.

1990s witnessed the end of major highway construction and many planning agencies had shifted to short-term planning to deal with the deterioration of transportation infrastructure and transportation congestion. Moreover, the growth in travel demand was no longer able to be accommodated by increasing supply levels. As a result, integrated TDM strategies, especially those enabled by advanced technology, had a greater role to play in revitalized long-term strategic transportation planning process.

In the last 10 to 15 years, the influence of new technology to transportation planning has been dramatic. Firstly, telecommunication technologies, such as smartphones, have changed the way people communicate and receive information; commercial and civil GPS usage has been increasing for navigation and location-based service. Secondly, Intelligent Transportation System (ITS) has created an infused technology and information layer to existing infrastructure, which helps to increase the level of operation and management. Thirdly, new vehicles (such as autonomous cars) and services (such as car-sharing economy) for ground transportation have re-shaped the definition of transportation systems and mobility. To incorporate the policies and management strategies enabled by those technologies has been a major issue in the new century.

The emerging and enriched contexts of transportation planning are majorly based on the experience of the United States (Weiner, 1997). However, the major trends also apply to other developed regions where transportation planning has shifted its focus from capital-intensive investment to travel demand and transportation system management and policies, from purely economic consideration to performance-based criteria with broader perspective such as social, environmental factors, and more advanced technologies are introduced to influence system management and travel behavior. Although developing countries are insufficient in transportation infrastructure and large capital investment still dominates their transportation planning process, the experience introduced here provides clues on their future directions.

As the core of the transportation planning process, travel demand modeling has been evolving to respond to the emerging planning contexts, which is introduced in the next section.

1.1.3 Evolving modeling regimes

Prior to 1950s, traffic counts were used to assess the use of transport systems. Demand modeling was either based on existing travel demand or predictions with uniform growth factors, which were derived with a consideration of historical trends and were coarse. The accelerated highway construction in 1950s required more sophisticated prediction tools, with which the future travel demand can be met and economic impacts can be assessed. The development of computers in this period allowed the implementation of large-scale travel demand models. The method emerged in this decade and formalized in 1960s is often referred to as the four-step travel demand models.

The four-step models predict travel demand and traffic flows using a sequential procedure. In the first step, trip generation involves the estimation of the number of home-based and non-home based person trips produced from, and attracted to, each analysis zone in the study area. The second step, trip distribution, determines the distribution of all trips among all origin-destination (OD) pairs. An aggregate travel demand represented as an OD matrix is derived from the first two steps. Then, a mode choice model is applied to create travel demand matrices by travel mode. Finally, trips are assigned to the road network and traffic flows are obtained.

By the end of 1960s, application of the four-step method had become popular. However, from contemporary planners' point of view, there are several issues and arguments that start to hinder the application of the four-step method since 1970s (Rasouli and Timmermans, 2014), when there was a trend towards inventory-based planning (e.g., low capital options, dynamic, policy-sensitive and demand-responsive systems) and broader social and environmental perspectives that the four-step method failed to incorporate:

- Lack of integrity: There is no consistency between various models in the four-step method. For example, travel time produced from traffic assignment is not consistent with travel time used in the mode choice model.
- Aggregate nature: All origins and destinations (zones) are treated as single points; all households are assumed to be identical in the same zone; dynamic traffic patterns in different time periods are not represented.
- Lack of interdependency between trips: The four-step method uses trips as the unit of analysis such that the trip chaining characteristics and related constraints cannot be represented.
- Inadequate behavioral realism: The belief that relationships found in natural science, such as gravitational attraction, could be extended to urban systems is questionable. Choice mechanisms and constraints on choice are not incorporated.

From 1970s to 1990s, there had been substantial improvement to the original four-step method in response to the emerging criticism. For example, integrity between models is enhanced by looping among individual models; disaggregation is enhanced by generating trips for individuals and incorporating traffic dynamics with time-dependent travel demand; and micro-simulation-based traffic assignment is adopted to take driving behavior into consideration. However, the improved four-step methods are still fundamentally aggregate. Moreover, the disaggregate techniques were only tested in limited applications while most planning agencies remained highly dependent on the aggregate four-step method in 1980s.

Into 1990s, the gap between planning contexts and available demand modeling tools was wider than before as the major focus of transportation planning in developed regions

had been shifted from capital-intensive investment to travel demand and transportation system management and policies. As a result, the advocation of disaggregate models and simulations started, which stimulated the development of activity-based travel demand models as promising alternatives to the four-step method.

Modeling travel demand from an activity-based perspective is originated from the principle that the need for travel is derived from the need for activity participation. Follow the basic principle, different groups of activity-based models have emerged in the literature to incorporate the principle from various viewpoints. The development and deployment of activity-based models is reviewed subsequently in Chapter 2. It is noted that in the last two decades, many cities and regions in North America (mainly the United States), Europe and Japan have supported the development and evaluation of activity-based travel demand models and the research into activity-based models is still ongoing to incorporate emerging theoretical achievements and planning scenarios.

1.2 Research Scope and Objective

The focus of the thesis is on the comprehensive development of activity-based travel demand modeling framework and moving the state-of-the-art into empirical and innovative implementations.

Figure 1.2 represents an example of integrated and disaggregate framework for activity-travel decisions. Based on distinct time frames of decision-makings, household and individual choices can be grouped into long-term (in a time frame of years) mobility and lifestyle choice, mid-term (on daily basis) activity and travel scheduling, and short-term (within a day) plan implementation and rescheduling (Ben-Akiva et al., 1996). The decision-makings of these levels determine the travel demand and affect the performance of transportation systems (supply), which in turn, affect simultaneously individual decisions and urban development. Within the framework of distinct time frames and demand-supply interactions, the scope of the thesis is characterized by the activity and travel scheduling on daily basis, where activity-based travel demand modeling frameworks are usually placed. As a result, other

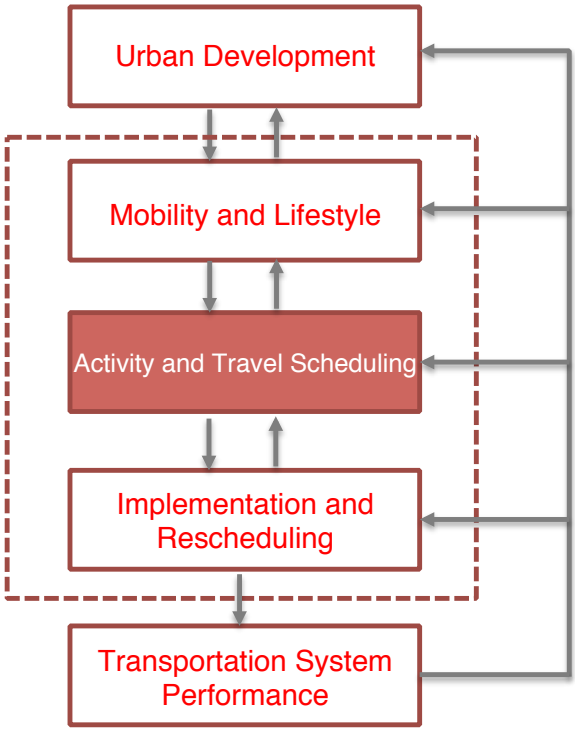


Figure 1.2: An example of integrated and disaggregate framework for activity/travel decisions (adapted from Ben-Akiva et al., 1996)

components are exogenous and not considered in the thesis. For example, long-term decisions are treated as given and not modeled.

This thesis is dedicated to the design and implementation of an activity-based modeling framework for Singapore. This core objective is served through Chapter 3 to Chapter 7 with specific issues to address. Figure 1.3 provides a schematic diagram of the thesis workflow, where the thesis and its contribution is divided into three parts.

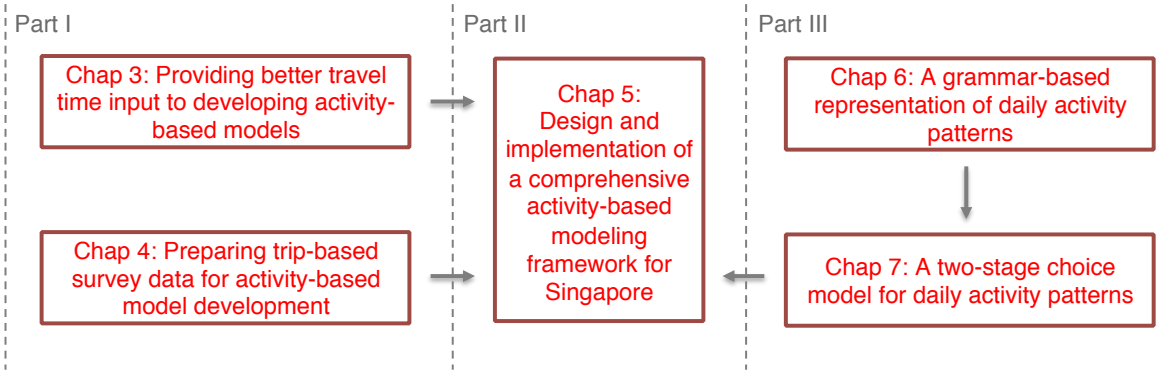


Figure 1.3: A schematic diagram of the thesis workflow

Part I focuses on the data (prerequisite) needed for the development of operational activity-based modeling frameworks. Two specific topics are isolated and investigated in Chapter 3 and Chapter 4, respectively. Transportation planning agencies usually maintain travel time matrices (or network skims) for a small number of time windows, which is low in resolution to develop TDM strategies that aim to shift time-of-travel patterns. Observing this gap, Chapter 3 tries to fuse trips collected from two sources and develop regression models to generate time-dependent travel time with fine resolution. Chapter 4 is dedicated to the intensive data processing efforts that aim to utilize existing trip-based travel surveys in Singapore and ready the surveys for the development of activity-based models. Data recollection efforts are avoided if it is proven in practice that trip-based survey is adequate for the implementation of activity-based travel demand models. Part I contributes to the model development efforts in Part II.

Part II is devoted to the design and implementation of an activity-based modeling framework, the pre-day activity-based model, and its simulator for Singapore. The pre-day activity-based model is formulated through a system of interconnected discrete choice models representing choices at distinct dimensions of daily activity schedule. The system design, estimation of individual components, key features of the simulator, interface between the model and the simulator and model calibration/validation process are introduced with details. The pre-day model intends to serve as an operational full-fledged benchmark, based on which, more advanced models of various activity-travel decision-making facets can be incorporated and tested. The pre-day model also represents one of the core functionalities in an integrated travel demand and supply simulation platform with a consistent individual-based representation of travelers.

Part III starts with an alternative definition and representation of daily activity patterns, which is different from the one in the pre-day model in Chapter 5. Chapter 6 intends to advance the studies on human daily activity patterns by providing new perspective and methodology in the modeling and learning of daily activity patterns using probabilistic context-free grammars. The heterogeneity in selecting daily activity patterns is captured in the several formulations proposed to estimate the probability of a context-free grammar. Practically, the proposed methodology sheds light on the issue of generating customized

choice sets for daily activity pattern models in certain activity-based modeling frameworks with hierarchical structures, such as the pre-day model. To validate the statement, Chapter 7 proposes a two-stage choice model for daily activity patterns based on the grammar-based representation of daily activity patterns, where a choice set is customized for each individual explicitly followed by a choice model. The two-stage choice model is implemented in the pre-day framework and able to replicate the activity participation behavior of the base year, thus able to replace the day pattern model introduced in Chapter 5.

1.3 Thesis Organization

The thesis is organized as follows:

Chapter 1 introduces the general research background, research scope and objective.

Chapter 2 provides a general literature review on activity-based travel demand modeling approach, focusing especially on the operational and implemented works and how daily activity patterns are represented and modeled.

Chapter 3 fuses trips collected from two sources and develops regression models to generate time-dependent travel time for developing activity-based models.

Chapter 4 prepares trip-based household travel survey ready for the development of activity-based models through trip-to-tour conversion process.

Chapter 5 introduces in detail the design and implementation of an activity-based modeling framework, the pre-day activity-based model, and its simulator.

Chapter 6 explores the grammar-based representation of daily activity patterns with the application of probabilistic context-free grammar.

Chapter 7 proposes and implements a two-stage model for daily activity patterns where a choice set is customized for each individual explicitly by taking advantage of the grammar-based representation, followed by a choice model.

Chapter 8 draws the concluding remarks and presents directions for future research.

Chapter 2

Literature Review

2.1 Overview

The thesis finds its root in the existing literature of activity-based travel demand models and intends to move the state-of-the-art into empirical and innovative implementations. The intention is fulfilled initially with a literature review on activity-based travel demand modeling approach, focusing especially on the operational and implemented works and how daily activity patterns are represented and modeled. The review is carried out in a general manner while the subsequent chapters on individual topics will give a short literature overview of respective problems.

It is mentioned in the first chapter that the trip-based approach, especially the four-step method was found to be inconsistent and unrealistic in several fundamental aspects. The impact of those fundamental weaknesses is relatively small in a supply-oriented planning process where the focus of transportation planning is to provide adequate transportation infrastructure with intensive capital investment. However, the new century has witnessed a transformation from supply-oriented planning to inventory-based planning where addressing policy concerns, environmental impacts, and management of travel demand with current level of supplies are the concentration of both researchers and practitioners. Under such circumstances, the weaknesses of the four-step method limit its further application and activity-based models are envisioned as promising alternatives to the four-step method.

The upward trend in the research of activity-based models is found in Figure 2.1 (two keywords, "activity-based" and "travel demand", are used for the search). The concept of

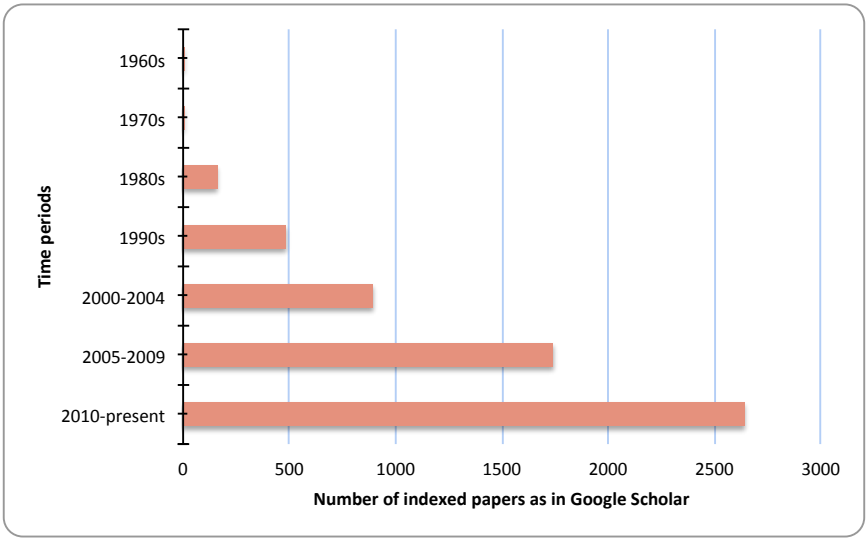


Figure 2.1: An upward trend in activity-based travel demand models (as in March 2015)

activity-based approach was picked up in 1960s-70s and found its origin in several studies. The link of travel and activity was established in a benchmark study of Mitchell and Rapkin (1954). The study also called for a comprehensive framework for travel behavior. However, it was not until 1970s, that the researchers and practitioners were motivated by several seminal studies in urban planning, sociology, and environmental studies to model the travel demand from the perspective of activity participation.

Hägerstrand (1970) provided the first explicit discussion on activity participation in the context of time-geographic and outlined the system of conceptual constraints for activity participation in time-space prism, in which geographical space is represented by a two-dimensional plane and time is defined on the remaining vertical axis. The representation allows activity-travel pattern definition in terms of a path through time and space. Chapin (1974) proposed a motivational framework in which activity demand is motivated by social constraints and individual desires. Fried et al. (1977) tried to address the household travel behavior in urban areas and understand the influence of social structures and rules. A first attempt to combine the elemental studies was observed in Cullen and Godson (1975), where the motivational approach and constraint approach were combined to create a model that depicts a routine and priority-based approach to activity generation and scheduling. Later, in a couple of more comprehensive studies by Jones and other researchers (see Jones, 1977

and Jones et al., 1983), the conceptual works by Hägerstrand, Chapin, Fried et al. were summarized and the relationship between activity and travel was explicitly considered. To be specific, the need for travel is derived from the need to participate various activities with social, temporal and spatial constraints. Since then, the principle that travel is a derived need from activity participation has become the common foundation and root for studies on activity-based approach. In additional to those studies, it is worth mentioning that at the same time period, McFadden (1973) set the foundation for Random Utility Maximization (RUM) theory, which has profound influence over the development of empirical activity-based models.

Into 1980s, researchers and practitioners started to pick up the concept of activity-travel pattern and the principle that travel demand stems from the need for activity participation. As a result, the conceptual and empirical development of activity-based models started to boom. So far, activity-based models have achieved considerable progress in terms of theorization, implementation, and most importantly, deployment. However, as the activity-based models slowly move to replace the traditional four-step method, the current state of the practice differs widely between countries (Rasouli and Timmermans, 2014).

In general, the activity-based models to date can be classified based on their modeling approach into several categories: (1) Constraints-based prototypes, (2) Rule-based models, (3) Econometric models and (4) Hybrid models. Most efforts in the development, implementation and deployment of activity-based models will be reviewed in the rest of the chapter from Section 2.2 to 2.5, focusing on their own characteristics, implementation and application details, and current status. A quick reference table can be found in Table 2.1. To summarize the review, some discussion on the concerns in the development of activity-based models is provided at the end of this chapter.

Framework/Model	Selected literature	Category	Daily activity pattern	Development/Deployment: Estimation	Validation	Implementation	Featured application	Current status
1. Constraints-based Prototype								
CARLA	Jones et al. (1983)	Constraints-based	Start/End time of a sequence of activities	N/A	N/A	N/A	Burford School study (Jones et al., 1983)	Not in use
2. Rule-based Approach								
STARCHILD	Recker (1995); Recker et al. (1986a,b)	Rule-based, Logit model	Activity schedule with purpose, timing, location and seuqnce	Yes (the Logit model in the final step)	N/A	N/A	Academic case studies in Orange County and Portland	Not in use
SCHEDULER	Gärling et al. (1989); Golledge et al. (1994)	Rule-based	Activity schedule with purpose, timing, location and seuqnce	N/A	N/A	N/A	Academic case studies (Golledge et al., 1994) by using a survey in Sacramento County by Kitamura et al. (1990)	Not in use
SMASH	Ettema et al. (1993, 1996)	Rule-based, Nested Logit model	Sequenced activities with purpose, timing, location, mode of travel, route	Yes (Nested Logit model)	N/A	N/A	Academic case studies with a sample in the city of Veldhoven, Netherlands (Ettema et al., 1996)	Not in use
AMOS (SAMS)	Kitamura et al. (1993, 1996); Pendyala et al. (1995, 1998); RDC Inc. (1995)	Rule-based, neural network	Sequenced activities with purpose, timing, location, mode of travel	Yes (Neural network is trained with AMOS Survey)	N/A	For Washington, D.C.	Evaluation of TDM strategies such as parking and congestion pricing, improved bicycle/pedestrian facilities (RDC Inc., 1995)	Abandoned
ALBATROSS (FEATHERS)	Arentze et al. (1999); Arentze and Timmermans (2000, 2004); Bellemans et al. (2010)	Rule-based, decision trees	Sequenced activities with purpose, timing, location, mode of travel	Yes (decision trees are derived with surveys in Rotterdam, 1997)	N/A	For Dutch MoT and Flanders, Belgium	Academic case studies: vehicle emission/air pollution (Beckx et al., 2009) and energy consumption (Yang et al., 2010)	Under development
3. Econometric Models with Individual Daily Activity Pattern								
Portland Metro Model	Bowman (1998); Bowman et al. (1998); Bradley et al. (1998); Bowman and Ben-Akiva (2001)	Utility-based	Abstraction of daily activity participation with sequence of tours, sub-tours and intermediate stops	Yes	No	For Portland Metro	Congestion pricing study on I-5 corridor (Bowman et al., 1998)	Abandoned
San Francisco (SFCTA) Model	Bradley et al. (2001); Cambridge Systematics, Inc. (2002); Freedman et al. (2006); Outwater and Charlton (2006)	Utility-based	Abstraction of daily activity participation with sequence of tours, sub-tours and intermediate stops	Yes	Yes	For SFCTA	Countryside transportation plan (Castiglione et al., 2006), New Central Subway project in SF (Freedman et al., 2006), Congestion pricing (San Francisco County Transportation Authority, 2010), Development of bike infrastructure (Hood et al., 2011), Evaluation of transit crowding (Zorn et al., 2012)	In use
Sacramento (SACOG) Model	Bradley et al. (2007, 2010); DKS Associate et al. (2012)	Utility-based	Occurrence of tours (0 or 1+) and intermediate stops (0 or 1+) for various activity purposes	Yes	Yes	For SACOG	Sacramento Region 2035 Metropolitan Transportation Plan (SACOG, 2008), Placer Vineyards transit-oriented development (SACOG, 2007), 2010 SB376 greenhouse gas emission analysis (California EPA Air Resources Board, 2010)	In use
Denver (DRCOG) Model	Sabina and Rossi (2006); Cambridge Systematics, Inc. (2010)	Utility-based	Occurrence of tours (0 or 1+) and intermediate stops (0 or 1+) for various activity purposes	Yes	Yes	For DRCOG	2010 Denver Regional Plan, Bike/Pedestrian Project Selection for Transportation Improvement Program, FTA New Starts Analysis	In use
Seattle (PSRC) Model	Nichols et al. (2014)	Utility-based	Occurrence of tours (0 or 1+) and intermediate stops (0 or 1+) for various activity purposes	Yes	Yes	For PSRC	Academic case studies on Automated Vehicles (Childress et al., 2015)	Under development

Oregon and Ohio SDT Model	Ohio Model: Parsons Brinckerhoff (2010), Oregon Model: Parsons Brinckerhoff et al. (2010)	Utility-based	Activity sequences (represented as a string of activities)	Yes	Yes	For Oregon DOT and Ohio DOT	Study of Eastern Oregon Freeway Alternatives, Modeling of transportation and land use alternatives, Oregon Freight Plan (Knudson and Weidner, 2010)	In use
			4. Econometric Models with Coordinated Daily Activity Pattern					
NYBPM	Parsons Brinckerhoff (2005b); Chiao et al. (2006)	Utility-based	Sequential determination: Number of journeys (tours) by person type for three activity types	Yes	Yes	For NYMTC	Case studies as mentioned in Chiao et al. (2006): Air quality conformity analysis, Southern Brooklyn transportation investment study, multiple bridge, corridor and expressway studies	In use
Columbus (MORPC) Model	Parsons Brinckerhoff (2005a); Anderson and Jiang (2013)	Utility-based	Sequential determination: Number of journeys (tours) by person type for three activity types	Yes	Yes	For MORPC	FTA New Starts Analysis (Schmitt, 2007), Sensitivity to gas price change in 2008	In use
Atlanta (ARC) Model	Parsons Brinckerhoff (2006); Rousseau (2012)	Utility-based	Simultaneous determination: Choice of three day patterns (mandatory/non-mandatory/at-home) for all household members	Yes	Yes	For ARC	Cloud computing platform for parter agencies (Rousseau, 2012)	In use
Bay Area MTC Model	Davidson et al. (2010)	Utility-based	Simultaneous determination: Choice of three day patterns (mandatory/non-mandatory/at-home) for all household members	Yes	Yes	For Bay Area MTC	Regional transportation plan (Metropolitan Transportation Commission, 2011b), Emission reduction strategies, Regional express lanes (Metropolitan Transportation Commission, 2011a), Transit Sustainability Project	In use
			5. Econometric Models with Activity-scheduling Process					
PCATS (FAMOS)	Kitamura and Fujii (1998); Kitamura et al. (1998); Pendyala et al. (2005)	Utility-based, time-space prism	Sequenced activities with purpose, location, mode of travel and duration	Yes	Yes	For FDOT	Academic case study: Evaluation of TDM strategies for CO2 emission reduction (Kitamura et al., 1998)	Developed in laboratory
CEMDAP	Bhat et al. (2004); Pinjari et al. (2006)	Utility-based	Sequenced tours with purpose (determined through an generation-allocation-scheduling process)	Yes	Yes	For NCTCOG and SCAG	Comparison of CEMDAP and DFWRTM 4-step method (Mirzaei and Eluru, 2007), Modeling traffic emissions with fine resolution (Isbell and Goulias, 2014)	In use
			6. Hybrid Models					
TASHA	Miller and Roorda (2003); Roorda et al. (2005); Roorda and Miller (2005); Roorda et al. (2008); Miller (2014)	Rule-based, utility-based	Activity agenda with activity purpose, start time and duration	Yes	Yes	For the City of Toronto	Urban form changes on green gas emission in the Greater Toronto Area, a wide variety of major rail transit investment options for the City of Toronto (Miller, 2015)	In use
ADAPTS	Auld and Mohammadian (2009, 2012)	Discrete event simulation, utility-based	Sequenced activities with purpose, location, mode of travel and duration	Yes	Yes	N/A	Academic case study on policy sensitivity: the teleworking scenario (Auld and Mohammadian, 2012)	Developed in laboratory

Table 2.1: Various activity-based travel demand modeling frameworks

2.2 Constraints-based Prototypes

Constraints-based prototypes are built directly on the seminal research of activity-participation under time-space constraints and they are the first type of activity-based models. As an example, Lenntorp (1977) implemented Hägerstrand (1970)'s approach and developed a model that calculates the total number of feasible paths in time-space prism given an activity program (set of activities with durations), transportation network, and activity distributions in time and space. The model was then evolved into a constraints-based prototype CARLA (Jones et al., 1983). Those models are called prototypes because they are not representation of contemporary activity-based models that intend to generate activity-travel patterns. Rather, they were used to determine the feasibility of given activity-travel patterns under temporal-spatial constraints and reschedule the patterns using ad-hoc behavioral assumptions.

Combinatorial Algorithm for Rescheduling Lists of Activities (CARLA) examines the feasibility of external activity programs and adjusts them based on the provided temporal-spatial constraints. The output activity schedule is based on the external input of activity programs and policies. When an external activity program is in conflict with policies, CARLA reacts to policies through schedule adjustment (by solving the Traveling Salesman Problem) and output the feasible pattern that is most similar to the current one. The rules to determine similarity are either tested in the algorithm or fed as input. While there are many reactions to policies, such as re-timing, activity cancellation, activity relocation, travel mode shift, etc., only re-timing (by altering duration, start and end time of activities) is accounted in CARLA. The algorithm was tested once in the Burford School study where CARLA was able to generate the identical schedule for 65 percent of 62 pupils after implementing a policy that set the school hours 30 min forward.

Constraints-based models deal with overall activity-travel demand comprehensively and use time-space prism to constrain the selection of activity patterns for people. However, there are a few issues with the classical time-space prism proposed by Hägerstrand. Firstly, the question of generating alternative activity patterns as input before determining the feasible ones is out of the reach of constraints-based prototypes. Secondly, the schedule

adjustment resembles little behavioral foundation and mechanism. Thirdly, the classical time-space prism takes a deterministic view of activity-travel geographic and is unable to deal with the uncertainty in travel and activity participation. In spite of those limitations, the constraints-based prototypes are still worth mentioning as the concept of time-space prism has been integrated to many contemporary activity-based models.

2.3 Rule-based Approach

Rule based approach, often referred to as Computational Process Models (CPM) in early literature (Gärling et al., 1994), uses heuristic production rules (if A then B) to mimic the underlying decision-making process and make decisions about activity participation and travel. Although production systems were first proposed by Post (1943) as a general computational mechanism, the later extension of the system has two diverse paths: The psychological modeling efforts that aim to mimic human decision-making process (Newell and Simon, 1972) and the performance-oriented AI systems that aim to create error-free procedures for certain tasks (see for example, Feigenbaum et al., 1970). The first path gave rise to the application of production systems in activity-based travel demand modeling.

STARCHILD (Recker et al., 1986a,b) is the first of its kind and is considered as the first rule-based activity-based modeling framework, although its application is restricted to research purposes. It works in three stages. First of all, individual activity programs are developed in reflection of activity needs and desire along with household interactions and environmental constraints. Then, activity pattern choice sets are enumerated and generated based on CARLA. In the second step, similar patterns are further grouped together with various production rules, resulting a small choice set. At last, a Logit model is applied for pattern choice. The framework was tested based on activity data from Orange County and Portland. It was then extended in Recker (1995) to model the household activity-travel patterns as a mathematical programming problem (HAPP). To date, STARCHILD has not been operationalized largely due to the fact that it requires data with temporal, spatial and inter-personal constraints, which is still not available nowadays.

SCHEDULER (Gärling et al., 1989) is a conceptual framework of activity-based activity scheduling process. In the framework, short-term activities are selected and sequenced based on their priority and constraints from priori commitments. The locations of these short-term activities are determined based on a heuristic procedure to minimize travel distance. The conceptual framework was operationalized in Golledge et al. (1994) with a GIS interface for predicting activity patterns of commuters before/after introducing tele-commuting to those who participated voluntarily in a tele-commuting program organized among state employees in Sacramento County (Kitamura et al., 1990).

SMASH (Ettema et al., 1993, 1996) is a rule-based model with a step-wise activity scheduling process. It predicts daily activity patterns as an ordered list of activities with destination, mode, route choice, time of day and choice of companion. It iteratively builds up the pattern from an empty schedule under the assumption that activity scheduling is a step-wise (stop and go) decision-making process. The stop and go process allows a decision maker to add, delete or reschedule an activity or terminate the scheduling process, based on the current schedule. To make a decision on the four actions, a Nested Logit model is applied.

AMOS (Kitamura et al., 1993; Pendyala et al., 1998) is a key component for Sequenced Activity Mobility Simulator (SAMS) (Kitamura et al., 1996). The key concept of AMOS is adaption by trial-and-error and its four modules work in an iterative procedure. First of all, a set of observed baseline patterns (including activity purposes, frequencies, priority list, timing, duration and location) are extracted from survey and synthesized, searching for possible adaptions. The response option generator then generates and prioritizes options with a calibrated neural network (Pendyala et al., 1995) before they are fed to the activity travel adjuster where each one is simulated individually. Finally an evaluation routine will determine the selected option.

The one aspect that distinguishes AMOS from the frameworks mentioned above is that it is the first applicable activity-based models. In fact, AMOS (along with SAMS) was implemented for Washington, D.C. under the supervision of Metropolitan Washington Council of Governments (MWCOG), and applied to evaluate the performance of various travel demand management (TDM) strategies, such as parking pricing, improved bicycle/pedestrian facilities

and congestion pricing (RDC Inc., 1995). Although AMOS heavily relies on the baseline activity-travel patterns (from household travel survey) as input, it is nevertheless the first activity-based model implemented and utilized by regional MPO.

For STARCHILD, SCHEDULER, SMASH and AMOS, the majority of decision-making heuristics are ad-hoc, with the exception that AMOS applies a neural network model estimated from the survey data. However, all of them are developed as incomplete prototypes and rely on exogenous forecast of important dimensions of activity-travel decisions. On the contrary, the following model, ALBATROSS, uses observed data to endogenously derive all the rules or heuristics.

ALBATROSS (Arentze et al., 1999; Arentze and Timmermans, 2000, 2004) is a household activity-based modeling framework. It is built onto a scheduling module with models connected by conditional rules. The input to the model is a skeleton of schedule with fixed activities to be conducted along with a list of different types of constraints. The scheduling process will add flexible activities. The output of the scheduling module will be a list of unordered fixed and flexible activities. In the next step, the activities will be sequenced and mode, destination, time-of-day will be modeled using decision trees (derived using observed data) after taking all kinds of constraints into consideration.

ALBATROSS was initially developed experimentally for the Dutch Ministry of Transportation, Public Works and Water Management, in an attempt to explore the possibility of the rule-based approach and develop a travel demand model for policy impact analysis. A case study was carried out for the Rotterdam region, confirming the potential of the rule-based approach. Since then, ALBATROSS has been evolving and improving (see for example, Arentze and Timmermans, 2007 and Anggraini et al., 2007). In line with the improvement, ALBATROSS has been integrated into an integrated activity-based framework FEATHERS (Bellemans et al., 2010), which has been implemented for Flanders, Belgium and applied to simulate vehicle emission, air pollution and energy consumption (see for example, Beckx et al., 2009; Yang et al., 2010).

While there are some more recent developments using rule-based approach, such as TASHA and ADAPTS, they will be reviewed later as examples of hybrid models since both production

rules and econometric models are blended together in their frameworks.

Among the frameworks/models reviewed in this section, none of them has been put into operation to replace the traditional four-step models, although there are some case studies on the application of these frameworks found in the literature. There are several possible reasons. Firstly, some early rule-based frameworks are quite conceptual and require input data (such as temporal and spatial constraints) that is difficult to collect, which hinders the implementation of these frameworks. Secondly, most of these frameworks are incomplete and require observed activity programs, activity episodes as input and focus on only the scheduling and sequencing of those patterns. Thirdly, the scheduling process of the reviewed frameworks are mostly hardcoded with pre-determined ad-hoc production rules to mimic the human behavior of activity scheduling (the only exception is ALBATROSS), which makes it less convincing when those frameworks are to be used to assess TDM strategies.

2.4 Econometric Approach

Activity-based models that adopt econometric approach mainly find their theory foundation in Random Utility Maximization (RUM) theory where individuals are rational in making decisions and complete information on all possible alternatives are available. Advances in discrete choice models pave the road for the application of RUM in activity/travel related decision-makings (Ben-Akiva and Lerman, 1985). Early application of discrete choice models in the context of activity-based modeling focuses on single-faceted decisions, such as mode choice[1]. An early extension of single-faceted model to multi-faceted activity-travel patterns can be found in Adler and Ben-Akiva (1979). In their work, alternative activity-travel patterns are characterized by many facets, including trip chaining, tour characteristics, travel modes, travel time and destination characteristics. Discrete choice models are used to choose from a set of such defined patterns. The choice of activity-travel patterns is modeled as a single discrete decision, where the choice set is taken from observed patterns in survey data.

[1] As indicated in Table 2.1, some rule-based models have applied discrete choice models in their frameworks. However, those models are standalone and single-faceted. Similarly, standalone discrete choice models are also used in trip-based models for mode choice. Those applications are mainly for the convenience of the modeling and less systematic compared to the application of discrete choice models reviewed in this section.

The idea of modeling the choice of activity-travel patterns as a single model embeds the constraints and interactions of trips and tours in the choice process.

To date, activity-based travel demand modeling frameworks that adopt econometric approach are among the most comprehensive and well-developed ones. To be clear and comprehensive, activity-based models adopting econometric approach reviewed subsequently are classified into two groups based on how daily activity patterns are defined, represented and modeled. Models in the first group follows the philosophy of Adler and Ben-Akiva (1979). However, given the large number of activity-travel patterns in the choice set, choice facets are isolated and modeled separately while a hierarchical structure is applied to nest all choice facets. The concept of daily activity patterns in this group can be defined as reflection and abstraction of individual activity participation on daily basis, and is placed at the top of the hierarchy to constrain the decisions of other facets, or dimensions. Based on how daily activity patterns are coordinated within households, the first group can be further divided into two sub-groups: models with fully individual-based patterns and models with household-coordinated patterns. In the second group, instead of modeling daily activity patterns at the top level, the models assume an activity-scheduling process: a sequential decision process to generate daily activity-travel patterns as output. Section 2.4.1 to 2.4.3 will review models with individual daily activity patterns, coordinated daily activity patterns and activity-scheduling process respectively.

2.4.1 Models with individual daily activity patterns

Ben-Akiva and Bowman (1995) estimated a utility-based hierarchical choice model of daily activity-travel patterns comprising a Nested Logit model of daily activity pattern choices (i.e., purposes, priorities and structure of daily activities) and travel-related choices (mode choice, destination choice, number of stops in tours, and departure time from home and from the primary activity in tours), which is the first working prototype of a full-day model that includes and integrates the activity participation and travel decisions spanning a day, and also includes the dimensions of destination, mode and timing of the associated travel. The approach was then summarized as Day Activity Schedule model (Bowman, 1998) and has

been applied in several cities and regions in the United States.

The Portland Metro Model (Bowman et al., 1998; Bowman and Ben-Akiva, 2001) is the first large-scale operational activity-based modeling framework based on Bowman's approach. The nested hierarchical structure is applied to sequentially model activity-travel decisions at day pattern, tour and intermediate stop levels (Bradley et al., 1998). From the top down, decisions of lower levels are conditional on the outcome of higher levels. From the bottom up, the accessibility measurements are fed upwards to reflect their influence. The Metro Model is the first activity-based model with individual daily activity patterns. The daily activity patterns are defined as a skeleton or abstraction of a detailed plan (see Figure2.2) and modeled separately for each individual at the top level. With this definition, trip chaining within a tour is represented as a sequence of activities.

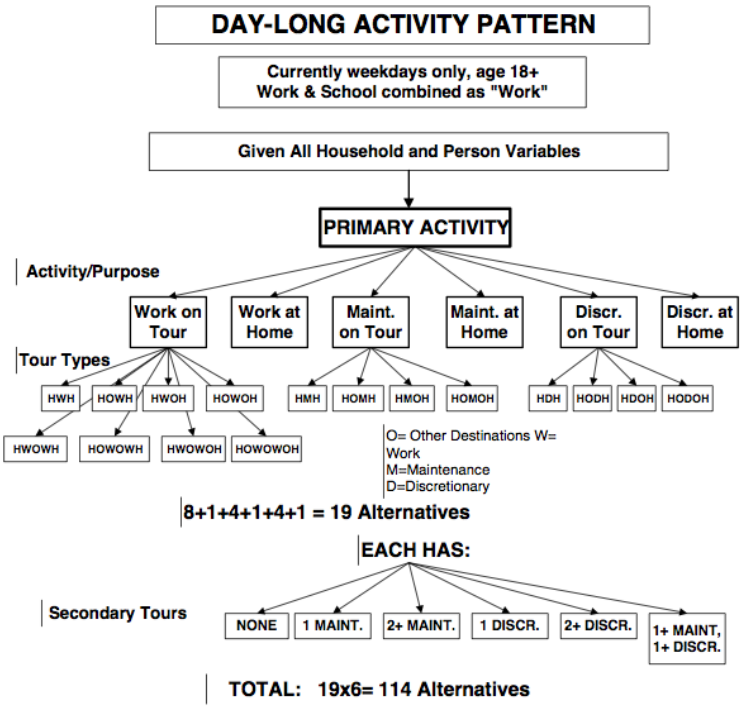


Figure 2.2: Definition and choice set of daily activity patterns for the Portland Metro Model (source: Bradley et al., 1998)

The Metro Model was developed during 1996 and 1997. It has not been used actively by Portland Metro as the results were found not satisfactory (Vanasse Hangen Brustlin, Inc., 2006) probably due to the fact that the model was not calibrated. Besides, back then, the results from activity-based models were merely used to support the re-estimation of

trip-based models. Having said that, however, the implementation of the Portland Metro Model is remarkable as it explores the possibility of developing an activity-based modeling framework without imposing heavy burdens on data re-collection. Moreover, it marks that MPOs and modeling practitioners are seriously considering replacing the state-of-the-practice trip-based models with activity-based ones and the process has speed up since then.

San Francisco County Transportation Authority (SFCTA) Model (Bradley et al., 2001; Cambridge Systematics, Inc., 2002; Freedman et al., 2006; Outwater and Charlton, 2006) keeps the basic design as in Portland with its own innovation and breakthrough. The full-day activity pattern model is improved from the Portland one with more primary, second tour combinations, resulting a more complex choice structure. Moreover, the choice sets are adjusted for each of the four considered person types and a full day pattern model is estimated individually for each person type, resulting a more diverse and realistic model response.

The SFCTA Model was estimated mainly using household survey data collected in 1990 and data from 1990 Census was used for model calibration. As noticed in Outwater and Charlton (2006), SFCTA Model is the first activity-based model that goes through a calibration process and is put into active use after the development phase. It was used for an equity analysis to estimate impacts on mobility and accessibility for different population segments to support development of a countryside transportation plan (see Castiglione et al., 2006). It was used in two major transit projects, where in one of them, New Starts, SFCTA Model was used to evaluate the benefits of the proposed New Central Subway project in downtown San Francisco. This project is also the first application of an activity-based travel demand model in the United States to a major infrastructure project in support of a submission to the Federal Transit Administration (FTA) for project funding (see Freedman et al., 2006). The model was re-estimated using data from year 2000 and enhanced to support the modeling of more policies, such as congestion pricing (San Francisco County Transportation Authority, 2010), development of bike infrastructure (Hood et al., 2011), evaluation of transit crowding (Zorn et al., 2012), etc. It is still in active use for San Francisco.

The SACSIM Model developed for Sacramento (California) Area Council of Governments

(SACOG) (Bradley et al., 2007, 2010; DKS Associate et al., 2012) is another model in active use. It reformulates the individual activity patterns as the occurrence of tours (0 or 1+) and intermediate stops (0 or 1+) for various activity purposes with a second model to predict the exact number of tours for various activity purposes. The overall framework is clearly divided into three layers: day pattern level, tour level and intermediate stop level. Vertical integration with accessibility measurements is applied to nest models of different layers. It is the first model with parcel-level spatial resolution and 30-min temporal resolution (Bradley et al., 2010). The activity-based modeling framework of SACSIM has been coded into a software, DaySim, to simulate resident daily activity patterns. SACSIM has also been used to prepare land use and transportation analysis for several transportation or development projects (refer to Table 2.1).

The development of SACSIM Model with DaySim reveals the possibility of rapid development and deployment of activity-based models. Most recent efforts for the development and deployment of activity-based models with individual daily activity patterns including Denver (DRCOG) Model (Sabina and Rossi, 2006; Cambridge Systematics, Inc., 2010), Seattle (PSRC) Model (Nichols et al., 2014), Jacksonville (NFTPO) Model(Lawe, 2010) and Houston (HGAC) Model (Rossi, 2012; Rossi et al., 2013) are the variations of the Sacramento Model. However, different development strategies are adopted: while Denver and Houston developed their own modeling frameworks and software platforms (see Rossi, 2012), Seattle and Jacksonville, on the other hand, directly interfaced with DaySim for rapid implementation.

Besides activity-based models developed for metropolitan area, two states in the United States, Oregon and Ohio[2], are found to have included an activity-based model in their statewide land use and transportation demand models to predict short-distance person travel (SDT) (see Parsons Brinckerhoff et al., 2010 and Parsons Brinckerhoff, 2010). While the overall framework of the SDT module is a variation of Bowman's approach with a series of hierarchically nested models, the individual daily activity patterns are uniquely defined as a sequence (string) of activities and modeled with a Logit choice model. The activity sequences contain information on the number and sequence of tours, sub-tours and intermediate stops.

[2]The statewide model for Ohio was built based on the Oregon's.

The detailed representation of number and sequence of tours results in a large choice set. The SDT model for Oregon contains over 3,000 observed alternatives. In the Ohio Model, the number is brought down to around 100 by generalizing patterns of low occurrence.

2.4.2 Models with coordinated daily activity patterns

In the last section, the reviewed activity-based modeling implementations consider full individual day patterns and treat individuals as the unit for activity-travel decision-makings. As a result, the interactions between household members are not well represented. Into 2000s, such interactions have drawn substantial attention and been emphasized in a number of studies. Intra-household interactions have many facets, to name some:

- Household activity/task allocation, see for example, Bhat and Pendyala (2005) and Srinivasan and Bhat (2005).
- Joint activity participation, see for example, Srinivasan and Bhat (2008).
- Vehicle allocation, see for example, Golob et al. (1996), Hunt and Petersen (2004), Petersen and Vovsha (2005) and Petersen and Vovsha (2006b).
- Trip escorting, see for example, Vovsha et al. (2003) and Roorda et al. (2006).
- Some researchers have already stepped further and taken social interaction into consideration, see for example, Walker et al. (2011) and Ronald et al. (2012).

In terms of incorporating intra-household interactions into activity-based models, Wen and Koppelman (2000) proposed the first conceptual framework based on discrete choice models. Generally, the framework is under the track of utility maximization and models household activity-travel patterns conditioning on the household's maintenance needs. To date, there have been a number of empirical models that adopt such philosophy and apply an enhanced version of the full individual day pattern approach by accommodating intra-household interactions in activity-travel engagement.

New York Best Practice Model (NYBPM) (Parsons Brinckerhoff, 2005b and Chiao et al., 2006) is the first operational activity-based modeling framework that uses the contemporary

conceptual framework of daily activity patterns accounting for intra-household interactions between household members and constraints on travel in terms of both time and space. In particular, intra-household interactions are taken into consideration in two levels: activity allocation, joint activity participation, trip escort in the Journey Frequency Model and vehicle allocation, shared ride at mode/destination level. At day pattern level, the Journey Frequency Model is illustrated in Figure 2.3. It puts household members into the same space with time, space and member constraints. There are 3 person types and 6 journey purposes, resulting in 13 journey frequency models with different priorities[3]. Although the household structure is pre-defined and static, it is nevertheless the first operational daily activity pattern model that explicitly considers and models intra-household interactions.

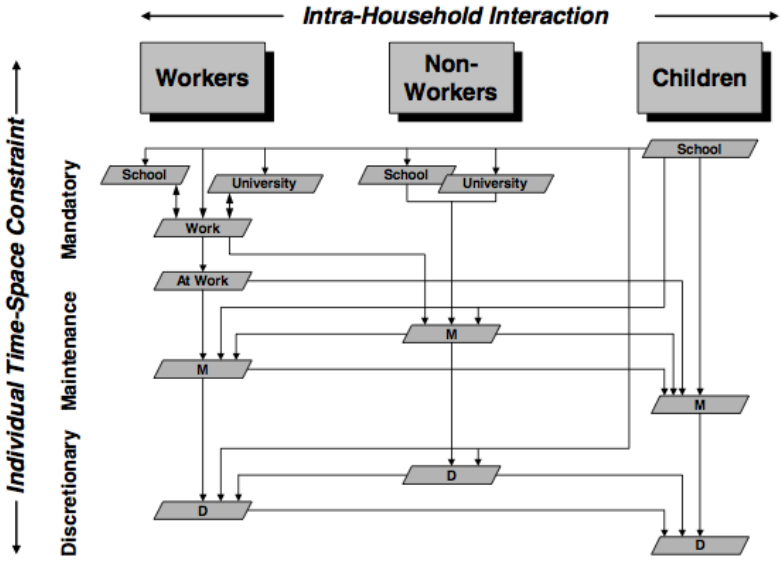


Figure 2.3: The Journey Frequency Model in NYBPM with intra-household interactions (source: Parsons Brinckerhoff, 2005b)

NYBPM was estimated to represent the 1996 base year condition. No regional model of the geographic scale and functional coverage as with NYBPM has ever been successfully implemented in this very large metropolitan area. Consequently, NYBPM keeps evolving to meet the refinement of detailed specifications to this region. The model was re-estimated

[3]In Figure 2.3, an arrow points to a model of lower priority. A model with higher priority is modeled earlier. For example, journeys to school for children are modeled first. The arrow will also provide the results of previous model to the next one, with the expectation that the decisions made will have influence on the subsequent models.

as the 2002 base year model. Finished in 2005, the new model incorporated more recent modeling efforts in modeling journey frequency, mode, destination and time-of-day choice. Since its preview in 2002, NYBPM has been used to support air quality conformity analysis and a series of single-mode and multimodal transportation studies (refer to Table 2.1).

With the successful development, deployment and application of NYBPM, several regions started their own activity-based models, which directly leads to a family of activity-based modeling frameworks called the Coordinated Travel and Regional Activity Modeling Platform (CT-RAMP) (see Davidson et al., 2010 and Vovsha et al., 2011). Among all the characteristics of the CT-RAMP, the most distinguished one is the explicit representation of intra-household interactions across many activity and travel dimensions (Vovsha et al., 2004; Bradley and Vovsha, 2005) and the focus here is on the coordinated daily activity patterns.

The Columbus Model developed for Mid-Ohio Regional Planning Commission (MORPC) is the first activity-based model of the CT-RAMP (Parsons Brinckerhoff, 2005a; Anderson and Jiang, 2013). The overall framework is enhanced based on NYBPM with better time-of-day resolution. Similar to NYBPM, the Columbus Model defines daily activity patterns as the number of tours for three activity types (mandatory, non-mandatory and at-home). The daily activity patterns are sequentially modeled for all household members according to a pre-specified order of processing.

The Atlanta Model developed for Atlanta Regional Commission (ARC) (Parsons Brinckerhoff, 2006; Rousseau, 2012) generally adapts the design of the MORPC Model. A major distinction is found in terms of how daily activity patterns are modeled. In the ARC Model, three basic daily activity patterns are defined for each individual (mandatory/non-mandatory/at-home), the day pattern model will simultaneously determine the choice for all household members taking into account all possible cross-impacts of different household members on each other (Bradley and Vovsha, 2005). The simultaneous approach for the modeling of daily activity patterns is also adopted in the Bay Area MTC Model, which is developed in parallel with the ARC Model (Davidson et al., 2010).

More recent development and deployment has taken place in San Diego (SANDAG), Phoenix (MAG), Chicago (CMAP) and Miami (SERPM). While all of them belong to the CT-RAMP

family, special features are incorporated to address the regional concerns in respect of their own focus.

For econometric activity-based models with individual and coordinated daily activity patterns, with the experience gained through intense research, collaboration between government sectors and consultancy, and continuous assessment and evaluation, they have become the most popular regime to activity-based perspective of travel demand modeling. Although a model with coordinated daily activity patterns is more general in terms of explicit consideration of household interactions, to adopt it or not is more of a practical concern. For example, the CT-RAMP family of models consider the household interactions at day pattern level with a cost of either pre-assuming an order to process all household members (which increases the number of models) or simplifying the choice set for daily activity patterns (in order to simultaneously model the choice of all household members with discrete choice analysis). Moreover, the study area may be lack of activity-based household travel survey data considering intra-household interactions explicitly throughout the survey.

2.4.3 Models with activity-scheduling process

So far, the reviewed econometric activity-based models all adopt a hierarchical framework where daily activity patterns are modeled as an abstraction of daily activity participation at the top level, followed by subsequent decisions (such as mode, destination) adding details to the skeleton in order to generate activity-travel plans. Alternatively, there are a few frameworks that adopt a sequential decision process to generate daily activity-travel plans from the beginning. In such frameworks, the daily activity pattern is not found at the beginning of the modeling process but at the end of it as output.

PCATS (The Prism-Constrained Activity Travel Simulator) is a comprehensive econometric model that uses temporal-spatial and other constraints to model activity engagement decisions sequentially, conditioned upon past activity engagement (Kitamura and Fujii, 1998). It first divides the day into open and blocked periods. While blocked periods are occupied by fixed activities (such as work, education), open periods are filled with flexible activities generated sequentially in consideration of time-space prism. For each activity, activity type,

destination, mode and duration are determined in sequence by discrete choice models. When generating the sequence of activities, it is assumed that the individual's activity decision is dependent on the past, but not dependent on the future, except for the presence of prism constraints. PCATS was once adopted in an academic case study to evaluate the effectiveness of several TDM strategies for CO_2 emissions reduction in Kyoto, Japan (Kitamura et al., 1998). Later, PCATS was re-estimated and interfaced into the Florida Activity Mobility Simulator (FAMOS) as the core module to simulate individuals' activity and travel in urban space (Pendyala et al., 2005).

CEMDAP (Bhat et al., 2004; Pinjari et al., 2006) is another framework characterized by its unique activity generation-allocation-scheduling process and continuous representation of time through hazard-based duration. The decision for mandatory activities for workers (students), allocation of household maintenance activities, escorting activities, and other flexible activity purposes are carried out with the generation-allocation system. Then, two scheduling process are used to generate activity-travel plans for workers and non-workers respectively. For non-workers, the tours and associated attributes (number of stops, duration, location, mode and stop-level information) are sequentially scheduled. For workers, the daily activity-travel plan is divided into five periods: before work (BH), home-to-work (HW), work-based (WB), work-to-home (WH), after-work (AW). For each period, a scheduling process is adopted. First of all, for the WH period, mode, number of stops and activity duration are modeled. Then the HW period is modeled similarly. If the worker has also decided to participate in other activity purposes, the number of tours to be undertaken during each of the AW, WB and BW periods is modeled followed by a scheduling process similar to non-workers to model subsequent decisions at tour and intermediate stop level.

CEMDAP was initially developed in the research community. The framework was then adapted in the Simulator of Activities, Greenhouse Emissions, Networks, and Travel (SimAGENT) developed for Southern California Association of Governments (SCAG) (Goulias et al., 2012)

2.5 Hybrid Approach

There are certain activity-based models that use more than one concept of computational process, production rules, and utility maximization to handle the different behavioral and temporal-spatial constraints in modeling different facets of activity-travel patterns. It allows the modeling of different condiments of the overall activity-based framework with different types of techniques.

Travel/Activity Scheduler for Household Agents (TASHA) (Miller and Roorda, 2003; Roorda et al., 2008) is an example of the hybrid approach. Several modeling techniques are combined: First, activity episodes and associated attributes, such as start time and activity duration are generated by drawing samples from empirical distributions observed in household survey. Second, person schedules are constructed from scratch based on the concept of projects and production rules are used for scheduling and rescheduling to resolve the scheduling conflicts (Roorda and Miller, 2005). The scheduling process and associated rules are verified in a tailored survey where the actual scheduling process are observed (Roorda et al., 2005). Third, location choice as well as mode choice (including individual/joint activity mode choice, vehicle allocation and ride share determination) is carried out with discrete choice analysis. A validation study was carried out in Roorda et al. (2008) mainly for the validation of the activity generation and scheduling process against the activity-travel behavior of the base year and forecast year.

Another example is the Agent-based Dynamic Activity Planning and Travel Scheduling (ADAPTS) Model (Auld and Mohammadian, 2009, 2012), where computational process, discrete event microsimulation and econometric models are blended together. For both rule-based models and econometric models previously reviewed, regardless of modeling daily activity patterns as a sequential scheduling process or as the outcome of an econometric model, some a priori planning orders for specifying the activity attributes, such as mode, destination, time of day, need to be pre-assumed. ADAPTS, on the other hand, by accepting the claim in Doherty et al. (2004) that those assumptions are unrealistic, adopts the framework in which activity planning events are treated as individual discrete events and an activity schedule is created and modified over time, and that the individual attributes of each

activity (for example, destination and mode) are not necessarily planned in any given order. The framework allows the model to explicitly represent planning dynamics, in which each attribute is planned at a discrete time and is conditioned on the previously planned attributes and schedule. In turn, these decisions constrain future attribute planning and scheduling decisions, creating a truly dynamic model of activity planning and scheduling.

2.6 Summary

Modeling travel demand from an activity-based perspective is originated from the principle that the need for travel is derived from the need for activity participation. Follow the basic principle, different groups of activity-based models have emerged in the literature to incorporate the principle from various viewpoints. The last two decades have witnessed the evolving of activity-based models from conceptual ones to operational ones and the paradigm shift from the traditional four-step method to the activity-based approach. Currently, many cities and regions in North America (mainly the United States), Europe and Japan have supported the development and evaluation of activity-based travel demand models and decided (1) to replace trip-based models with activity-based ones or (2) to use both models in parallel or (3) not to proceed and stick to trip-based models. The experience gained in the process and the sharing of these experience greatly speed up the model development and institutional training process, which in turn attracts more cities and regions to seek the potential benefits brought by activity-based travel demand models.

However, from the reviewed literature, it is concluded that the methodologies used for various activity-based modeling frameworks are quite different and far from converging to a well-accepted common ground. For example, rule-based approach and econometric approach are different in terms of modeling human behavior and the assumptions embedded. Even for models with econometric approach, the modeling of daily activity patterns is still model-specific. It is expected that the evolution of activity-based models will continue in the literature and greater behavioral realism will be found and balanced with modeling performance to better represent the response to numerous transportation-related policies. Moreover, although a common ground for activity-based models is appreciated for fast

implementation and deployment of operational activity-based models, the methodologies and features of activity-based models will continuously evolve in order to be consistent with emerging theoretical achievements and tailored to specific regions with distinct policy focus and practical considerations.

As an expansion of this summary, several concerns raised during the literature review are discussed below.

Daily activity patterns It is observed clearly from the reviewed frameworks that daily activity patterns are treated in two ways. First, daily activity patterns are modeled as a decision process with underlying behavioral mechanism. Usually, either heuristics or production rules are used to represent the decision-making process in terms of activity generation/scheduling. Second, daily activity patterns are defined as the abstraction of daily activity participation behavior and modeled as the outcome of one of several choice models. While all rule-based models feature the first treatment, econometric models reviewed in Section 2.4.1 and 2.4.2 feature the second treatment. Although it is claimed that the first treatment is able to explicitly model the decision process, the behavioral mechanisms that lead to the decision process are not validated. Among the reviewed frameworks that feature the first treatment of daily activity patterns, only ALBATROSS, TASHA and ADAPTS are able to derive the heuristics, rules that mimic the underlying behavioral mechanisms from different perspectives (far from converging to a common theory foundation). However, their superiority over the econometric ones is not clear in terms of replicating base year activity-travel patterns and predicting the response to various TDM strategies. On the other hand, the second treatment has been proven in practice to be amendable to the development and deployment of operational activity-based travel demand models. In fact, the models reviewed in Section 2.4.1 and 2.4.2 represent the majority of activity-based models currently in operation.

Demand-supply integration The ultimate purpose for travel demand modeling is to measure the traffic flow and assess the level of service of transportation systems under different planning scenarios. In the traditional four-step method, the traffic flow is calculated

with an equilibrium-based approach in the final step after aggregate travel demand is derived from the first three steps. Generally speaking, the efforts to improve the four-step method have been focused on moving from aggregate into disaggregate framework to ensure a better behavioral realism and model resolution. However, the research efforts that attempt to improve the traditional four-step model have been disjointed historically (Vovsha, 2009). On the supply side, most of the early efforts are made to the replacement of equilibrium-based assignment of traffic with simulation-based dynamic traffic assignment (DTA) process where the travel demand is exogenous and not explicated modeled. Usually, a DTA process will replace the final traffic assignment in the four-step method and the aggregate measure of travel demand represented as (time-dependent) OD metrics will be fed into the DTA process. On the demand side, the research efforts are focused on replacing the trip-based approach in the four-step method with more sophisticated activity-based travel demand modeling frameworks reviewed in this chapter. However, in practice, the disaggregate travel demand derived from many operational activity-based models is often aggregated into time-dependent OD matrices for traffic assignment. The disaggregate representation of individual travelers and their decisions gained through the application of activity-based models are lost in the integrated OD matrices and the explicit representation of rescheduling and re-routing decisions, as well as the individual-based performance measures are not allowed (Balmer, 2007). It was not until recently that efforts seeking an integration of dynamic and disaggregate travel demand and supply within the same framework and a consistent individual-based representation of travelers throughout the whole process started to emerge (see the case examples in Castiglione et al., 2015). For a newly deployed activity-based model, to better facilitate the policy decision-makings in an inventory-based planning scenario, it is essential to have an integrated demand and mobility framework such that the dynamic interaction between travel demand and transportation systems can be modeled and reflected.

Data Activity-based models require a large amount of data for model estimation, calibration and validation. For some early activity-based models, although the conceptual frameworks are elegant, the critical question that what sort of data is necessary and sufficient in developing those frameworks has remained largely unanswered. For example, early rule-based models

require temporal-spatial constraints that are not measured in surveys, not to mention that the heuristics and rules are not validated due to the lack of data that can reveal the behavioral process. In recent years, with the deployment of operational activity-based models, the issue has been partially addressed in practice. Essentially, the data required to develop an activity-based model is not significantly different from the data required to develop a trip-based model (including household travel survey, economic and demographic information about the spatial distribution of employment and households, and representation of transportation networks). In fact, it has been demonstrated that the same household surveys used to develop trip-based models can be used to develop activity-based models. Although it requires that all the survey data to be more consistent internally across all the trips for all individuals in each household, the continuity of data collection efforts during the transition from trip-based models to activity-based ones involves little data re-collection and allows the practitioners to switch between modeling paradigms with greater flexibility. However, it remains a challenge that how emerging data sources and surveys enabled by technology (for example, data from multi-day surveys and vehicle trips captured in GPS data) can contribute to the development of operational activity-based models.

Chapter 3

Travel Time Modeling with GPS and Household Travel Survey Data

Historical time-dependent travel time is one of many essential data sources needed for the development of activity-based travel demand models. Transportation planning agencies all over the world usually maintain travel time matrices (or network skims) for a small number of time windows. In order to predict travelers' response to congestion mitigation strategies, which is one of the key motivations behind the rising of activity-based demand models, it becomes essential to develop time of day choice models that require travel time estimates at a finer time resolution.

In this chapter, we develop regression models to relate travel times collected from taxi GPS data to the network travel times and compare the results to a similar model estimated with household travel survey data. The rationale behind this procedure is to develop a formula that allows the calculation of travel time for any origin-destination pair and for any time of the day, given the network travel times for three time periods (AM peak, PM peak, and off-peak). The two data sources, survey and GPS data, are compared based on descriptive statistics and by plotting the variation of the predicted speed by time of day. Statistical tests are performed to investigate whether the two data sources can be pooled together.

The test results indicate that though there are significant differences in the estimated coefficients, which do not vary across time of day (for example, coefficients of distance, and central business district indicators), the two data sources exhibit comparable profiles of time-of-day variation in speed up to certain scales.

3.1 Introduction

Urban transportation planning agencies usually maintain travel time matrices (or skims) for a small number of pre-determined time windows, such as AM peak, PM peak, and off-peak, generated after the calibration of volume-delay functions (based on measurements from e.g. floating car data) and the assignment of OD matrices to the network for a number of time periods. The transportation planning community is recognizing the need for models that can predict temporal patterns of travel at a finer time resolution to enable greater sensitivity to policies that aim at shifting time-of-travel patterns, such as congestion pricing. As a result, several time of day models have been developed in the last few years using time period intervals as small as half-hour (e.g., Abou-Zeid et al., 2006; Hess et al., 2007; Kristoffersson and Engelson, 2008; Popuri et al., 2008; Cambridge Systematics, Inc., 2010; DKS Associate et al., 2012). Such models require estimates of travel time by time of day (e.g., by half-hour) as input. Network travel time skims usually fail to provide such input because they are available only for a few aggregated time periods. Estimation models that use time categories as dummy variables (see for example, Fu and Rilett, 2000) are not enough. In Abou-Zeid et al. (2006) and Popuri et al. (2008), a method was proposed to relate through regression analysis the reported zonal travel time extracted from household travel surveys to network travel time skims, time of day, and land use variables. The method is capable of generating time-dependent zonal travel time as a continuous function of time.

Besides trips reported in household travel surveys, Global Positioning System (GPS) provides an alternative approach for collecting trip information. Firms have been using GPS to track commercial vehicles for a long time. As an example, in Walsh et al. (2008) and Rahmani et al. (2010), taxis equipped with GPS receivers were used as floating vehicles to measure the average speed and level of service of road networks in urban areas. However, no literature is found to have developed the mentioned regression model with trips collected by GPS. This research attempts to estimate a travel time regression model with taxi GPS data from Singapore and compare the results to a similar model estimated with Singapore household survey data to evaluate the potential of using GPS data for travel time prediction in transportation planning models.

Compared to self-reporting, GPS receivers can record time and location information with better accuracy. In Stopher et al. (2007), the authors assessed the accuracy of household travel surveys by using a GPS survey conducted simultaneously. It was found that for car trips, people tend to overestimate travel time by 20 and 40 percent for trips by car drivers and passengers, respectively.

In this study, since we use in-service taxi GPS data collected in Singapore, another issue is whether taxi trips are equivalent to car trips reported in a household travel survey. As shown in Miwa et al. (2008), in-service taxis travel faster than out-of-service taxis. Also taxi drivers are more aggressive in route choice, which could result in a shorter travel time. The difference between car trips reported in household travel surveys and taxi trips recorded by GPS should be taken into consideration when evaluating the estimation results of both models.

The next section describes the methodology including model specification and evaluation. Section 3.3 introduces the datasets used in the study and data processing steps. Section 3.4 shows estimation results as well as comparison of results estimated using the household survey dataset and taxi GPS dataset. Section 3.5 provides a summary and directions for future research.

3.2 Methodology

3.2.1 The model

In Singapore, network travel time matrices (network skims) for three time periods in the day (AM peak, PM peak, and off-peak hours) are available from a four-step travel demand model. We develop a regression model that relates reported travel time data from the Household Interview and Travel Survey (HITS) or travel time extracted from taxi GPS data to those network skims.

Before introducing the model specification, it is necessary to distinguish two travel times we estimate: (1) travel time given departure time from the origin, and (2) travel time given

arrival time at the destination. This distinction is needed in time of day model application where arrival time at and departure time from the primary activity of a tour are modeled jointly; travel times given arrival time at the primary activity are then needed for scheduling stops made before the primary activity, while travel times given departure time from the primary activity are needed for scheduling stops made after the primary activity (See Abou-Zeid et al., 2006 and Ben-Akiva and Abou-Zeid, 2013). In the rest of the section, we will use the terms departure-time-based model and arrival-time-based model to distinguish the two. These two models have the same specification yet are estimated independently. The final specification takes the form in Equation 3.1 where i denotes the i^{th} observation in a dataset.

$$\begin{aligned}\ln\left(\frac{V_i}{V_{\text{off-peak},\ i}}\right) = {} & \beta_0 + \beta_1(\text{CBD origin dummy})_i + \beta_2(\text{CBD destination dummy})_i \\ & + \beta_3\left[\ln(\text{distance})_i\right] \\ & + (\text{delay})_i\left\{\beta_4 + \sum_{k=1}^{M_1}\left[\beta_{2k+3}\sin\left(\frac{2k\pi t_i}{24}\right) + \beta_{2k+4}\cos\left(\frac{2k\pi t_i}{24}\right)\right]\right\} \\ & + \left\{\sum_{k=1}^{M_2}\left[\beta_{2k+2M_1+3}\sin\left(\frac{2k\pi t_i}{24}\right) + \beta_{2k+2M_1+4}\cos\left(\frac{2k\pi t_i}{24}\right)\right]\right\} + \varepsilon_i\end{aligned} \tag{3.1}$$

Instead of using travel time as dependent variable, we use speed, which incorporates the influence of distance on different OD pairs. Thus, our model could be used for all OD pairs. Trip travel speed (V_i) is calculated as the ratio of traveled distance divided by reported (or extracted in case of GPS) in-vehicle time, and off-peak speed ($V_{\text{off-peak},\ i}$) is calculated as off-peak zonal centroid distance divided by off-peak zonal travel time.

Origin and destination CBD dummies try to capture the influence of zonal geometric and economic characteristics on travel speed. We expect to observe negative coefficients for these two terms as worse traffic conditions are more frequent in CBD areas and will decrease the speed of trips that depart from or arrive at CBD areas. Trip distance also has an influence on travel speed; thus, $\ln(\text{distance})_i$ is introduced into the specification where distance is the minimum of AM peak, PM peak, and off-peak distance measured in kilometers. We expect comparatively larger speeds for long distance trips than for short distance trips, because in long distance trips there would be larger road segments without congestion or signal control.

The variable "delay" is defined as

$$\text{delay}_i = 1 - \frac{V_{\text{peak, } i}}{V_{\text{off-peak, } i}} \tag{3.2}$$

where $V_{\text{peak, } i}$ is the peak speed. Both peak speed and off-peak speed come from network skims. Peak speed is defined as the minimum of AM peak speed and PM peak speed. The delay variable captures the congestion effect on each OD pair. Since network peak speed is smaller than off-peak speed, the delay could take any value between 0 and 1. The larger the delay, the greater is the congestion for that OD pair.

Interacted with the first trigonometric function in the specification, the delay variable is projected to any time in a day. We use a trigonometric function because it is cyclical and continuous. t in the trigonometric function is trip departure time for departure-time-based model and arrival time for arrival-time-based model. The product of the delay variable and its coefficient given t (the sum of the terms in the first parentheses) is the contribution of delay to the dependent variable at t. Since a larger delay should result in a smaller ratio of travel speed to off-peak speed, the delay coefficient is expected to be negative for all t from 0 to 24. The first trigonometric function may also be called congestion-sensitive trigonometric function.

A second trigonometric function that does not interact with the delay variable is added to the model specification. It can be seen as a time dependent constant. Without including the second trigonometric function, it was found that from midnight to around 6 am, the delay coefficient became positive for both models estimated with HITS and GPS datasets, which is counterintuitive and indicates that travel time variation over time cannot be explained by only considering a delay variable. Therefore, a second trigonometric function is introduced to capture the variation that the delay variable fails to capture and test results show that this trigonometric function is significant for both HITS and GPS models. Since it does not interact with the delay variable, the second trigonometric function may be called congestion-free trigonometric function.

The truncation points of the trigonometric functions, M_1 and M_2 in the specification, need to be determined empirically. An appropriate combination of M_1 and M_2 will result in a

good fit of the model to the data as well as a reasonable time of day profile of the two trigonometric functions. Finally, ε_i in Equation 3.1 is a normally distributed error term.

This specification does not account for potential spatial correlation in the dataset, which could reduce the efficiency of the model. The correlation between the computed residuals from the ordinary least squares regression of Equation 3.1 was found to be in the order of 0.2 to 0.3 and significant. It is noted that with large samples, such as the one we use in this research, statistical tests may fail to reject the hypothesis that spatial correlation does not exist in the dataset. Since the parameter estimates produced by ordinary least squares regression are consistent and given the moderate level of spatial correlation, the results produced by ordinary least squares regression are sufficient for our purpose, which is to develop regression models to generate time-dependent travel time estimates.

3.2.2 Statistical tests

Since we estimate the model with two data sources, HITS dataset and taxi GPS data, it is interesting to know the "compatibility" of the two data sources. We investigate whether the data from these two sources can be pooled together. This can be done using the F-test where the unrestricted models are the ones estimated separately with taxi GPS data and HITS dataset and the restricted model is the one estimated with the combined dataset (taxi GPS and HITS data). The following restrictions are tested:

1. All the coefficients are restricted to be equal across the two datasets.
2. The coefficients of the congestion-sensitive and congestion-free trigonometric functions are restricted to be equal across the two datasets.
3. Similar to specification 2, but the coefficients of the trigonometric series of the HITS dataset are allowed to have a different scale from the coefficients of the trigonometric series of the taxi GPS data (whose scale is normalized to 1).
4. Similar to specification 3, but two different scale parameters are applied to the HITS data: one for the coefficients of the congestion-sensitive trigonometric series and another for the coefficients of the congestion-free trigonometric series.

5. In the final specification, four different scales are allowed in HITS: one for the congestion-sensitive sine functions, one for the congestion-sensitive cosine functions, one for the congestion-free sine functions, and the last one for the congestion-free cosine functions.

The idea behind the tests of specifications 2-5 is that even though the predicted speeds may be different between the GPS and HITS data, the profile of the temporal variation in speed may be similar between the two datasets up to a certain scale (or more than one scale).

3.3 Data

Various datasets are involved in the study: network skims, Singapore Household Interview Travel Survey (HITS), and taxi GPS data. The trip records used for the regression analysis are extracted from HITS and GPS data. After attaching zonal information, trip distance, peak and off-peak speed from network skims, a data cleaning process is applied before model estimation. This section describes these datasets and the data cleaning process.

3.3.1 Network skims

Singapore island is divided into 1,092 traffic analysis zones. Zone-to-zone travel time and network distance are provided for three time periods: AM peak from 07:30 to 09:30, PM peak from 17:30 to 19:30, and off-peak hours for the rest of the day in the network skims. Network skims provide a rough picture of time-dependent travel time for each OD pair for a small number of time periods. However, skims are not able to capture the variation of travel time within these time periods.

For each OD pair, network off-peak speed and peak speed are calculated. The network off-peak speed for each OD pair is defined by the following equation:

$$V_{\text{off}-\text{peak}} = \frac{\text{off} - \text{peak centroid distance}}{\text{off} - \text{peak travel time}} \tag{3.3}$$

The off-peak travel time in Equation 3.3 includes only in-vehicle time. Off-peak speed is similar to free flow speed as off-peak hours are expected to suffer less from congestion in

comparison with morning and evening peaks. The assumption that off-peak speed is larger than peak speed as we define it is true for almost all 1,191,372 OD pairs. We do, however, observe OD pairs that violate this assumption.

Another speed necessary for calculating the delay variable is peak speed. For AM and PM peaks, it was initially defined as AM speed and PM speed, respectively. For off-peak hours, it was defined as the minimum of AM speed and PM speed. Following this definition, the delay variable becomes time-dependent which results in discontinuity of predicted speed at the boundary between peak and off-peak hours. To address this problem, we decided to make the delay variable not time-dependent by redefining the peak speed in Equation 3.2 as the minimum of AM speed and PM speed.

3.3.2 HITS dataset

HITS is conducted by the Land Transport Authority (LTA) in Singapore every 4 years. The HITS in the study was conducted in 2008 (the most recent one is HITS 2012) and is used for extracting trips conducted by car drivers and passengers. Among all trips, 21,060 are made by either car drivers or car passengers.

While the delay variable, zonal CBD dummies, peak speed, and off-peak speed of a trip record are attached from the network skims, departure time, arrival time, and in-vehicle travel time are self-reported in the survey and included in the HITS dataset. Another important variable used in the regression analysis is trip travel speed. In HITS, participants are not required to report trip distance. Therefore, we use the zonal shortest path distance (the minimum of AM peak, PM peak, and off-peak distances) of the same OD pair as trip distance to calculate travel speed.

The following data cleaning steps were performed:

1. **Quality of data**: Trip attributes should not be empty or non-numerical.
2. **Illogical travel time**: In Singapore, any OD pair could be traversed by car in 2 hours. Beyond that, the trip duration is considered to be illogical.
3. **Carpooling**: Since trips are reported by each member in the same household, it is

possible for two or more members to report the same trip when they share the same vehicle. Carpool trips involving several household members should be considered only once in the regression model. In the HITS survey, participants need to specify the number of passengers in a car if a trip is conducted by car driver or passenger. But there is no direct question on whether the trip involves more than one household member or not. We therefore assume that if trips conducted by members in a household have the same origin, destination, departure time, and number of persons in the vehicle, those trips are identified as carpool trips, and only one trip is preserved in the dataset.

4. **Illogical travel speed**: Trips with extremely high travel speed, e.g., 1000 km/h, should not be included.

Table 3.1: Data cleaning procedure and results for car trips in HITS dataset

Step	Cleaning step description	Number of excluded trips	Remaining records (percentage)
1	Initial data	0	21,060 (100.0%)
2	Screen trips with duration or in-vehicle time greater than 120 minutes	362	20,698 (98.3%)
3	Screen intra-zonal trips	279	20,419 (97.0%)
4	Screen trips with 0 network distance and null travel time	37	20,382 (96.8%)
5	Screen trips with network peak speed greater than off-peak speed	267	20,115 (95.5%)
6	Screen carpool trips	4,556	15,559 (73.9%)
7	Screen trips with illogical travel speed	136	15,423 (73.2%)

The data cleaning procedure, number of trips excluded at each step, and remaining records are summarized in Table 3.1.

3.3.3 Taxi GPS data

Taxi GPS data are collected by GPS receivers installed in Comfort taxis. The receivers are active only when passengers are on board. Each record is a trip with departure time, arrival time, origin, and destination. Vehicle plate number and driver ID are recorded as well. The available data are for August 2011. During that month, 11,888,644 trips with passengers on board were recorded.

Compared with the HITS dataset, GPS data have several advantages. Firstly, they cover a majority of taxi trips on the island over all time periods of the day. Secondly, they record time information with better accuracy. Thirdly, they cover a greater number of OD pairs than does the HITS dataset. Despite all these advantages, a shortcoming that cannot be neglected is that the travel distance cannot be acquired by adding flying crow distance between GPS coordinates. Map matching techniques are required to map the GPS coordinates to road networks and derive the traveled distance from origin to destination. Given that, network skim distance (the minimum of AM peak, PM peak, and off-peak distances) is used.

Here a data cleaning procedure is also applied. Cleaning steps and results are summarized in Table 3.2. It should be noted that only weekday trips are considered as the network skims only apply to weekdays and travel demands are different between weekdays and weekends. After the cleaning procedure, the dataset is still too large to be used in the regression model from a computational standpoint. A random sample of around 20,000 trips is selected from the 8,454,879 trips that pass the data cleaning procedure.

For the two cleaned datasets, some descriptive statistics like trip rate, average travel speed, and average ratio of travel speed to off-peak speed are plotted by departure time period in Figures 3.1 to 3.3. In Figure 3.1, trip rate is measured as number of trips per hour. For HITS, the rate reaches a peak in the AM and PM peak hours. For GPS data, on the other hand, the rate remains almost constant after 8 AM. In terms of travel speed, average travel speed calculated with GPS data is always greater than that calculated with HITS data for all time periods. Two possible explanations are that (1) HITS participants tend to overestimate trip travel time as indicated in Stopher et al. (2007), or (2) taxi drivers tend to drive faster than other passenger vehicles and are more aggressive in route choice. Figure 3.3 plots the

Table 3.2: Data cleaning procedure and results for car trips in taxi GPS dataset

Step	Cleaning step description	Number of excluded trips	Remaining records (Percentage)
1	Initial data	0	11,888,644 (100.0%)
2	Screen trips conducted in weekends	2,920,438	8,968,206 (75.4%)
3	Screen trips with 0 network distance and null travel time	272,619	8,695,587 (73.1%)
4	Screen intra-zonal trips	165,755	8,529,832 (71.7%)
5	Screen trips with network peak speed greater than off-peak speed	60,179	8,469,653 (71.2%)
6	Screen trips with illogical travel speed	14,774	8,454,879 (71.1%)

average ratio of travel speed to off-peak speed by departure time period for both datasets. It is shown that while the ratio is below 1 for HITS data at all time periods, it is greater than 1 and remains almost constant for GPS data from midnight to around 6 am. From Figure 3.2 and 3.3, we can observe some sudden drops of speed or ratio in the early morning (2:30 AM to 4 AM) for the HITS dataset. For those early morning time periods, very few trips were reported; thus, the values at those time periods are very sensitive to the quality of those few trips. Also, we observe drops in both the average speed and in the average ratio of speed to off-peak speed for taxi GPS data during AM and PM peak periods, but such variations are absent in the plots of the HITS dataset, which may be attributed to inaccurate travel time reporting by HITS respondents.

In terms of the spatial distribution of trips, it is found that zones in the CBD area of Singapore generate and attract more trips than other zones for taxi GPS trips. Particular origin or destination zones also generate or attract more taxi trips, such as Changi Airport and Sentosa Island as destination. Such concentration of trips is not observed in HITS trips.

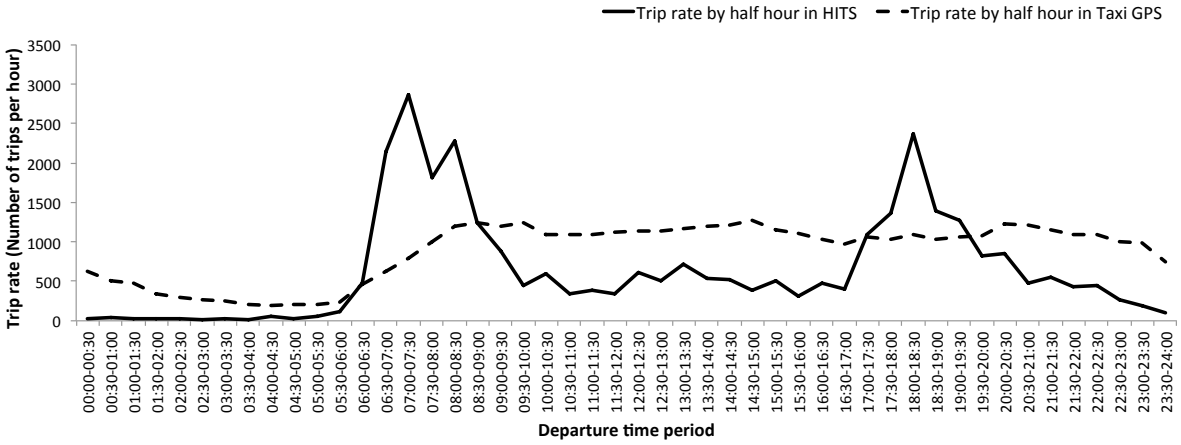


Figure 3.1: Trip rate by departure time period

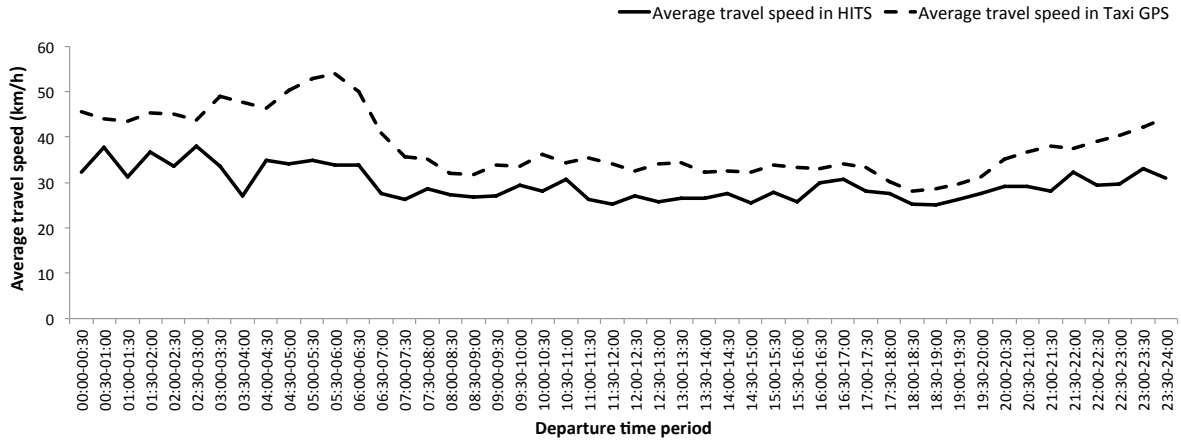


Figure 3.2: Average travel speed by departure time period

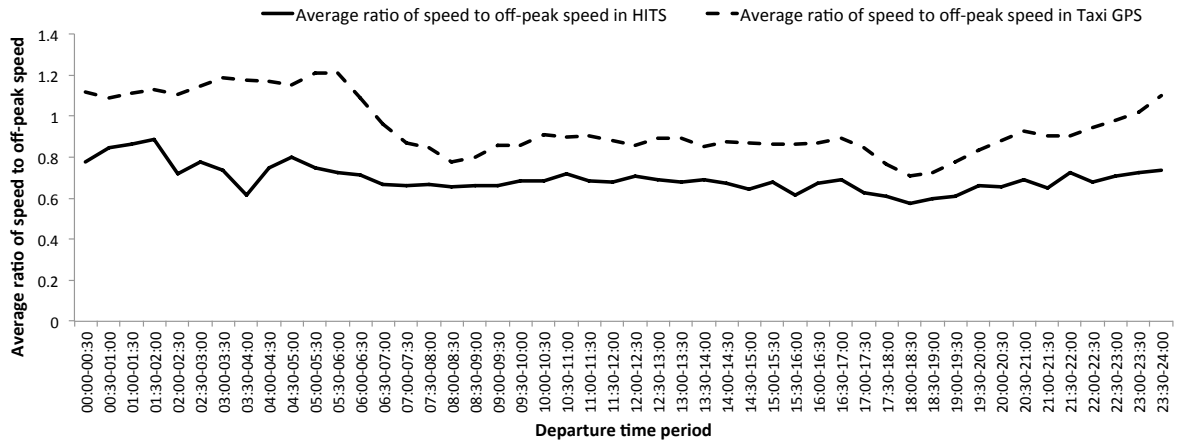


Figure 3.3: Average ratio of travel speed to off-peak speed by departure time period

3.4 Model Estimation and Analysis

3.4.1 Unrestricted model

In this section, the HITS and GPS models are estimated separately using the same specification. To decide on the truncation points M_1 and M_2, we tried many combinations in an attempt to get a good model fit as well as a reasonable delay profile. It turns out that no combination tried can produce reasonable delay profiles for both HITS and GPS datasets. And in most cases, the delay coefficient of the departure-time-based model of HITS becomes positive during the early morning.

By pooling the two datasets, we take one step back and consider the best specification for the pooled dataset. The specification should produce a reasonable delay profile for both departure-time-based and arrival-time-based models estimated with the pooled dataset. $M_1 = 10$ and $M_2 = 12$ are used as truncation points for the two trigonometric functions after trying a number of combinations. In this section, we use this specification to estimate the unrestricted models with the HITS and GPS datasets.

The estimation results are summarized in Table 3.3 and Table 3.4. For both data sources, distance has a positive influence on the travel speed indicating that a higher speed is expected for longer trips. CBD dummies on the other hand have a negative influence on the travel speed as expected.

In terms of the trigonometric functions, it is more intuitive if we plot them and compare the results between models estimated with the two data sources. Figure 3.4 shows the congestion-sensitive trigonometric function that interacts with the delay variable (delay coefficient). Except for the departure-time-based model of HITS, other models produce reasonable delay profiles. During daytime, there is a good offset relationship between the departure-time-based model and the arrival-time-based model for both HITS and GPS datasets (i.e., an arrival-time-based peak closely follows a departure-time-based peak) except for the two models of HITS in the early morning. In fact, we found that by changing the specification, the shape of the delay profiles in the early morning for the departure-time-based model and the arrival-time-based model will change greatly. In contrast, for the GPS dataset,

Table 3.3: Estimation results: Departure-time-based model

Variable	Parameter : HITS (t-stat against 0)	Parameter : GPS (t-stat against 0)
Intercept	-0.7577 (-47.26)	-0.0859 (-11.55)
Ln(distance)	0.2147 (47.96)	0.0687 (20.17)
CBD Origin Dummy	-0.0959 (-9.07)	-0.1069 (-19.00)
CBD Destination Dummy	-0.0289 (-2.72)	-0.0503 (-8.78)
Delay	-0.4078 (-9.75)	-0.4461 (-20.18)
sin(2πt/24)*Delay	0.1826 (3.13)	0.0049 (0.17)
cos(2πt/24)*Delay	0.1471 (2.22)	0.0492 (2.01)
sin(4πt/24)*Delay	0.2612 (3.98)	0.0632 (2.39)
cos(4πt/24)*Delay	0.0169 (0.28)	0.0192 (0.72)
sin(6πt/24)*Delay	-0.1166 (-1.93)	-0.0377 (-1.44)
cos(6πt/24)*Delay	-0.1436 (-2.19)	-0.1140 (-4.35)
sin(8πt/24)*Delay	-0.1763 (-3.30)	-0.0716 (-2.82)
cos(8πt/24)*Delay	-0.1050 (-2.12)	0.0041 (0.17)
sin(10πt/24)*Delay	-0.0133 (-0.31)	0.0246 (1.05)
cos(10πt/24)*Delay	0.0222 (0.52)	-0.0195 (-0.82)
sin(2πt/24)	-0.0060 (-0.28)	0.1149 (12.09)
cos(2πt/24)	-0.0113 (-0.47)	0.0966 (12.04)
sin(4πt/24)	-0.0600 (-2.44)	0.0735 (8.47)
cos(4πt/24)	0.0884 (4.22)	0.0379 (4.30)
sin(6πt/24)	0.0208 (0.93)	-0.0178 (-2.06)
cos(6πt/24)	0.0554 (2.44)	0.0075 (0.87)
sin(8πt/24)	0.0617 (3.18)	-0.0257 (-3.07)
cos(8πt/24)	0.0428 (2.37)	0.0073 (0.88)
sin(10πt/24)	0.0421 (2.81)	0.0348 (4.51)
cos(10πt/24)	0.0038 (0.25)	0.0223 (2.82)
sin(12πt/24)	0.0247 (3.92)	0.0281 (8.06)
cos(12πt/24)	0.0108 (1.73)	0.0079 (2.29)
Number of Observations	15,423	21,147
Number of Parameters Estimated	27	27
SSE	2,216.99	2,324.14
R-Square	0.1605	0.1489
Adjusted R-square	0.1591	0.1479

Table 3.4: Estimation results: Arrival-time-based model

Variable	Parameter : HITS (t-stat against 0)	Parameter : GPS (t-stat against 0)
Intercept	-0.6988 (-48.83)	-0.0845 (-11.32)
Ln(distance)	0.2175 (48.40)	0.0674 (19.78)
CBD Origin Dummy	-0.0848 (-8.05)	-0.1065 (-18.90)
CBD Destination Dummy	-0.0268 (-2.53)	-0.0501 (-8.74)
Delay	-0.5988 (-16.87)	-0.4395 (-19.80)
sin(2πt/24)*Delay	0.0211 (0.37)	0.0193 (0.66)
cos(2πt/24)*Delay	-0.0997 (-1.71)	0.0392 (1.63)
sin(4πt/24)*Delay	0.0433 (0.74)	0.0788 (3.01)
cos(4πt/24)*Delay	-0.0464 (-0.76)	-0.0086 (-0.31)
sin(6πt/24)*Delay	-0.2155 (-3.54)	-0.0681 (-2.56)
cos(6πt/24)*Delay	0.0527 (0.85)	-0.0962 (-3.69)
sin(8πt/24)*Delay	-0.1179 (-2.16)	-0.0469 (-1.85)
cos(8πt/24)*Delay	-0.0092 (-0.19)	0.0474 (1.88)
sin(10πt/24)*Delay	-0.0020 (-0.05)	0.0055 (0.23)
cos(10πt/24)*Delay	-0.0272 (-0.63)	-0.0681 (-2.87)
sin(2πt/24)	0.0678 (3.21)	0.1207 (12.49)
cos(2πt/24)	0.0543 (2.53)	0.0897 (11.33)
sin(4πt/24)	0.0470 (2.12)	0.0734 (8.51)
cos(4πt/24)	0.0686 (3.25)	0.0276 (3.08)
sin(6πt/24)	0.0490 (2.22)	-0.0152 (-1.73)
cos(6πt/24)	-0.0053 (-0.24)	0.0106 (1.24)
sin(8πt/24)	0.0419 (2.16)	-0.0270 (-3.23)
cos(8πt/24)	0.0312 (1.76)	0.0086 (1.03)
sin(10πt/24)	0.0399 (2.72)	0.0400 (5.14)
cos(10πt/24)	-0.0202 (-1.34)	0.0165 (2.10)
sin(12πt/24)	0.0300 (4.71)	0.0279 (8.01)
cos(12πt/24)	-0.0237 (-3.78)	-0.0088 (-2.54)
Number of Observations	15,423	21,147
Number of Parameters Estimated	27	27
SSE	2211.99	2,330.08
R-Square	0.1624	0.1467
Adjusted R-square	0.1610	0.1457

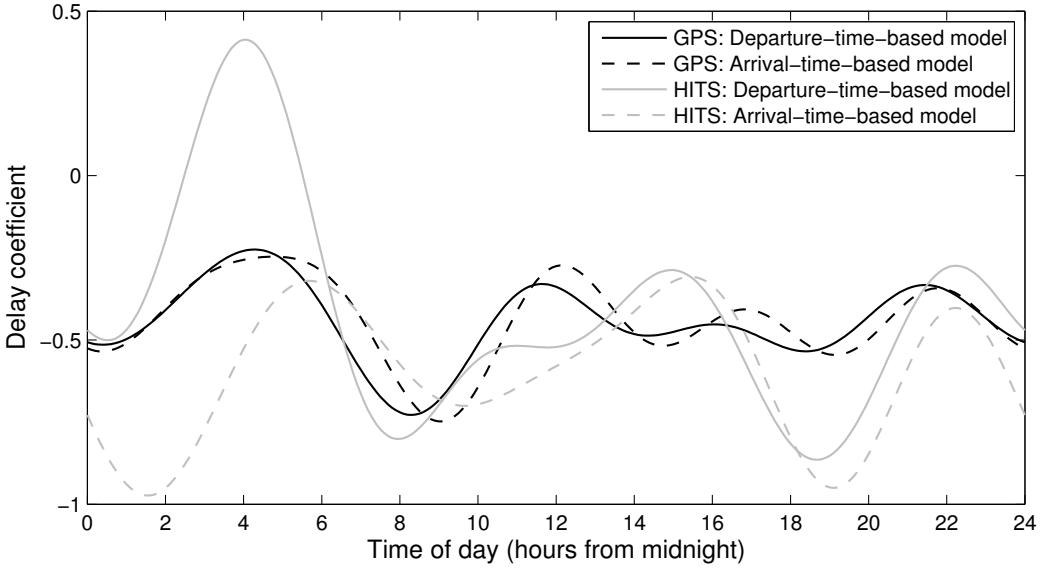


Figure 3.4: Congestion-sensitive trigonometric function

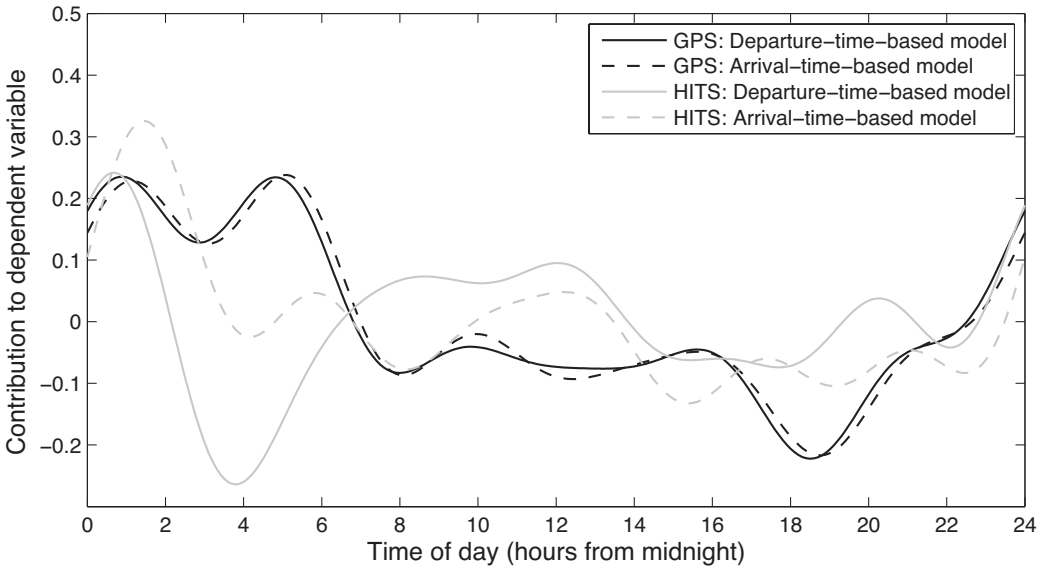


Figure 3.5: Congestion-free trigonometric function

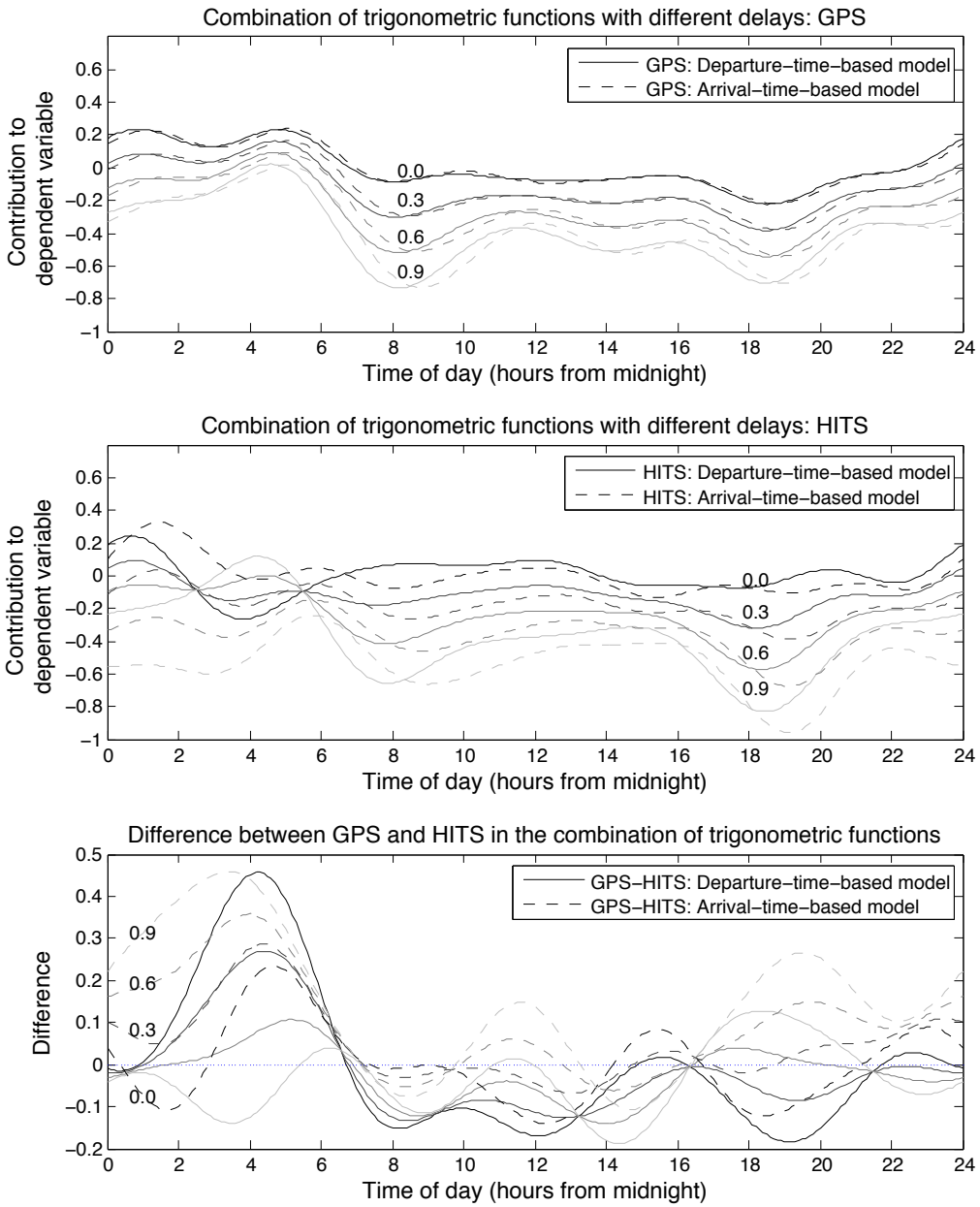


Figure 3.6: Combination of congestion-sensitive and congestion-free trigonometric functions when delay varies from 0 to 1. The first plot is for GPS data and second plot is for HITS data. The third plot is their difference.

the shape is less sensitive to different specifications.

Figure 3.5 plots the congestion-free trigonometric function that does not interact with delay. The Y-axis is the contribution to natural logarithm of speed over off-peak speed. The figure shows the time-dependent variation of speed that cannot be captured by the delay variable. As shown by the GPS model, people tend to drive faster after midnight.

By varying delay from 0 to 1, we plot the combination of the congestion-sensitive trigonometric function multiplied by delay and the congestion-free trigonometric function in Figure 3.6 for each of the GPS and HITS models as well as the difference between the two models. The y coordinate being the contribution to the dependent variable, measures the change of the dependent variable when delay takes a value in $[0, 1]$. While for both data sources, AM and PM peaks appear at around the same time, the GPS dataset results in a lower AM peak speed and the HITS dataset results in a lower PM peak speed when the delay variable is close to 1. From the first plot for the GPS data, we can observe a larger offset between the departure-time-based model and the arrival-time-based model as delay increases. As shown in Figure 3.4 and Figure 3.5, the offset between the departure-time-based model and the arrival-time-based model is mainly brought by the congestion-sensitive trigonometric function, indicating that the more severe the congestion is on the road, the larger the difference between travel time predicted by the departure-time-based model and the arrival-time-based model.

To further compare the model estimated with HITS and GPS data, we plot the weighted average ratio of speed to off-peak speed by time of day in Figure 3.7. For time t, the ratio of speed to off-peak speed is first calculated with the estimated models for each OD pair (including the effects of distance and CBD dummies). Then this ratio is weighted by the daily demand for every OD pair (obtained from the four-step model), from which a weighted average ratio of speed to off-peak speed is calculated for time t. Both models estimated with the HITS and GPS datasets capture the peak effect in the AM and PM peak periods. The peaks in the GPS model are more clearly distinguished, indicating that taxi trips tend to be concentrated in congested areas. Compared with the HITS model, the GPS model gives a higher ratio of speed to off-peak speed at all times, which may imply that taxi drivers are more familiar with the road network and make better route and lane choice decisions.

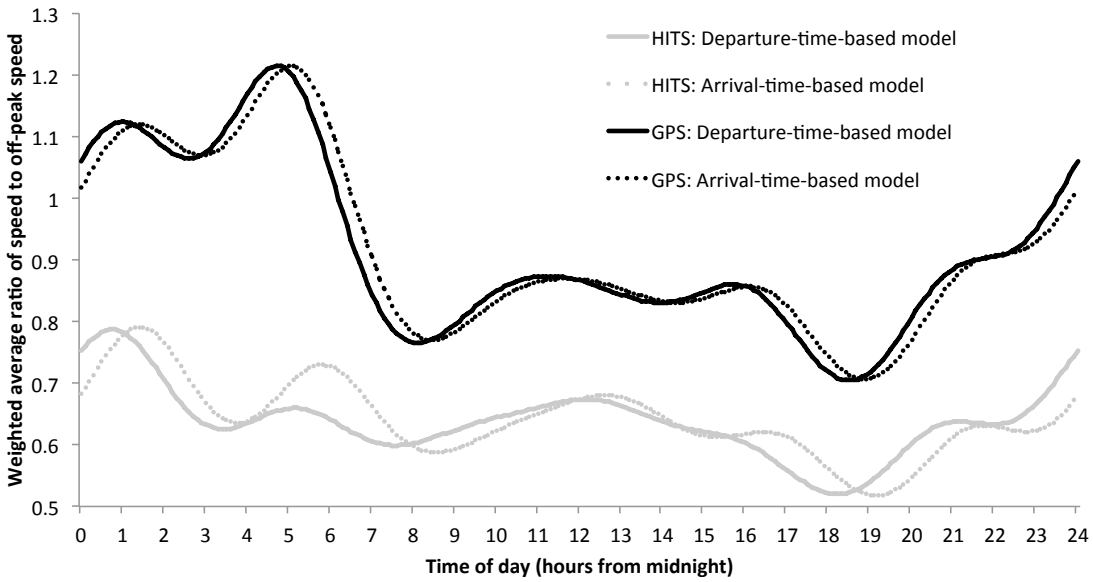


Figure 3.7: Weighted average ratio of speed to off-peak speed by time of day

Also the ratio is greater than 1 during the early morning for the GPS model. However, by looking at the part of the plot after 8 am only, we could observe some similarities between the patterns given by two models during daytime. To further explore the similarities, we next consider the restricted models and conduct a few hypothesis tests.

3.4.2 Restricted models and hypothesis tests

A summary of the statistical tests performed to investigate whether the two data sources can be combined or not is presented in Table 3.5 and Table 3.6 for the departure-time-based and arrival-time-based model, respectively. The statistical test results indicate that there are significant differences between the two data sources. The hypothesis that the coefficients of all the explanatory variables are equal across the two data sources can be rejected. However, considering just the temporal pattern, the difference between the individual (unrestricted) models and the combined model is found to reduce significantly (manifested by decreased Sum of Squared Errors (SSE) and F-statistics) from specification 1 to specification 2 and it continues to decrease gradually as we move down from specification 2 to specification 5 for the combined/restricted model. Also significant scale differences are observed among the

Table 3.5: Statistical test results of departure-time-based model

Model Description	Scale of the Trigonometric Series in the HITS Dataset	Estimated Scale (t-stat against 1)	SSE	F-stat	Dof_1	Dof_2	F-critical at $\alpha = 0.05$
Unrestricted Model	–	–	4541.12	–	–	–	–
Restricted Models							
1. All coefficients equal	One	Fixed	5391.00	253.11	27	36,516	1.49
2. Coefficients of the trigonometric series equal	One	Fixed	4673.47	11.86	22	36,516	1.55
3. Coefficients of the trigonometric series equal (one scale parameter)	μ for both congestion-sensitive and -free trigonometric series	0.7565 (-6.39)	4569.53	10.91	21	36,516	1.56
4. Coefficients of the trigonometric series equal (two scale parameters)	μ_{CS} for congestion-sensitive trigonometric series	0.3427 (-10.25)	4565.22	9.72	20	36,516	1.57
	μ_{CF} for congestion-free trigonometric series	1.0391 (0.51)					
5. Coefficients of the trigonometric series equal (four scale parameters)	$\mu_{CS,sine}$ for congestion-sensitive sine functions	0.1380 (-7.95)	4558.15	7.61	18	36,516	1.61
	$\mu_{CS,cosine}$ for congestion-sensitive cosine functions	4.0775 (1.66)					
	$\mu_{CF,sine}$ for congestion-free sine functions	0.5651 (-8.15)					
	$\mu_{CF,cosine}$ for congestion-free cosine functions	0.2438 (-7.05)					

Table 3.6: Statistical test results of arrival-time-based model

Model Description	Scale of the Trigonometric Series in the HITS Dataset	Estimated Scale (t-stat against 1)	SSE	F-stat	Dof_1	Dof_2	F-critical at $\alpha = 0.05$
Unrestricted Model	–	–	4542.07	–	–	–	–
Restricted Models							
1. All coefficients equal	One	Fixed	5400.69	255.66	27	36,516	1.49
2. Coefficients of the trigonometric series equal	One	Fixed	4564.47	8.18	22	36,516	1.55
3. Coefficients of the trigonometric series equal (one scale parameter)	μ for both congestion-sensitive and -free trigonometric series	0.8022 (-5.12)	4561.66	7.52	21	36,516	1.56
4. Coefficients of the trigonometric series equal (two scale parameters)	μ_{CS} for congestion-sensitive trigonometric series	2.8447 (2.51)	4554.19	4.89	20	36,516	1.57
	μ_{CF} for congestion-free trigonometric series	0.4547 (-9.5)					
5. Coefficients of the trigonometric series equal (four scale parameters)	$\mu_{CS,sine}$ for congestion-sensitive sine functions	1.4137 (1.39)	4550.69	3.85	18	36,516	1.61
	$\mu_{CS,cosine}$ for congestion-sensitive cosine functions	2.6047 (-2.24)					
	$\mu_{CF,sine}$ for congestion-free sine functions	0.5957 (-6.47)					
	$\mu_{CF,cosine}$ for congestion-free cosine functions	0.6260 (-3.55)					

coefficients of the congestion-sensitive and congestion-free sine, cosine functions between the taxi GPS data and HITS dataset. The pattern of decreasing F-statistic as more scale parameters are added suggests that the taxi GPS data and HITS data result in a similar profile of time-of-day variation in speed up to certain scaling parameters. The null hypothesis of equal coefficients is however not accepted even when 4 scaling parameters are used possibly because of the very large number of observations used.

3.5 Summary

In this research, we estimated a travel time regression model using household survey (HITS) and taxi GPS data from Singapore with the objective of comparing the consistency and the quality of travel time data in both datasets. Differences are observed with respect to the overall predicted speeds as well as the delay profiles. We attribute such differences to both the quality of the datasets and the different characteristics of car trips and taxi trips. Reported trip attributes, like departure time, arrival time, and in-vehicle time in HITS are not as accurate as those recorded in the GPS data. Furthermore, while the HITS model may perform well in predicting travel time during daytime, the results for early morning from the HITS model may not be reliable since only a few trips are reported during early morning. In terms of different characteristics between car trips and taxi trips, travel speed of taxi trips in GPS data is greater than that of car trips in HITS. Further effort should be made to confirm whether the difference is caused by people's overestimating trip duration in household travel surveys, or taxis' traveling faster than ordinary passenger cars, or both.

Despite these differences, we observe some similarities between the models in terms of weighted (by number of trips by OD pair) average ratio of speed to off-peak speed during daytime, which inspires us to estimate restricted models on a pooled dataset combining HITS and GPS data and conduct several hypothesis tests. Different restrictions are imposed on the combined model and lead to different specifications for restricted models. Although for each of these specifications we fail to accept the hypothesis that all or some coefficients are equal across the two datasets, it is found that the temporal patterns of speed predicted from both datasets follow a similar profile up to some scaling constants.

Chapter 3. Travel Time Modeling with GPS and Household Travel Survey Data

Regression analysis has been performed with household travel survey data before (in Abou-Zeid et al., 2006 and Popuri et al., 2008) to model travel time. The available taxi GPS data enable us to reveal the difference and similarity between models estimated using both survey and GPS data. Such models would further benefit from the availability of future household travel surveys carried out with GPS-enabled technology or evaluated with a GPS survey conducted simultaneously to help answer the question whether the differences observed in this study are the results of using travel time of two different modes – taxis and passenger cars.

This research serves as one of the first steps of developing activity-based demand models in Singapore. Besides car trips, similar regression models are developed for public transportation modes where household survey data and smart card data are utilized. With regression models, travel times by mode and time of day for every half-hour are generated and serve as input to estimating time of day models in activity-based models.

Chapter 4

Preparing Household Travel Survey Data for Activity-based Modeling

Although various sources of data are involved in the development of activity-based models, household travel survey remains the main data source as the individual models in an activity-based modeling framework are all estimated with household travel survey. While trip-based traditional household travel surveys are still dominant, it is argued in this chapter that more advanced surveys may not be necessary and the information captured in an ordinary trip-based survey may be enough for the implementation of activity-based travel demand models. Sufficient efforts, such as data checks, trip-to-tour conversion, and work-based sub-tour detection are introduced in detail. Then the processed data are presented with a descriptive analysis.

4.1 Introduction

Although there has been a list of studies exploring replacing traditional diary based one-day or two-day household travel surveys with GPS-based surveys (see for example Ohmori et al., 2005; Bricka and Bhat, 2006; Bellemans et al., 2008; Giaimo et al., 2009; Abdulazim et al., 2013 and Shen and Stopher, 2014), traditional household travel surveys remain the major source of data needed for activity-based models. Most of the large-scale empirical development and deployment of activity-based modeling frameworks have been relying on the traditional surveys. For example, the development of NYBPM (Parsons Brinckerhoff, 2005b) utilized a survey including approximately 11,000 households in a one-day survey; the MORPC model (Parsons Brinckerhoff, 2005a) was built with a one-day survey with 5,555 households; The SACSIM developed by SACOG (DKS Associate et al., 2012) and DRCOG model (Cambridge Systematics, Inc., 2010) used surveys with 3,941 and nearly 5,000 households respectively.

There are several reasons to explain the prevalence of traditional household travel surveys in the development of empirical activity-based modeling frameworks.

Firstly, the development of activity-based models requires a great number of explanatory variables. The effect of explanatory variables is tested in the process of model estimation. However, those variables need to be considered in the survey in the first place such that the test becomes possible. Putting a large quantity of unnecessary details in the survey will result in a low response rate, increased budget and decreased quality of data. Furthermore, a survey conducted in Area A with certain set of variables may not work perfectly in Area B and a customized survey for activity-based model will actually hinder the progress of model development. Using existing surveys seems to be the most practical solution.

Secondly, the traditional household travel survey is the main data source for the development of models based on the four-step method. It is usually carried out regularly with a relatively stabilized format, which may suggest that both the survey practitioners and the public have become familiar with the traditional data collection efforts. Moreover, while more regions have shown interests in the activity-based modeling approach, most of them are very prudent

toward the deployment of such models or even replacing the ones based on the four-step method. As a result, regional planning authorities will usually require a comparison study between the traditional model and the activity-based one. In such circumstances, having the same data source used to estimate both models is an efficient way to control the error introduced by using separate datasets for developing different models.

Thirdly, it was estimated that only about 3 percent of the information obtained from travel surveys is used in the traditional four-step models (Petersen and Vovsha, 2006a). The fact may suggest that many regional planning agencies have already possessed the data required for developing activity-based models. As shown in Sabina et al. (2006), complex, advanced and specially designed surveys are not necessary to develop reasonable tour codes to support the development of activity-based models and it is possible to acquire the data and modeling units in activity-based models using traditional trip-based surveys.

Given the above reasons and following the spirits of Sabina et al. (2006), we would like to explore the possibility of applying traditional trip-based household travel surveys to the development of activity-based models. Specifically, we will rely on 2008 Household Interview and Travel Survey in Singapore (HITS2008) as the main data source of the activity-based travel demand model introduced in the subsequent chapter.

Land Transport Authority (LTA) of Singapore has conducted a household interview and travel survey every four years since 2000. The latest one available at the moment of model development was HITS2008. HITS is conducted by means of an internet-based interview with all eligible members of the household to determine their travel patterns on a typical weekday. The survey is designed to capture details of the household, personal characteristics and trip-making/activity-participation of each person in the household. All legal residents in Singapore are the target participants[1].

The household sampling rate of HITS is around 1 percent. For the case of HITS2008, stratified random sampling of the Primary Sampling Unit (PSU) by districts in Singapore was applied and 10,840 households were sampled compared with a target of 10,500. After

[1]Those included in the survey are current legal residents of Singapore: namely Singapore citizens, Permanent Residents, EP Holders, Student Pass Holders, Work Permit Holders and Dependent Pass Holders.

Table 4.1: Summary of the sampled households in HITS2008

	Frequency (percentage)
1. HITS households	
* Total (screened out)	10,840 (199)
* Accepted	10,641
* Weighted households	1,144,409
2. Household size	
* 1	105,706 (9.24%)
* 2	208,160 (18.19%)
* 3	236,461 (20.66%)
* 4	292,292 (25.54%)
* 5	185,337 (16.19%)
* 6+	116,453 (10.18%)
3. Household with childern under 15 years old	
* Yes	466,708 (40.78%)
* No	677,701 (59.22%)
4. Household private vehicle availability[1]	
* 0	698,688 (61.05%)
* 1	391,519 (34.21%)
* 2	50,240 (4.39%)
* 3+	3,962 (0.35%)

[1] Personal-registered private cars available to the household

Table 4.2: Summary of the sampled population in HITS2008

	Frequency (percentage)
1. HITS population	
* Total	38,053
* Under 4 yrs old + Not eligible[1]	1,760 + 1,170
* Accepted sample size	35,123
* Weighted population	3,764,326
2. Gender	
* Male	1,752,660 (46.56%)
* Female	2,011,666 (53.44%)
3. Age group	
* 4-9 years old	303,546 (8.06%)
* 10-20 years old	597,547 (15.87%)
* 20-30 years old	549,219 (14.59%)
* 30-40 years old	635,609 (16.89%)
* 40-50 years old	658,039 (17.48%)
* 50-60 years old	506,509 (13.46%)
* 60 years old and above	513,857 (13.65%)
4. Driving license	
* Yes	1,202,871 (31.95%)
* No	2,561,455 (68.05%)
5. Personal monthly income	
* Refused	250,054 (6.64%)
* No income	1,848,086 (49.09%)
* S$1-S$2,000	763,720 (20.29%)
* S$2,000-S$4,000	579,953 (15.41%)
* S$4,000-S$6,000	187,439 (4.98%)
* S$6,000-S$8,000	54,703 (1.45%)
* S$8,000+	80,371 (2.14%)

[1] Those two groups of people were not interviewed during the survey period. Among them, children under 4 years old are assumed not able to make any trips during the day by themselves.

the quality control procedure conducted by LTA, data from 199 households were eliminated and the remaining 10,641 households were accepted as summarized in Table 4.1. Compared to the island-wide total household number of 1,144,400, the actual sampling rate is right under 1 percent. Table 4.2 presents some of the basic social-demographic characteristics associated with the sampled population.

HITS datasets have traditionally been used for developing trip-based travel demand models in LTA. However, its application to the development of activity-based models is not straightforward. In the rest of the chapter, sufficient efforts, such as data checks, trip-to-tour conversion, and work-based sub-tour detection are introduced in detail. Then the processed data are presented with a descriptive analysis.

4.2 Methodology

This section introduces definitions, assumptions and methods used in the data processing and tour encoding of HITS survey.

4.2.1 Survey data structure

The original data acquired from LTA is sorted in the following increasing order.

- H1_HHID
- Pax_ID
- Trip_ID
- Stage_ID
- msno_main

where msno_main is the internal ID used during the survey and may not be necessary for further application of the data. H1_HHID provides the household unique ID and Pax_ID is the person ID in the household. Trip_ID and Stage_ID are used to locate a specific stage inside a trip since a trip may consist of several stages involving different travel modes.

The sorted file ensures that the records of a particular participant are consecutive in the table and are presented in the order in which the person's trips took place. Each row of the table stands for a particular stage of a trip made by a participant. Columns in the table provide information of the participant and the information falls in three categories: Household particulars, Person particulars, and Trip particulars.

4.2.2 Data checks

Although the survey is internet-based and the structure of it is carefully designed to ensure acceptable answers in each section, further data checks are still necessary to maintain the correctness and consistency of the data, which is especially valued if the data will be used for tour encoding as more complicated structures (such as trip chains or tours) have strict requirement of data consistency.

On one hand, the purpose of running data checks is to provide us with some insights about the data and prepare the data for future applications. On the other hand, before converting trips into tours we need these checks to determine which participant should be eliminated from further consideration because some of his/her checks are indicated as wrong and make it impossible to derive tour information from the database.

File structure check We need to check that the data are sorted: all trips of a given participant are in increasing order (Trip_ID & Stage_ID). Given the fact that some stage IDs are missing, we also need to check the stage ID itself and create a unique column containing the correct stage ID. The column is called Stage ID Modified.

Time/duration checks The survey system did not check if the time information typed in by the participants is correct and has no logical errors. After changing the format of time from HHMM to minutes accumulated from midnight, we ran 5 checks. The details and results of the checks are shown in Table 4.3. Basically, the time/duration checks ensure that the trip chain of each individual is consistent and there are no inner contradictions in terms of time and duration of both reported trips and activities.

Table 4.3: Results of time/duration checks

Check ID	Details	Results	
		Result/error	Explanation
1	The departure time of the following trip should be greater than the arrival time of the preceding trip. Also the activity duration is not greater than 15 hours. In the database we use "Flag2" to indicate the result of the check.	534	534 entries have activity durations either greater than 15 hours or with negative numbers.
2	It is plausible that the longest duration of a single trip is 2 hours in Singapore. Any trip with duration greater than 2 hours is considered wrong. "Flag_duration" is used to indicate the result.	543	543 entries have durations greater than 2 hours.
3	For any given trip, the arrival time should be greater than the departure time. "Flag1" is used to indicate this result.	1,310	1,310 entries have negative trip durations.
4	Check if we have start time and end time for each trip.	0	Every trip has start time and end time.
5	Check if we have estimated time and calculated time for each trip.	0	Every trip has estimated and calculated time.

Origin/destination checks In each trip, the origin postal code and destination postal code are to be provided. And we need to check these postal codes to find if they are consistent and make sense. We ran 6 checks. The details and results are shown in Table 4.4. Check 3 and 4 are not the indication of errors as intermediate trips in a trip chain are not home-originated or home-destined.

Table 4.4: Results of origin/destination checks

Check ID	Details	Results	
		Result/error	Explanation
1	For particular person, the origin of the following trip must be the destination of the preceding trip. We use "Flag3" to indicate the result.	815	We have 815 entries that the origin of the following trip and the destination of preceding trip are different.
2	For any given trip, the destination and origin can not be the same. We use "Flag4" to indicate the result.	874	We have 874 entries that have the same origin and destination in a trip.
3	Check how many entries in the database have non-home-originated trips. We use "Flag5" to indicate the result.	12,670	12,670 entries contain non-home-originated trip.
4	Check how many entries in the database have non-home-destined trips. We use "Flag6" to indicate the result.	11,821	11,821 entries contain non-home-destined trip.
5	Check if the characteristic of the last trip location does not fit with the purpose. For example the location type is working yet the purpose is to return home. We use "Flag12" to indicate the result.	1	We have 1 case where the purpose is return home yet the type of the location is not residential.
6	Check if we have origin and destination postal codes for each trip.	0	We have origin and destination postal codes for each trip.

Mode checks Mode checks are to check if the mode combination makes sense in each trip and is compatible with the purpose of the trip. Here we only have 2 checks. Details and

results are shown in Table 4.5.

Table 4.5: Results of mode checks

Check ID	Details	Result	
		Result/error	Explanation
1	Activities with purpose "drop-off/pick up someone" or "accompany someone" should have appropriate number of persons in vehicle either for the preceding trip or the following trip if the mode is to drive. We use "Flag7" to indicate the result.	57	57 entries have trip purpose "drop-off/pick up someone" or "accompany someone" yet the number of persons in the vehicle is not correct.
2	In a trip, bicycle cannot follow car as a mode. We use "Flag11" to indicate the result.	6	6 entries have mode combination "car+bike".

Among the data checks we carried out, some will affect the further application of HITS in developing activity-based models. As a result, Individuals failed time/duration check 1-3, origin/destination check 1,2,5, and mode check 1,2 are excluded at the stage of model development. Moreover, some checks will have direct impacts in the stage of trip-to-tour conversion and related individuals are to be excluded at this stage.

4.2.3 Trip-to-tour conversion

While the traditional four-step method uses trips as the primary unit of analysis, activity-based models need trips, tours and day patterns as the unit of analysis. To filling the gap between a trip-based survey and the need for developing activity-based models, a procedure called trip-to-tour conversion, or tour encoding is necessary.

The first part of this conversion is to clarify the term "tour". Usually, a tour is a sequence of trips that start and end at home. Yet in the HITS database we have many occasions where the first trip of a person is not started from home, the last trip of a person is not ended at home or both. Instead of putting those tours aside, two terms "non-home-originated tour" and "non-home-destined tour" are created to describe "tours" with non-home-originated first trip or non-home-destined last trip. Besides the tours defined above, a special tour called "work-based sub-tour" needs attention. A work-based sub-tour is a tour that starts and ends at a work location. The reason to have home-based tours and work-based sub-tours is that home and work locations are the places people spend most of their time in a work day and trip decisions made during work are somewhat similar to the decisions made at home.

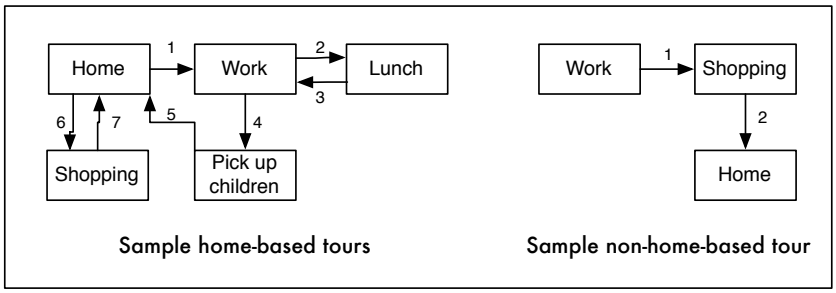


Figure 4.1: Tour patterns illustration

Figure 4.1 illustrates home-based tours, non-home-originated tours and work-based sub-tours. In the example to the left of the figure, the work tour (trip 1 to 5) includes an intermediate stop to pick up children and a work-based sub-tour for lunch (trip 2 to 3). The example also includes a simple shopping tour (trip 6 to 7). In the right part, the person starts the day at a non-home location and returns to home at the end of day. With the basic concept of tours introduced, we next describe the methods of tour encoding from HITS.

Definitions Before the trip-to-tour conversion, some definitions and terms used in the following sections are introduced here.

- **Stages of a trip**: A stage in a trip is part of the trip and is associated with a travel mode. Although people likely will have less than 5 stages in a trip, HITS2008 actually provides 10 potential stages for each trip.
- **Trip mode**: The trip mode is the main mode that the person uses to accomplish the trip. It is not as obvious as stages of a trip, since a trip may need to combine several modes in a sequence to finish. The sections below cover the hierarchy used to obtain the trip mode.
- **Tour mode**: The tour mode is the primary mode of the tour. Usually the tour mode is obtained by assigning a unique priority to each mode used in the trips of the tour.
- **Primary activity**: Primary activity is the activity a person conducts at the primary location of a tour. There are several different approaches to determine the primary

activity of a tour. Two ways are introduced in the following section and used to determine the primary activity.

- **Secondary stops**: Secondary stops or intermediate stops are the stops where a person conducts non-primary activities.

- **Primary work location**: For workers with one or more work locations, a work location, either the only one (fixed one) or the one having the largest duration of work activity, should be chosen as the primary one.

Imported data and tour table entries The imported data falls in two categories: data from the sorted database and flags. The following table shows which columns in the database are actually used in the encoding. Table 4.6 provides a summary of all imported data from HITS original trip table with column names in brackets. Notice that the five flags generated during the data checks are required for tour encoding.

Table 4.6: Imported data from HITS2008 trip table

	Data from sorted database			Flags
	HouseHold Particulars	Person particulars	Trip particulars	
1	H1_HHID (B)	PaxID (AC)	Trip_ID (BD)	flag1 (DE)
2	H1_Pcode (C)	P5a_Education (AM)	Stage_ID (BE)	flag2 (DF)
3		P5_EconActivity (AN)	P13d_OriginPcode (BH)	flag3 (DG)
4		P6a_FixedWkpl (AS)	P13e_DepartureTime (BI)	flag4 (DH)
5		P6c_FixedwkplPcode (AT)	T2_DestnPcode (BK)	flag12 (DZ)
6		P10_MakesTrip (AZ)	T3_StartTime (BL)	
7		P13_1stTripOrigin_Home (BF)	T4_EndTime (BM)	
8		P13b_1stOriginPcode (BG)	T5_PlaceType (BN)	
9		P14_1stTripStartTime (BJ)	T6_Purpose (BO)	
10			T10_Mode (BP)	
11			ActivityDuration(min) (DW)	

The tour table and work-based sub-tour table to be generated from HITS 2008 original trip table contain entries as shown in Table 4.7.

Methods of trip-to-tour conversion The logic flow shown in Figure 4.2 is translated into VBA macros that can be applied to the sorted database. The steps in the flow chart are briefly introduced as follows:

- **Data checks**: Before we store a person's trip information, we need to check if the related flags are correct. First of all, for those who report no trips or the trip ID

Table 4.7: Entries in tour table and work-based sub-tour table

	Tour table	Work-based sub-tour table
1	HHID (A)	HHID (A)
2	PAX_ID (B)	PAX_ID (B)
3	Make_tour (C)	Make tour (C)
4	Tour ID (D)	Tour ID for sub-tour (D)
5	Begin trip ID (E)	Sub-tour ID in a tour (E)
6	End trip ID (F)	Begin trip ID (F)
7	Begin time of tour (G)	End trip ID (G)
8	End time of tour (H)	Sub-tour location (H)
9	Tour duration (I)	Is sub-tour location primary (I)
10	Primary activity (J)	Begin time of tour (J)
11	Primary activity duration (K)	End time of tour (K)
12	Primary activity location code (L)	Sub-tour duration (L)
13	PA first arrival time (M)	Primary activity (M)
14	PA last departure time (N)	Primary activity location (N)
15	Trip ID of PA (O)	Primary activity trip ID (O)
16	Primary work location (P)	Primary activity duration (P)
17	Secondary stops section (Q-AF)	Mode index (Q)
18	Primary mode (AG)	
19	Primary mode index (AH)	
20	Detected work-based sub-tour (AO)	
21	Non-home-originated flag (AP)	
22	Non-home-destined flag (AQ)	

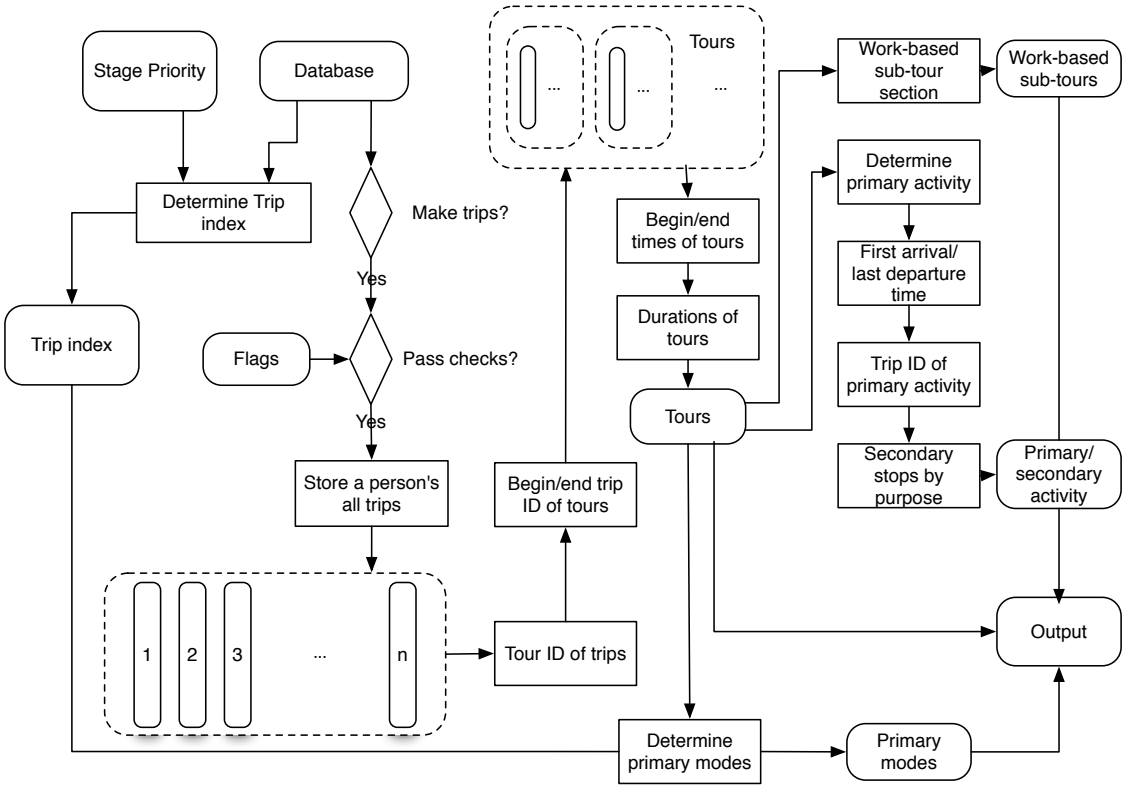


Figure 4.2: Logic flow of trip-to-tour conversion

column is left blank, they are eliminated from the input; Secondly, in the previous section, we run a few data checks. Among these checks, some will have a great impact in the process of trip-to-tour conversion due to inconsistency (for example, negative trip duration). Therefore, the data need to pass flag1, flag2, flag3, flag4, and flag12 before entering the trip-to-tour conversion process.

- **Store a person's trips**: It is noticed that each entry in the database is a stage of a trip. We do not need stage (mode) information to group trips into tours. So we only take the last stage of a trip to store related trip information as input. It contains all the information needed to link the trip to a tour.

- **Tour ID of trips**: Assigning a tour ID to a trip is simply done by first assigning a tour flag for each trip sequentially. 1 for "begin trip of the day", 2 for "end trip of the day", 3 for "begin trip of the tour", 4 for "end trip of a tour" and 0 for "non begin/end trip of a tour". tour ID has initial value 1 and does not increase until tour flag equals 4.

- **Non-home-originated/destined tour flag**: An advantage of assigning a tour flag to a trip is that we can easily get the information whether the tour is a home based one or a "non-home-originated/destined" one.

- **Begin/end time and duration of a tour**: The begin time of a tour is the begin time of the first trip in a tour. Likewise, the end time of a tour is the end time of the last trip in a tour. The duration is the time period in between.

- **Stage/trip priority**: Stage priority is the unique priority assigned to a travel mode. In HITS2008, the participants can choose from a choice set containing 15 modes (besides walking). They are public bus/company bus/school bus/MRT/LRT/taxi/car driver/car passenger/van, lorry driver/van, lorry passenger/cycle/motorcycle driver/ motorcycle passenger/shuttle bus/others. Only one mode is used in a stage. Since a trip may contain several stages, determining the mode of a trip is not simple.

  First of all, we need to define the modes of trips. 16 modes are defined and assigned a unique priority number, which is show in Table 4.8. Similar definition of trip modes and priority scheme is observed in the Denver (DRCOG) Model (Cambridge Systematics,

Inc., 2010). To determine the mode for a trip, first, we assign the priority to its stages with single mode. Then, we detect if these assigned stages belong to a combined mode like kiss & ride. After these two steps, the stage mode index is determined and the trip mode index is the highest priority in the trip stages.

Table 4.8: Trip mode priority scheme

Priority	Mode
1	Company/School/Shuttle bus
2	Bus (Company/School/Shuttle bus) & ride (MRT/LRT/Public bus)
3	MRT/LRT
4	Public bus
5	Kiss (Car/Van/Lorry/Motor passenger) & ride (MRT/LRT/Public bus)
6	Park (Car/Van/Lorry/Motor driver) & ride (MRT/LRT/Public bus)
7	Bike & ride (MRT/LRT/Public bus)
8	Car/Van/Lorry/Motor passenger
9	Car/Van/Lorry/Motor driver
10	Taxi
11	Motor passenger
12	Motor rider
13	Cycle
14	Walk
15	Others
16	Unknown

- **Primary mode**: Primary mode of a tour is the highest trip mode priority index among all of its trips.

- **Primary activity**: Primary activity may have several different definitions. It can be determined by the duration of activity or the priority of activity (by purpose) or both. For HITS2008, the primary activity of a tour is determined by both the priority and duration of activity. Specifically, it is assumed that the primary activity of a person is associated with its social economic characteristics and has an inherent hierarchical structure. For example, the primary activity of workers should be going to work. If we encounter a situation where two activities conducted in the same tour have the same priority, we then come to compare the activity duration. For HITS2008, the following simple hierarchy is adopted.

  If the person is a student, then the trip priorities are as follows:

  1. Education

  2. Go to work or work-related business

3. Shopping (maintenance)
4. Everything else
5. Return home

If the person is not a student, then the trip priorities are as follows:

1. Go to work or work-related business
2. Education
3. Shopping (maintenance)
4. Everything else
5. Return home

After assigning each trip a trip priority, we can select those with the highest priority. If multiple trips in a tour share the highest priority, then we use duration of the related activity to determine the primary activity.

- **Trip ID(s) of primary activity**: It is worth noting that tour primary activity may be separated in time (for example, a work tour with a work-based sub-tour), and has multiple trip IDs.
- **Secondary stops by purpose**: Besides primary activities, secondary activities or intermediate stops also need to be recorded in the tour table. We are interested in the number and purpose of these secondary stops.
- **Output results**: The output consists of a tour table in CSV format. The tour table, trip table, household and personal information are then coded into a relational database MySQL.

4.2.4 Work-based sub-tour detection

This section covers the work-based sub-tour detection in the logic flow in Figure 4.2. A work-based sub-tour by all means is similar to a tour. The only difference is that the origin

and destination of a work-based sub-tour is a work location. For a work-based sub-tour, we also need to provide the primary activity, primary mode and the duration of it.

During the coding process, it is found that there are two different kinds of work-based sub-tours. The first one is driven by the need for eating. Therefore, only 2 or 3 trips exist in the sub-tour. The other one is driven by the characteristics of the work. For example, a worker in a logistic company may go to office in the morning, then go to work outside the office until evening. In such process, more than 3 trips may be involved in the work-based sub-tour.

Different from the trip-to-tour conversion process, sub-tours need to be detected first from the trip chains. Then, the methods used in the trip-to-tour conversion process to obtain primary activity and primary mode information can be applied to sub-tours.

Definitions Some definitions that are unique to work-based sub-tours are listed below.

- **Sub-tour ID in a tour**: Without loss of generality, we consider the situation where people may conduct more than one sub-tour during a tour. Therefore, it is necessary that every sub-tour in a tour should have a unique ID. It turns out that the multiple-sub-tour situation indeed appears in HITS2008.
- **Sub-tour location**: A sub-tour location is the place where the sub-tour starts and ends. It is worth noting that this sub-tour location may not be the primary work location. We do have such examples for workers to conduct sub-tours at non-primary work locations.
- **Sub-tour duration**: Sub-tour duration is simply the time period between the beginning of the first trip and the end time of the last trip in the sub-tour.

Entries in work-based sub-tour table The work-based sub-tour table contains entries as shown in table 4.7.

Methods The sub-tour detection begins with determining the primary work location. For people claiming that they have a fixed work location, the location is by default the primary

one. For others, the work location with the longest work activity duration is nominated as the primary work location.

In doing this, all work locations of a person are stored. Then, one work location by one work location, in each work tour of the person we detect if sub-tours exist. This search is exhaustive and all sub-tours will be detected. When the start/end trip ID of a sub-tour is determined by the detection, the methods used in the trip-to-tour conversion process are applied to determine other characteristics of a work-based sub-tour.

4.3 Descriptive Analysis and Insights

In the previous section, trip-to-tour conversion and work-based sub-tour detection are carried out on the traditional trip-based travel survey. The processed survey reveals more perspectives for researchers to understand travelers' diverse life styles and trip-making decisions. Based on the new data structure, such as tours, descriptive analysis is carried out in this section and insights gained in the process are extremely valuable for grasping the whole picture of traveling and activity-participation in the study area. The analysis in this section is carried out at three levels: trip level, tour level and person-day level, representing different levels of abstraction of the original trip information recorded in HITS2008.

4.3.1 Trip level analysis

On average, 1.81 trips are made by each individual on daily basis. Figure 4.3 shows the average number of trips made for each person type. Generally speaking, workers and students make more trips than the rest. Domestic workers live and work at customer's home thus have the least desire for traveling.

Table 4.9 reveals the trip mode share by different trip purposes. Overall, public transportation (including bus and MRT/LRT), vehicle driver, vehicle passenger and walk are among the several popular travel modes in Singapore. On the other hand, combined modes (such as kiss and ride, park and ride) are chosen by few. Private bus (company/school/shuttle bus) is popular for work and education trips. Although the overall mode share for public

Table 4.9: HITS2008 trip mode share by trip purpose

Modes	Selected trip purposes (mode share)						Total trips (mode share)
	Go to work	Education	Shopping	Meal	Personal errands	Escort	
Company/school/shuttle bus	118,115(7.76%)	118,728(13.39%)	2,680(1.21%)	3,677(2.15%)	1,113(1.87%)	3,503(0.72%)	429,984(6.30%)
Bus and ride	2,847(0.19%)	778(0.09%)	1,500(0.68%)	201(0.12%)	0(0.00%)	734(0.15%)	65,810(0.96%)
MRT/LRT	356,976(23.44%)	109,289(12.33%)	43,465(19.66%)	16,519(9.68%)	9,659(16.23%)	7,979(1.64%)	1,162,717(17.02%)
Public bus	326,079(21.41%)	214,689(24.21%)	82,479(37.31%)	23,328(13.67%)	15,933(26.78%)	20,557(4.23%)	1,547,653(22.66%)
Kiss and ride	10,467(0.69%)	3,203(0.36%)	203(0.09%)	527(0.31%)	109(0.18%)	102(0.02%)	23,892(0.35%)
Park and ride	1,172(0.08%)	0(0.00%)	93(0.04%)	0(0.00%)	0(0.00%)	100(0.02%)	1,667(0.02%)
Bike and ride	3,953(0.26%)	1,230(0.14%)	208(0.09%)	0(0.00%)	100(0.17%)	0(0.00%)	6,988(0.10%)
Car/van/lorry passenger	120,480(7.91%)	150,250(16.95%)	25,955(11.74%)	46,655(27.34%)	8,941(15.03%)	58,241(11.99%)	774,789(11.34%)
Car/van/lorry driver	382,819(25.14%)	6,320(0.71%)	38,455(17.40%)	52,422(30.72%)	15,278(25.68%)	337,191(69.43%)	1.517,514(22.22%)
Taxi	33,087(2.17%)	7,724(0.87%)	8,860(4.01%)	3,663(2.15%)	3,904(6.56%)	9,616(1.98%)	184,706(2.70%)
Motor passenger	2,885(0.19%)	1,789(0.20%)	678(0.31%)	631(0.37%)	0(0.00%)	213(0.04%)	13,361(0.20%)
Motor rider	57,492(3.77%)	1,287(0.15%)	2,469(1.12%)	3,301(1.93%)	773(1.30%)	7,801(1.61%)	152,283(2.23%)
Cycle	23,514(1.54%)	8,221(0.93%)	2,831(1.28%)	2,261(1.32%)	560(0.94%)	4,796(0.99%)	88,759(1.30%)
Walk	47,814(3.14%)	233,630(26.35%)	4,481(2.03%)	1,947(1.14%)	1,204(2.02%)	17,145(3.53%)	621,711(9.10%)
Other/missing	35,267(2.32%)	29,482(3.33%)	6,702(3.03%)	15,528(9.10%)	1,921(3.23%)	17,677(3.64%)	238,380(3.49%)
Total	1,522,967(100%)	886,620(100%)	221,059(100%)	170,660(100%)	59,495(100%)	485,655(100%)	6,830,214(100%)

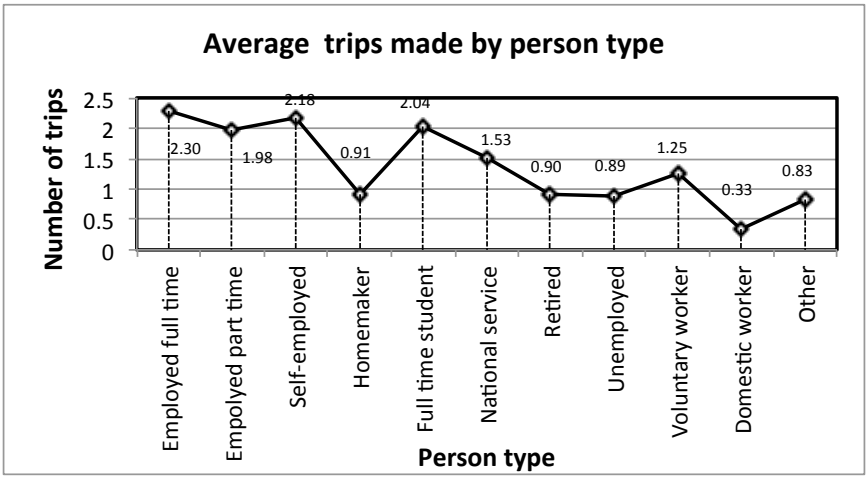
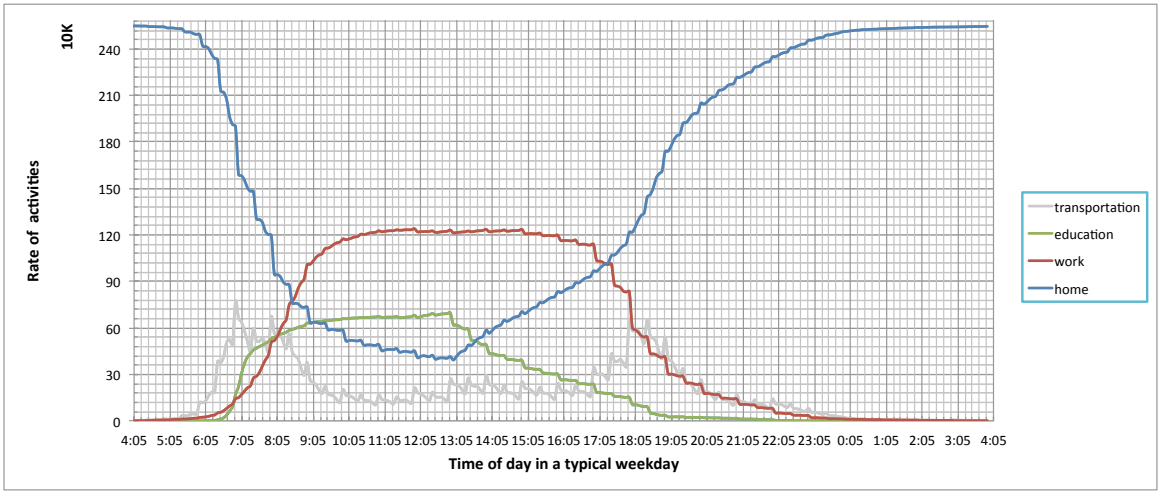


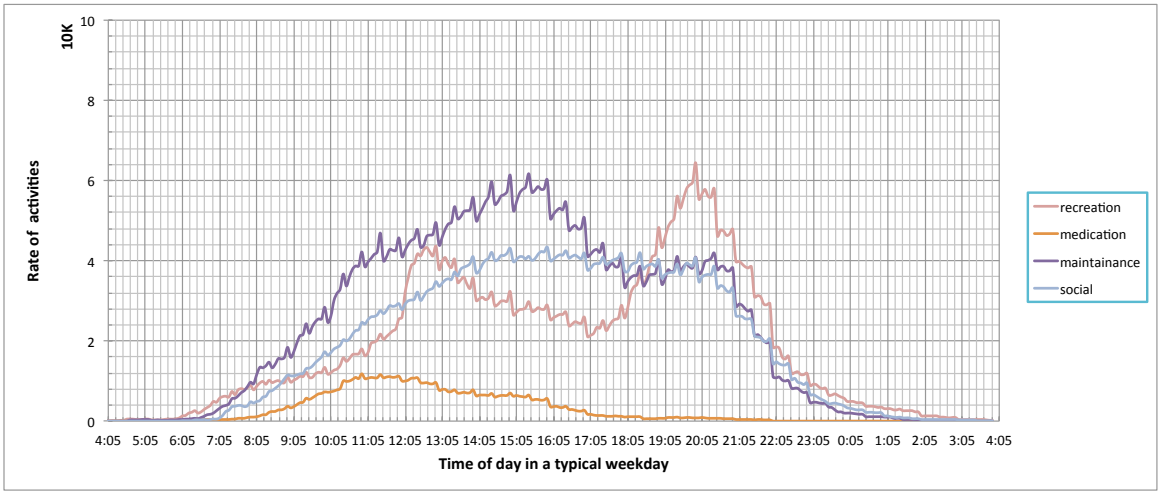
Figure 4.3: Average number of trips for each person type

transportation is high, the relative share of public bus and MRT/LRT is different for different trip purposes. While MRT/LRT is more popular for work trips for its higher level of service during peak hours, public bus has a higher share for education, shopping, meal and personal errands trips. Besides, education trips have the largest share for walk, which may reflect the fact that the walkability from home to school is a major consideration for parents.

People make trips to perform certain activities. Given the nature of the participated activities, the temporal patterns will be unique. Among all those who make at least 1 trip during the survey day, the temporal patterns of different activities are explored. Slices of time are taken and the number of a particular type of activities that are in progress is recorded. Figure 4.4 displays the time profile of 8 different activities (notice that only people with at least 1 trip are recorded in the figure). It can be easily spotted that there are two major peaks of travel: morning peak and evening peak. However, there are two individual peaks at morning peak hours with a time difference of one hour. It is very interesting to notice that the first morning peak (at around 7 am) coincides to the time when the increasing speed of on-going education activities is fastest and the second morning peak (at around 8 am) coincides to the time when the increasing speed of on-going work activities is fastest. Same phenomenon is observed at evening peak when the peak is caused by a fast decreasing of work activities. Although the peak for education activities appears earlier than work activities, education activities generally last less hours than work activities and start to fall at 1pm. It is also



(a) Major activities (home, work, education, travel)



(b) Minor activities (recreation, social, maintenance, medication)

Figure 4.4: Temporal patterns of activities in HITS2008

interesting to observe that the decreasing speed of home activities in the morning is higher than the increasing speed of home activities in the evening, indicating a shorter and tougher peak hour for Singapore in the morning. In terms of the minor activities, although over 200k expanded shopping activities are observed, the peak for shopping is only 60k, indicating a well-spread time span for shopping. At last, the two peaks for recreational activities appear at noon and 8pm, probably right after lunch and dinner, respectively.

4.3.2 Tour level analysis

In HITS2008, out of 63,433 trips (expanded to 6,830,214 trips for the whole island), 28,020 tours are generated (expanded to 3,013,910 for the whole island). On average, 2.26 trips are made in a tour. As can be seen in Figure 4.5, over 83 percent of tours contain only 2 trips and for work tours and education tours, this percentage becomes 81 and 93 respectively. 78.8 percent of people make simple tours with no intermediate stops, which may due to the high coverage of public transportation in Singapore since private motorized vehicles tend to make more intermediate stops in tours. Among all the tour purposes, work-related business tours have the highest average number of trips per tour, which is 2.68. Generally speaking, the trip chaining characteristic observed in HITS is simpler than it was observed in Sabina et al. (2006).

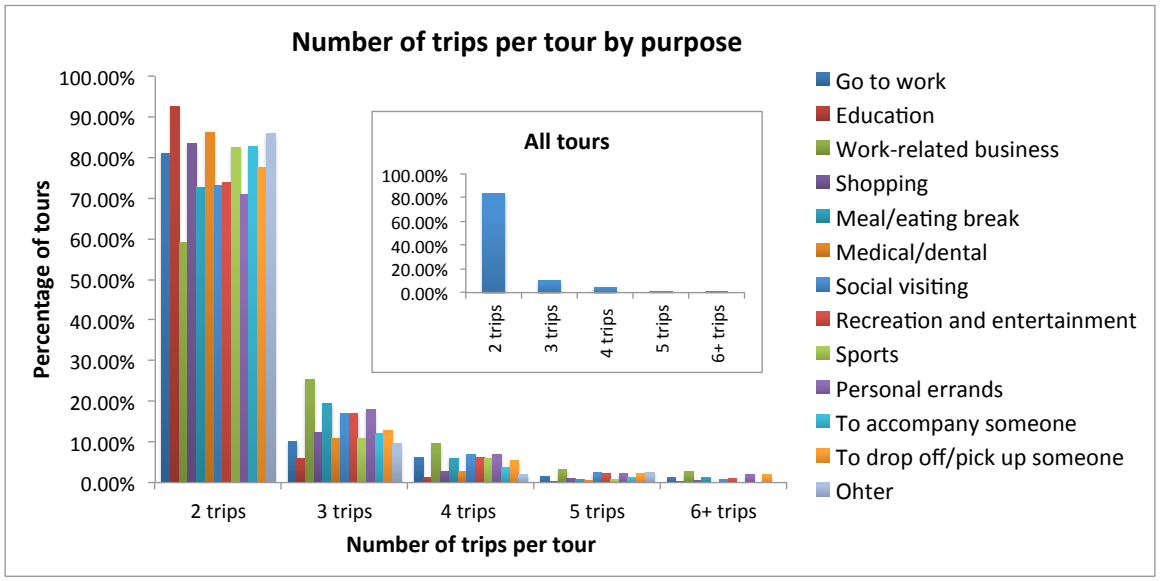


Figure 4.5: Number of trips per tour by purpose of tour

Out of 28,020 generated tours, only 405 work-based sub-tours are detected. The low rate of work-based sub-tour may be due to the fact that respondents underreported walk trips during work tours. All sub-tours are made by full-time, part-time employees and self-employed personals and over 87 percent of the sub-tours are made by full-time employees. Part-time employees are the least likely to make sub-tours, probably because of their shorter duration of stay at work.

Table 4.10: HITS2008 tour mode share by tour purpose

Modes	Selected tour purposes (mode share)						Total tours (mode share)	Total sub-tours (mode share)
	Go to work	Education	Shopping	Meal	Social visiting	Escort		
Company/school/shuttle bus	122,591(8.77%)	127,414 (15.05%)	1,862(1.15%)	431(0.65%)	2,408(2.49%)	901(1.05%)	265,304(8.80%)	2,211(5.02%)
Bus and ride	7,641(0.55%)	507(0.06%)	324(0.20%)	0(0.00%)	95(0.10%)	0(0.00%)	9,779(0.32%)	0(0.00%)
MRT/LRT	383,380(27.41%)	126,493(14.94%)	31,348(19.29%)	6,003(9.12%)	25,161(26.02%)	3,170(3.69%)	624,903(20.73%)	2,926(6.64%)
Public bus	316,731(22.65%)	252,207(29.78%)	68,623(42.23%)	13,584(20.64%)	29,458(30.46%)	11,022(12.85%)	762,493(25.30%)	2,460(5.58%)
Kiss and ride	2,834(0.20%)	95(0.01%)	110(0.07%)	0(0.00%)	0(0.00%)	0(0.00%)	3,359(0.11%)	0(0.00%)
Park and ride	119(0.01%)	0(0.00%)	0(0.00%)	0(0.00%)	0(0.00%)	0(0.00%)	119(0.00%)	0(0.00%)
Bike and ride	219(0.02%)	0(0.00%)	0(0.00%)	0(0.00%)	0(0.00%)	0(0.00%)	219(0.01%)	0(0.00%)
Car/van/lorry passenger	83,962(6.00%)	76,998(9.09%)	18,255(11.23%)	19,017(28.90%)	18,237(18.86%)	3,665(4.27%)	257,078(8.53%)	5,324(12.08%)
Car/van/lorry driver	341,514(24.42%)	5,509(0.65%)	25,737(15.84%)	20,363(30.94%)	15,210(15.73%)	48,366(56.37%)	583,741(19.37%)	21,434(48.63%)
Taxi	11,969(0.86%)	2,614(0.31%)	5,495(3.38%)	1,133(1.72%)	3,260(3.37%)	2,164(2.52)	45,625(1.51%)	2,206(5.01%)
Motor passenger	1,595(0.11%)	958(0.11%)	436(0.27%)	422(0.64%)	109(0.11%)	0(0.00%)	4,022(0.13%)	0(0.00%)
Motor rider	52,415(3.75%)	1,184(0.14%)	1,810(1.11%)	409(0.62%)	966(1.00%)	1,134(1.32%)	68,386(2.27%)	1,068(2.42%)
Cycle	21,003(1.50%)	7,767(0.92%)	2,418(1.49%)	1,827(2.78%)	670(0.69%)	1,916(2.23%)	42,053(1.40%)	550(1.25%)
Walk	44,810(3.20%)	228,482(26.98%)	4,154(2.56%)	1,538(2.34%)	1,041(1.08%)	8,953(10.43%)	304,737(10.11%)	331(0.75%)
Other/missing	7,660(0.55%)	16,608(1.96%)	1,914(1.18%)	1,077(1.64%)	95(0.10%)	4,508(5.25%)	42,092(1.40%)	5,564(12.62%)
Total	1,398,443(100%)	846,836(100%)	162,486(100%)	65,804(100%)	96,710(100%)	85,799(100%)	301,3910(100%)	44,074

The distribution of tour mode among all given modes and mode share for work-based sub-tours are shown in Table 4.10. For home-based tours, 3 most common modes are MRT/LRT, public bus and vehicle driver. 46 percent of tours are made by public transportation (MRT/LRT and public bus). Notice the difference in mode shares between trips (shown in Table 4.9) and tours (shown in Table 4.10). The difference is rooted in the prioritization scheme used to define the primary mode of tours. For example, vehicle passenger is of lower priority than public transportation. For a tour containing a trip made by vehicle passenger, if any other trip is made by public transportation, the tour mode will not be vehicle passenger. In terms of the mode share for work-based sub-tours, vehicle driver and passenger are the dominant modes.

4.3.3 Person-day level analysis

At person-day level, the trip-chaining behavior for each individual on daily basis is of interest. As indicated in Table 4.11, number of tours made during the day shows significant heterogeneity by person type. Respondents make an average number of 0.8 tours per day. 25.7 percent of all respondents make no tour at all. 69.2 percent of all respondents made 1 tour and 5.1 percent of all respondents make 2+ tours. Full time employees and students are likely to make at least 1 tour during the day. The retired and unemployed show great similarities in terms of number of tours made on daily basis.

Table 4.11: Number of home-based tours by person type

Person type	Number of tours (%)			
	0	1	2	3+
Employed full time	7.03	87.84	4.60	0.53
Employed part time	18.83	73.34	6.34	1.49
Self-employed	26.85	61.31	9.76	2.08
Homemaker	68.36	24.84	4.87	1.93
Full time student	6.64	90.09	3.15	0.12
National service	21.52	76.58	1.90	0.00
Retired	64.01	31.49	3.74	0.76
Unemployed	64.18	32.58	3.00	0.24
Domestic worker	87.87	8.92	3.07	0.14

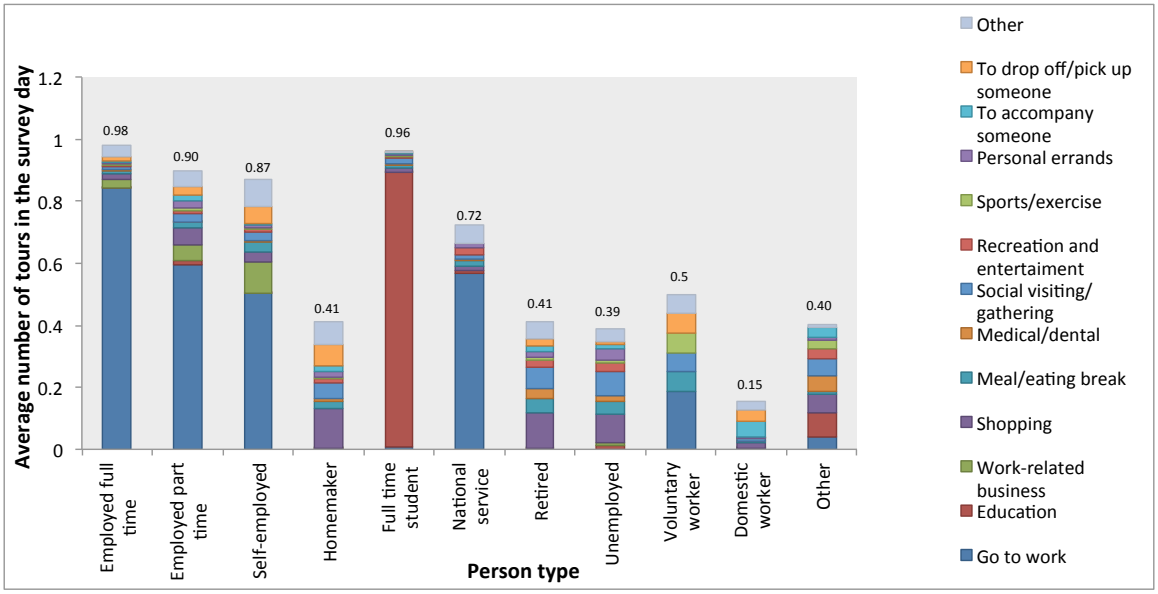


Figure 4.6: Average number of tours by person type and tour purpose

Figure 4.6 shows the average number of tours by person type and tour purpose, which again shows great heterogeneity among all person types. Employees and students make more tours than other people and the main purpose of these tours is work and education respectively. For person types other than employees and students, tour purposes are not concentrated on one particular purpose.

To further illustrate the relationship between trip-making/activity participation and person type, we refer to daily activity patterns. Follow the practice of the reviewed frameworks in 2.4.1, daily activity patterns are the abstraction of activity participation on daily basis. It is defined to create an abstraction of higher level than tours. Here we can define that the day pattern of a person is the occurrence of tours (0, 1+) and intermediate stops (0, 1+) for 10 given purposes[2]. By adopting this definition, the dominant day patterns for different person types can be determined. It is assumed that 10 activity purposes can be assigned to tours and trips as primary activity purpose of a tour and purpose of a trip respectively. One can make 0 or 1+ tours for each of the 10 purposes and 0 or 1+ intermediate stops for each of

[2]Similar definition of daily activity patterns can be found in the Sacramento (SACOG) Model, Denver (DRCOG) Model and Seattle (PSRC) Model, etc. See Table 2.1 for these models. Originally, HITS has 13 activity types as shown in Figure 4.6. Some of them are grouped together (for example, to drop off/pick up someone and to accompany someone are grouped together) such that we have 10 activity types left.

Table 4.12: Top 3 day patterns by person type

Person type	Top 3 day patterns					
	No. 1	%	No. 2	%	No. 3	%
Employed full time	1+ work tour 0 stop	64.68	0 tour 0 stop	7.03	1+ work tour 1+ escort stop	5.85
Employed part time	1+ work tour 0 stop	48.37	0 tour 0 stop	18.83	1+ shopping tour 0 stop	3.36
Self-employed	1+ work tour 0 stop	31.81	0 tour 0 stop	26.85	1+ work-related business tour 0 stop	4.62
Homemaker	0 tour 0 stop	68.36	1+ shopping tour 0 stop	9.79	1+ social/recreation tour 0 stop	4.81
Full time student	1+ education tour 0 stop	78.77	0 tour 0 stop	6.64	1+ education tour 1+ escort stop	1.73
National service	1+ work tour 0 stop	53.16	0 tour 0 stop	21.52	1+ other tour 1 stop	8.23
Retired	0 tour 0 stop	64.01	1+ shopping tour 0 stop	9.01	1+ social/recreation tour 0 stop	7.44
Unemployed	0 tour 0 stop	64.18	1+ social/recreation tour 0 stop	8.29	1+ shopping tour 0 stop	6.25
Domestic worker	0 tour 0 stop	87.87	1+ escort tour 0 stop	5.23	1+ shopping tour 0 stop	1.19

the 10 purposes, which theoretically results in 2^{20} alternatives for day pattern. Apparently a majority of these patterns are not feasible in reality. Only 579 unique day patterns are observed in HITS2008. As can be seen in Table 4.12, the top 3 day patterns show great heterogeneity by person type. The No.1 pattern for a person type can be viewed as the expected stereotypical pattern, but the percentage of choosing the stereotypical pattern is significantly different from 100 percent. For employees and students, the "0 tour, 0 stop" day pattern is an indication of the extent of telecommuting or absenteeism from work or school on a given weekday.

4.4 Summary

This study investigates a particular data issue related to the development of activity-based models: Whether trip-based surveys can be used to provide the data in need for the implementation of activity-based models. We give positive conclusion with a detailed walkthrough in this chapter. 2008 Household Interview and Travel Survey is the regular household travel survey carried out by the Land Transport Authority in Singapore. It was designed for the purpose of developing the four-step models. However, as we have shown in

the study, it is possible to encode tours with a trip-to-tour conversion process and detect work-based sub-tours. It is an immediate suggestion that the processed HITS data can be used to support the implementation of activity-based models in Singapore.

The trip-to-tour conversion and work-based sub-tour detection process is not unique for HITS2008. A program written in Python is further developed to automatically carry out the process with necessary data checks and data field clarification done. By using the code, the trip-to-tour conversion process for HITS2008 can be finished in a few seconds. Moreover, it can be applied to conduct activity-based analysis of consecutive HITS surveys.

Although it is shown that the current trip-based survey has been capable of tour encoding, a couple of suggestions come up during the study that can refine the survey design and better serve the purpose of activity-based model development.

1. The first one is applicable to all internet-based and trip-based household travel surveys. A number of checks are carried out before the trip-to-tour conversion process as the data inconsistency in the trip-based surveys can hinder the tour encoding process and has negative impact in the stage of mode estimation. It is suggested that with an interactive survey platform, such as internet survey, the system at backend could detect such inconsistency in the whole process and issue necessary warnings to the participants for corrections. Although the survey is still trip-based, it would has been justified for better data consistency.

2. The second suggestion is on the consistency of survey design. In HITS2008, respondents were asked to report if they made **any** motorized trips during the day before providing any trip information. For those who did not make motorized trips, the respondents were asked to provide trip information without indicating the trip mode (by default it is walk). However, this method puts those who made both motorized trips and walk trips into a dilemma. On one hand, they should proceed normally as the answer to the above question is positive. On the other hand, there is no option for walk when they need to choose trip mode in the section of trip particulars. The simple logic error was ignored and directly responsible for trip mode missing and trip underreporting. For example, the missing rate for work-based sub-tour is as high as 12.62 percent

(shown in Table 4.10).[3] The logic error did not occur at HITS2004 and was corrected at HITS2012. As a result, neither HITS2004 nor HITS2012 has such high trip mode missing rate. Although the logic error is unique to HITS, it applies to all consecutive surveys that good consistency of survey design is the key to correctly understand and interpret travel and activity patterns in a large time span.

3. The third suggestion is on the explicit recording of intra-household interactions in the trip-based survey. Currently, the HITS surveys used in Singapore fail to incorporate intra-household interactions in the survey, which will make it difficult to model those interactions in the stage of model development. The trip-to-tour conversion method introduced in this chapter will not solve the issue and the data records will remain intra-household disconnected. Therefore, intra-household connected trip-based surveys are appreciated and the trip-to-tour conversion method can further check the consistency of such interactions.

4. The last suggestion is on the transferability of this study. Although the data preparation work presented in this chapter is very specific to Singapore, the method of trip-to-tour conversion can be applied to other trip-based surveys as long as the basic elements of trip, such as timing, location, and mode are identified from those surveys.

[3] 207 out of 405 work-based sub-tours are for meal, which usually take place within a reasonable distance to walk. Trips associated with such work-based sub-tours are forced to have an empty mode. In fact the work-based sub-tour detection leads us to find the logic error.

Chapter 5

On the Design and Implementation of an Activity-based Travel Demand Model for Singapore

SimMobilityMT is one of the very recent integrations of a fully utility-based activity-based travel demand modeling framework with a simulation-based Dynamic Traffic Assignment (DTA) package. As part of the research and development efforts for SimMobilityMT, this chapter presents the design and implementation of a full-fledged activity-based travel demand simulator, or the pre-day simulator in SimMobilityMT[1].

The pre-day activity-based model behind the simulator is formulated through a system of interconnected discrete choice models representing choices at distinct dimensions of daily activity schedule. The system design, estimation of individual components, and interface between the model and the simulator are introduced in detail. Key features in terms of the implementation of the pre-day simulator, such as modularization, parallelization and multiple running modes are highlighted. With the pre-day model estimated and the simulator implemented, model calibration/validation process is carried out against the activity participation and travel behavior of the base year. It is shown that the pre-day

[1]The development of the pre-day model has been team-work efforts. In terms of the design, I am mainly responsible for preparing data, developing pre-day framework and estimation of the majority of models in the pre-day modeling framework and the design of simulation flow while two Post-docs, Carlos Carrion and Muhammad Adnan, have been working on refining model specifications and model re-estimation, developing model validation modules. We work closely in the stage of model development. In the stage of implementation, A software engineer, Harish Loganathan, is involved for implementing the framework and all the team have been involved in clarifying the implementation details. Finally, the whole development process has been under the supervision of Moshe Ben-Akiva and other research PIs. The author would like to thank them for their consent to put related contents and results in Chapter 5.

simulator is capable of correctly replicating the behavior of the base year. Finally, avenues of future works are mentioned to improve the capabilities of the pre-day model as well as SimMobilityMT.

5.1 Introduction

Travel demand estimation is the first step towards regional travel planning. In order to achieve system objectives such as efficiency, equity, reliability, it is necessary to obtain accurate and precise estimation of future travel demand and its response to various supply-side (e.g., adding new roads to the network) and policy (e.g., free transit for certain hours) changes. A good travel demand model should be theoretically sound, empirically verifiable, cheap to implement, capable of prediction, and policy-sensitive (Hensher and Button, 2000). Activity-based models come at the crossroads of an increasing need to analyze responses to supply and policy changes as well as increasing investment of transportation in urban areas, and a booming of microscopic simulation in demand modeling with special attention to the understanding of activity-travel patterns. Moreover, as shown in the literature review in Chapter 2, the activity-based approach to demand modeling is at the edge of replacing the traditional four-step method and becoming a powerful planning tool for MPOs worldwide.

Historically speaking, the research efforts that attempt to improve the traditional four-step method have been disjointed (Vovsha, 2009). On one hand, most of the early efforts are made to the replacement of equilibrium-based assignment of traffic with simulation-based dynamic traffic assignment (DTA) process. To name some, CORSIM (Owen et al., 2000), MITSIM (Yang, 1997), AIMSUN2 (Barceló et al., 1994) and VISSIM[2] are micro-simulation based DTA packages. And CONTRAM (Leonard et al., 1989), DYNASMART (Mahmassani et al., 1992), DYNAMIT (Ben-Akiva et al., 1998) are the examples of mesoscopic DTA packages. Despite the scope of the DTA packages, all those efforts focus on travel supply, and travel demands are not explicitly modeled (usually adopt the demand generated from the first three steps of the four-step method). On the other hand, from the perspective of travel demand, the research efforts are focused on replacing the trip-based approach in the four-step method with more sophisticated activity-based travel demand modeling frameworks, which have been reviewed in Chapter 2. However, empirically, the outputs from activity-based models are aggregated into time-dependent OD matrices that can be fed into DTA packages for traffic assignment (see for example, DKS Associate et al., 2012). The disaggregate representation

[2] http://vision-traffic.ptvgroup.com/en-us/products/ptv-vissim/

of individual travelers and their decisions gained through the application of activity-based models are lost in the integrated OD matrices, and the explicit representation of rescheduling and re-routing decisions, as well as the individual-based performance measures are not allowed (Balmer, 2007).

It was not until recently that the efforts of seeking an integration of travel demand and supply within the same framework and a consistent individual-based representation of travelers throughout the whole process started to emerge. Some examples include TRANSIMS (Gliebe, 2006), MATSim (Balmer et al., 2006), FEATHERS (Bellemans et al., 2010) and SimTRAVEL (Pendyala et al., 2010). Moreover, the study presented in this chapter is part of the research and development efforts for SimMobilityMT (MT is abbreviation for Mid Term), which is one of the very recent integrations of a fully utility-based activity-based demand modeling framework with a simulation-based DTA package (Lu et al., 2015).

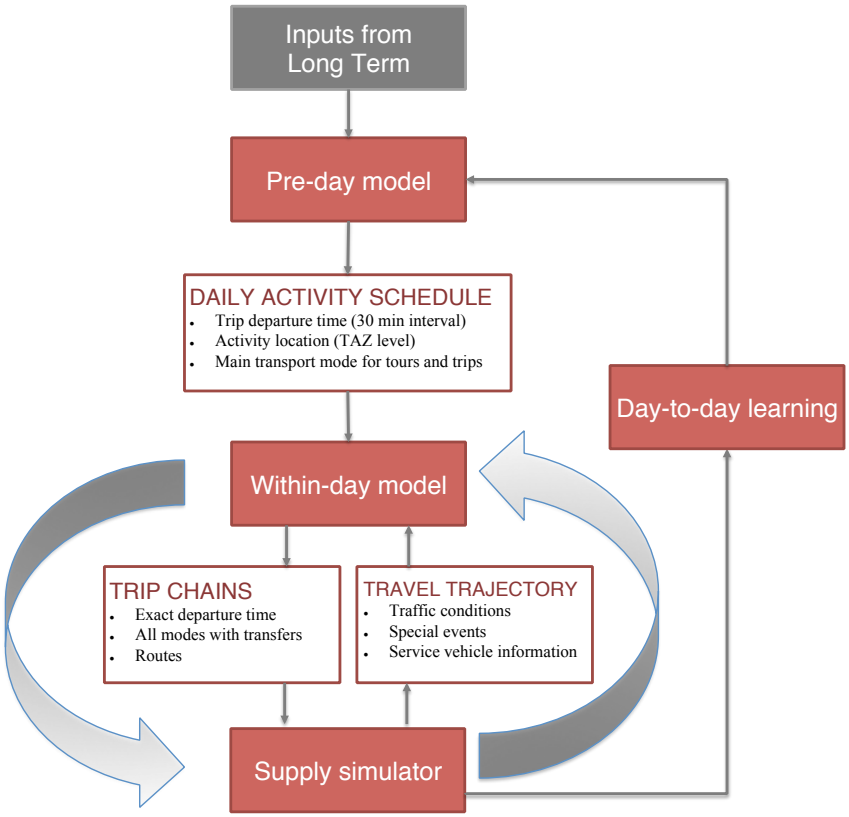


Figure 5.1: SimMobility Mid Term (SimMobilityMT) modeling framework (adapted from Lu et al., 2015)

Figure 5.1 shows the high level architecture of SimMobilityMT. The SimMobilityMT simulator

requires the population with socio-demographic characteristics, vehicle ownership, and household characteristics, along with land use information from the long-term simulator. SimMobilityMT has three major components: pre-day, within-day and supply simulator. From Figure 5.1, a workflow is observed: Firstly, the pre-day simulator is responsible for the generation of daily activity schedules comprising the planned activities with corresponding location, time of day and travel mode. Then, the output of the pre-day simulation is fed into the within-day simulator majorly for route choice and rescheduling. Finally, the detailed trip chains with exact departure time and routes will be fed to the supply simulator for DTA.

The focus of this chapter is on the design and implementation of a full-fledged activity-based demand simulator, or the pre-day simulator in SimMobilityMT, which is formulated through a system of interconnected discrete choice models representing choices at distinct dimensions. For the rest of the chapter, only components of the pre-day simulator will be discussed while other components of SimMobilityMT are exogenous. Section 5.2 provides an overview of the activity-based modeling framework and system design. Section 5.3 introduces the research efforts related to the development and estimation of the pre-day activity-based modeling system. Section 5.4 presents the results from model calibration and validation. Finally, the last section summarizes the chapter with discussion on the current progress as a benchmark and future works of the pre-day simulator and SimMobilityMT.

5.2 Model Framework and System Design

This section outlines the overall framework and model structure of the pre-day activity-based travel demand model. The pre-day model follows the Day Activity Schedule approach (Bowman, 1998 and Bowman and Ben-Akiva, 2001) and is formulated through a system of interconnected discrete choice models focusing on decisions related to daily activity and mobility. The overall model structure, overview of models of different levels, data requirements, accessibility measures and simulator design will be covered in this section.

5.2.1 Framework overview

People are making activity and travel decisions in many dimensions. Table 5.1 summarizes the decisions faced by individuals. Long-term decisions refer to the decisions that take longer period to prepare and take effect. Those decisions are often the results of accumulation of stimulus. For example, household location and household vehicle ownership are considered to be reversible at considerable costs. Mid-term decisions refer to the decisions that are made and updated on daily basis. Decisions in this group focus on daily activity and mobility. Short-term decisions capture the decisions made during implementing mid-term decisions. Those decisions are often the results of changing environment and unexpected events.

Table 5.1: Decision-makings in different dimensions

Long-term decisions	Mid-term decisions	Short-term decisions
Social role	Participation of activities	Adjustment
Residential location	Activity type	Route choice
Vehicle ownership	Activity location	Reaction to stimulus
Marriage status	Activity timing and duration	
Household composition	Trip chaining	
Lifestyle	Travel mode	
Social network		

The pre-day model is a disaggregate travel demand system designed to model the decisions at mid-term level. The long-term decisions are input to the system and thus are exogenous. Figure 5.2 and Figure 5.3 show the model components and process flow of the pre-day model, respectively. Synthetic population with known socio-demographic characteristics is an input to the system. Other inputs include network skims, land use characteristics, etc. Those inputs are discussed subsequently.

There are three different hierarchies in the system: day pattern level, tour level and intermediate stop level. Each level consists of several models. Given the complicated and multi-dimensional nature of the day activity schedule problem, curse of dimensionality will occur if every choice dimension is put into one single choice model. Therefore, choice dimensions are isolated, structured, and put into different choice models. The overall system can be viewed as a hierarchical (or nested) series of choice models. Figure 5.3 highlights the hierarchical structure of the model. The solid arrows indicate that models from lower

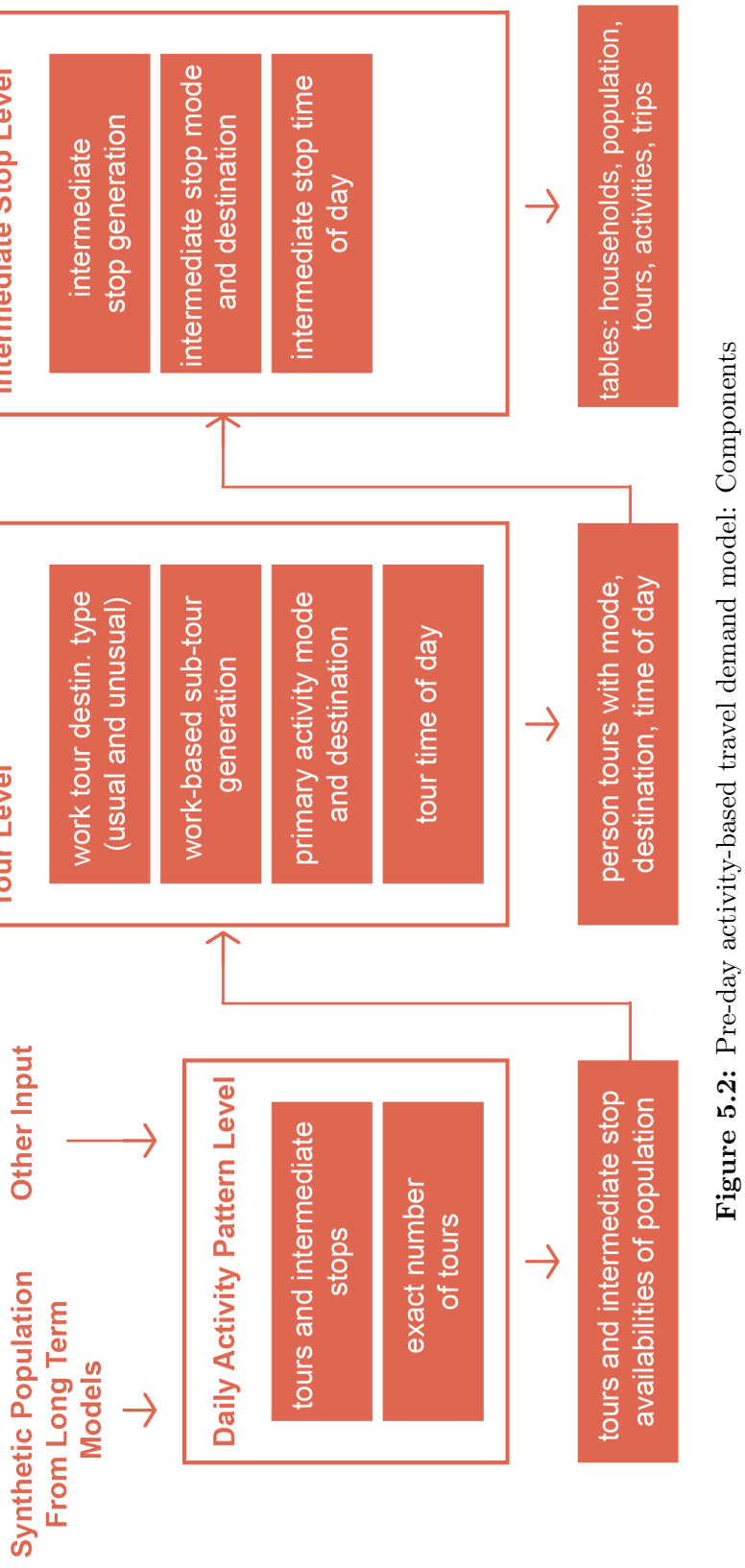


Figure 5.2: Pre-day activity-based travel demand model: Components

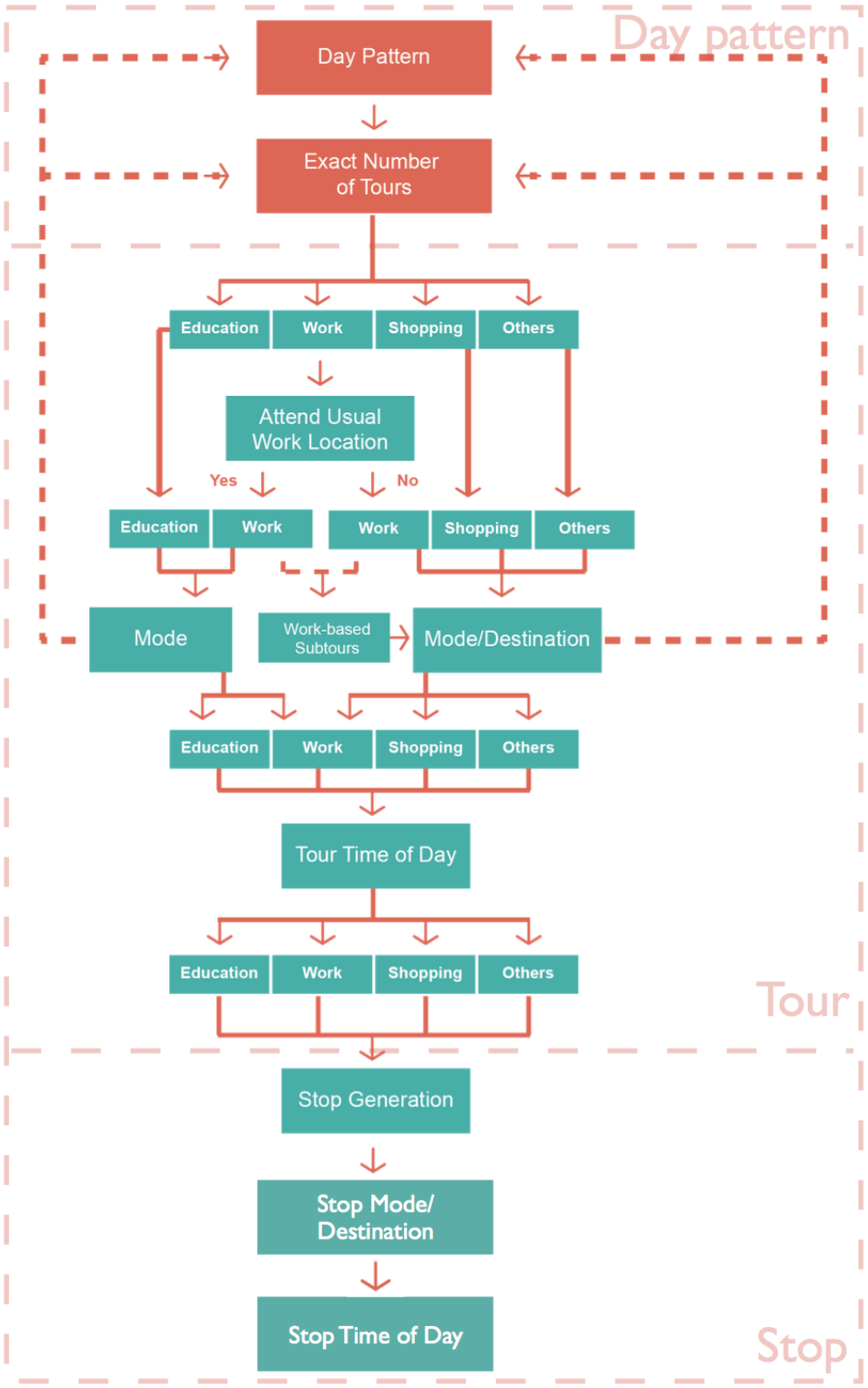


Figure 5.3: Pre-day activity-based travel demand model: Process flow

levels are conditioned on decisions made with models from higher levels. The dashed arrows represent the accessibility measures described subsequently.

Day pattern level This level distinguishes the pre-day model (an activity-based model) from tour-based models (see for example, Algers et al., 1995 and Cascetta and Biggiero, 1997) because it organizes tours and manages their sequence through the concept of day activity schedule. Day activity schedule is defined through the concepts of activity pattern, and activity schedule. Activity pattern defines the participation in activities as primary, and secondary. Primary activities are the anchors (e.g., home to work trip, and work to home trip represent a tour with work as primary activity) of tours, and secondary activities are intermediate stops within a particular tour (e.g., stopping for shopping at a work-to-home half-tour). Activity schedule adds detailed information about tours to the activity pattern such as sequence, timing, travel mode, destination of primary activity, and also the stops for secondary activities within tours.

In the pre-day activity-based travel demand model, the day pattern level (see Figure 5.2) includes two types of discrete choice models: day pattern model, and exact number of tours model for different primary activity purposes. The day pattern model predicts occurrence of tours for various purposes and availability of intermediate stops for various purposes. The purposes are defined by four activity types: work, education, shopping, and others. Tour purposes that are predicted to occur will be passed to a second model to determine the exact number of tours for that purpose. The predicted availability of intermediate stops has no immediate effect at day pattern level. However, the results will be provided to intermediate stop generation model to constrain the availability of each activity purpose. Day pattern level will generate a list of tours as well as intermediate stop availabilities for each individual in the synthetic population.

Tour level A tour is defined by a set of trips with the origin of the first trip and the destination of the last trip being home. In other words, tours are home-based, except for tours predicted by the work-based sub-tour model, which are work-based.

In the pre-day activity-based travel demand model, the tour level includes multiple discrete

choice models: usual/unusual work location; travel mode choice or travel mode/destination choice; work-based sub-tour generation; tour time of day. These models provide detailed information for each predicted tour. These details include destination, travel mode, time of day (arrival time and departure time). For every work tour, there is a specific model at this level to determine whether a sub-tour is going to be scheduled. For the purpose of modeling the tour-level decisions for work-based sub-tours, two additional sub models for work-based sub-tours, namely sub-tour mode/destination choice and sub-tour time-of-day choice, should be specified. Sequencing of tours is accomplished before modeling individual tours by assigning each of the tours predicted at day pattern level a priority number. The priority number is determined by the purpose of tour primary activity (as introduced in Chapter 4). It should also be noted that the time of day models for tours are based on the concept of cyclical and continuous indirect utility functions (Ben-Akiva and Abou-Zeid, 2013). In summary, this level provides activity and travel information for tours.

Intermediate stop level A basic tour contains only 2 trips: the first one from home to tour primary activity and the second one from tour primary activity to home. A more realistic tour structure should consider the existence of intermediate stops during a tour. The intermediate stop level will generate these stops. Trips for secondary activities are represented as intermediate stops within a tour, and the available types of secondary activities have been predicted in the day pattern model.

The intermediate stop level (see Figure 5.2) includes three types of discrete choice models: intermediate stop generation, mode/destination, and time of day. These models first generate intermediate stops for each tour and then predict the timing and destination of stops for secondary activities, as well as the travel mode. After applying the intermediate stop level models to the synthetic population, a daily activity schedule is generated for each individual in the population. The generated activity schedules provide the timing (arrival time and departure time) of each activity at a resolution of 30 minutes, the destination at zonal level and the travel mode for each trip/tour from a list of considered modes. The output is then fed to the within-day model to generate trip chains ready for simulation, discussion of which has gone beyond the focus of the pre-day activity-based travel demand model.

5.2.2 Accessibility measures

Disaggregate utility-based accessibility measures (Ben-Akiva and Lerman, 1985) originated from random utility theory are included within the pre-day activity-based modeling framework. These measures represent the expected maximum utility (or "worth") of a set of alternatives from a choice set of a discrete choice model and are consistent with random utility theory. In a hierarchical modeling system, accessibility measures are essential to capture the sensitivity of activity and travel decisions modeled in higher levels of the modeling hierarchy to the utility of opportunities associated with conditional and undetermined outcomes from lower level models.

In formal nested modeling hierarchies, such as the one for the pre-day model, the upward integrity comes from the composite measure of expected utility across the lower level alternatives, or the so-called "logsum". For example, in a destination choice model, a logsum variable can capture the expected utility of all the available travel mode alternatives. This is a very important aspect of model integration and can be referred to as upward vertical integration. Without it, the model system will not effectively capture the sensitivity to travel conditions.

The accessibility measures, or logsums, introduced in the pre-day model are shown in Figure 5.3 with dashed arrows. The pre-day model adopts a simple accessibility measure structure where disaggregate measures from tour mode or mode/destination choice models are fed to choice models in the day pattern level.

5.2.3 Data

This section describes the data required for the development and estimation of the components that make up the pre-day activity-based demand model.

Network skims The level of service variables used in estimating the mode choice models come from network skims. Urban transportation planning agencies usually maintain travel time matrices (or skims) for a small number of pre-determined time windows, such as AM

peak, PM peak, and off-peak, generated after the calibration of volume-delay functions (based on measurements from e.g., floating car data) and the assignment of OD matrices to the network for a number of time periods. In case of the pre-day model, the network skims are provided by the Land Transport Authority (LTA) in Singapore. Those skims contain zone-to-zone travel time, travel distance, and public transit fares matrices for 3 pre-defined time periods: morning peak from 7:30 am to 9:30 am, evening peak from 17:30 pm to 19:30 pm and off-peak hours for the rest of the day.

GPS/smart card data Travel time provided in the above skims remains the same for the whole time period (AM peak, PM peak and off-peak), which makes it impossible for decision makers to evaluate the efficiency of policies that aim to shift time-of-day choice patterns (e.g., congestion pricing). Especially for the pre-day model, the time of day model is designed to capture the timing of trip/activity at a resolution of 30-minute interval. The study introduced in Chapter 3 uses GPS-enabled travel time data as well as traditional survey-based travel time data, and applies data fusion techniques to generate a more realistic travel time for every 30-minute interval in the whole day for cars. Similar techniques are adopted to generate travel time for every 30-minute interval for public transportation by using transit data collected with smart cards.

Land use data Land use data is important for the modeling of destination choice. This data provides information used to build attraction variables (Daly, 1982) included in destination choice models. Currently, the destination choices within the pre-day model are as precise as traffic analysis zones, or so-called MTZs in Singapore. A list of land use parameters is attached to each MTZ in Singapore.

Household travel survey data As mentioned in Chapter 4, household travel survey data is the essential dataset required for model development and estimation. The specific data used by the pre-day model is Singapore Household Interview and Travel Survey 2008 or HITS2008. The efforts to encode tour information from HITS2008 have been discussed with details in Chapter 4. For the estimation of each individual component in the pre-day model,

a tailored dataset will be generated based on the processed HITS2008 data.

5.2.4 Simulator design

The design of the simulation process follows the sequential approach in respect of the process flow presented in Figure 5.3. Several expected design features must be incorporated in the design of the pre-day simulator:

- The capability to simulate the daily activity schedules for millions of individuals with a reasonable computational time, which requires the ability of parallel processing and distributed computing.
- A modular design of the pre-day simulator with the ability to update individual modules (data, model, algorithm) without modifying the whole framework of the pre-day simulator.
- The ability to conduct model calibration/validation using the pre-day simulator.

Those features are incorporated in the pre-day simulator as follows (Lu et al., 2015).

Firstly, the simulator is developed mainly in C++ and the performance improvements rely on parallel processing and distributed computing through *boost threads*, *boost mpi*, respectively.

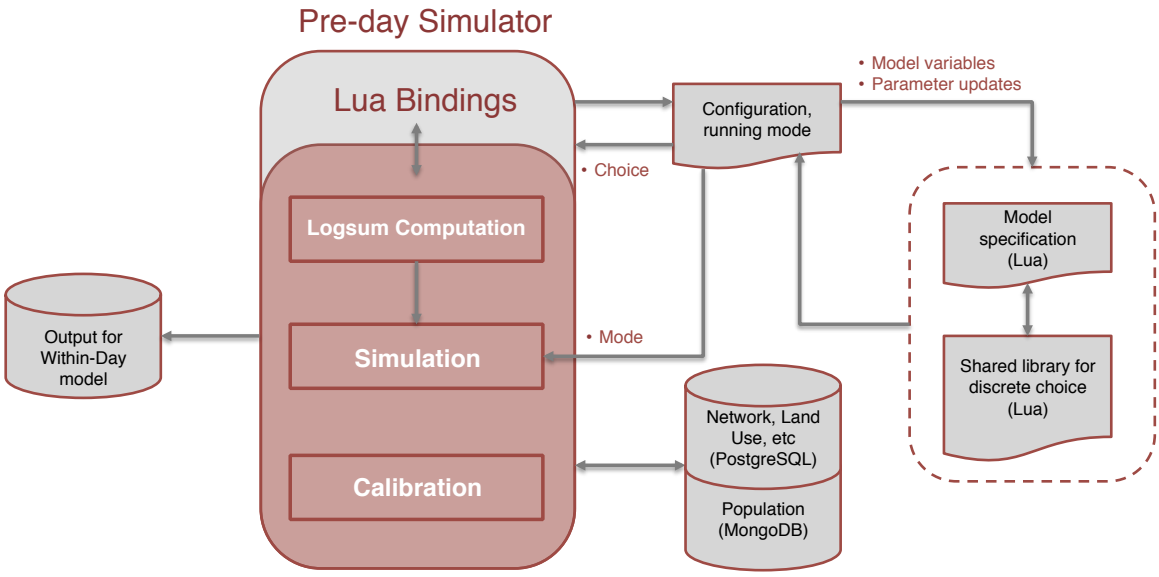


Figure 5.4: Modules of the pre-day simulator (shown as in simulation mode)

Secondly, the pre-day simulator is modularized as shown in Figure 5.4. A lightweight, embeddable scripting language Lua is used for model specification. Each of the individual models in the pre-day modeling framework has a Lua script, recording the model specification, estimated parameters, choice set, etc. Besides, a shared Lua library is created to hold all scripts related to discrete choice analysis (probability computation, simulation of choice, etc.). The Lua scripts are linked to the pre-day simulator with a model configuration file and fed to the simulator through a module called "Lua Bindings". The model configuration will determine the running mode of the simulator, which will be discussed later. The Lua Bindings mechanism makes it easier to maintain and update the simulator as a joint effort of both modelers and software engineers: On one hand, the modelers are able to test new model specifications by simply updating the relevant Lua scripts and linking the updated ones in the configuration file without touching the core of the simulator. On the other hands, software engineers are able to concentrate on optimizing the core of the simulator without messing up with the modeling details.

Thirdly, different databases are utilized to improve the performance by decreasing the I/O time. PostgreSQL database is used to store spatial-related data used in the pre-day simulator and MongoDB is used for population data and further storage of data used for model validation.

Last but not least, the pre-day simulator has three different running modes: logsum computation, simulation and calibration. The chosen mode is controlled by the configuration file. Logsum computation will provide the disaggregate accessibility measures needed for simulation and calibration by running a fraction of the simulation process. It should be noted that both logsum computation and simulation will be ran repeatedly for a calibration process, which also requires a reasonable computational time for each simulation run.

When the simulator is running in simulation mode, Figure 5.5 represents the skeleton logic of the pre-day simulator in pseudo code to generate a daily activity schedule for each individual in the synthetic population. For each individual in the synthetic population, daily activity pattern is firstly modeled, followed by a subsequent model to determine the exact number of tours for each activity purpose. The function tourSequencer is used to determine the modeling

```
1  input: simulation configurations, running mode
2  link: model coefficients (Lua scripts), population, zonal information, skims, trave time/cost by 30-minute
   interval
3  pre-calculate: disaggregate logsums
4  foreach household do
5      foreach person in household do
6          initiate: person daily activity schedule
7          predict: daily activity pattern
8          foreach purpose with 1+ tours do
9              predict: exact number of tours
10         end
           /* Also determine location for work tours (usual/unusual)                    */
11         list of tours ←tourSequencer ()
12         initiate: tour-related parameters
13         initiate: tour time windows
14         foreach tour in list of tours do
15             if location fixed then
16                 predict: tour mode
17             else
18                 predict: tour mode/destination
19             end
20             predict: tour time-of-day
21             update: tour-related parameters
22             update: tour time windows
23             if work tour then
                   /* Also determine mode/destination/time-of-day for generated sub-tours  */
24                 generate: work-based sub-tours
25             end
26             if intermediate stops available then
27                 initiate: stop-related parameters
28                 initiate: stop time windows
29                 generate: intermediate stops for first half-tour (before primary activity)
30                 generate: intermediate stops for second half-tour (after primary activity)
31                 foreach stop in first half-tour do
32                     predict: stop mode/destination
33                     predict: stop time-of-day
34                     update: stop time windows, stop-related parameters
35                 end
36                 foreach stop in second half-tour do
37                     predict: stop mode/destination
38                     predict: stop time-of-day
39                     update: stop time windows, stop-related parameters
40                 end
41                 update: tour time windows
42             end
43         end
44         generate: output, statistics for person-day, tours, and trips of each individual
45         generate: daily activity schedule for each individual
46     end
47 end
48 output: statistics for results verification
49 output: daily activity schedules to Within-day model
```

Figure 5.5: Logic flow of the pre-day simulator in pseudo code

priority (sequence) of the generated tours as well as the work location (usual/unusual work place) for each work tour. Then, each tour in the pattern is simulated in turn and each intermediate stop is simulated within each tour. Work-based sub-tours are modeled as tours but with different models to determine mode/destination/time-of-day. While it is possible to process different individuals in parallel, it is however not possible to simulate tours of the same individual simultaneously as the tour-related variables (such as number of remaining tours) and time windows will change dynamically and affect future decisions in the forthcoming tours. Thus, the sequential approach must be enforced in the simulation.

In this section, we have demonstrated the framework and overall design of the pre-day model, as well as the accessibility measures incorporated in the framework and data requirements. Based on the framework, we then move on to the design of the pre-day simulator with special consideration for several design features, such as parallelization and modularization. In the next section, development and estimation of individual models and relevant highlights will be presented.

5.3 Model Development and Estimation

This section walks through the model development for different components in the pre-day model. While HITS2008 is the essential dataset required for the model estimation in general, for each individual model, a specific dataset is isolated and dedicated (may require linking to information in other datasets, such as skims) to the estimation of a single model. Maximum Likelihood estimation is used to obtain parameter estimates of all the discrete choice models[3].

5.3.1 Day pattern level

The day pattern level defined in the pre-day modeling framework is a variation of Bowman's Day Activity Schedule approach. It predicts the number of home-based tours undertaken during the day for four purposes: work, education, shopping and others, and the availability

[3]Detailed model specification and estimation results can be accessed from https://github.com/solafishes/results.

of intermediate stops during the day for the same four purposes. We further separate the predictions by first predicting the daily activity pattern (defined as the occurrence of tours and stops as 8bit string) followed by a second model to predict the exact number of tours for each of the four tour purposes that will occur as least once during the day.

Day pattern choice model

The pattern choice is a function of various household attributes and personal socio-demographic characteristics. It predicts the occurrence of tours and intermediate stops. The occurrence is expressed as a binary variable (0 or 1+) for each of the considered purposes.

Choice set If the pattern choice model includes all combinations of 16 binary choices (8 for tours, 8 for intermediate stops, for all the 8 major activity types considered in HITS2008), there would be $2^{16} = 65,536$ alternatives, which is a large choice set for the application of discrete choice models. However, a majority of the patterns in the choice set are not realistic and not observed in the dataset. Thus, we need to generate a choice set to cover as many as day patterns observed in the dataset while preserving the size of the set to be as small as possible for the convenience of modeling and simulation.

In terms of generating the choice set, a choice set covering 8 major tour purposes is first generated to ensure a good coverage as well as a small choice set size. Then an aggregated choice set with aggregated tour purposes is generated from this one (with less activity purposes, some purposes are combined together, such as entertainment and recreation). The two stage approach can conveniently adjust to different ways of aggregating tour purposes (in the pre-day simulator, for example, 8 activity purposes are further aggregated into 4). Several sets of rules are evaluated based on the coverage of observed alternatives and the number of unobserved alternatives in the generated choice set to determine the best rule set for generating the choice set. The best set of rules is defined as follows:

1. The sum of occurrences of all tour purposes shall not exceed 3.

2. The sum of occurrences of all intermediate stop purposes shall not exceed 3.

3. The sum of rule 1 and rule 2 shall not exceed 3.
4. If there are no tours, there shall be no intermediate stops.
5. Intermediate work stops and education stops shall not appear together.
6. If there are no work tours, there shall be no intermediate work stops, which also can be applied to education tours and stops.

Following the above criteria, a choice set consisting of 457 alternatives is generated. We refer to Figure 5.6 to explain the generated choice set. In this figure, we list the 483 observed alternatives in the HITS2008 database and arrange them on the X-axis by the log value of frequency. The column for an observed alternative is depicted with dark color if it is included in the generated choice set. Otherwise, the column is depicted with bright color. In this choice set that we have generated, 246 alternatives can be observed in the dataset, which also suggests that 211 alternatives in the choice set cannot be observed. However, using this set we cover the choice of 98 percent of the respondents. What about the remaining 237 observed alternatives in HITS2008 yet not in the generated choice set? If we take a look at Figure 5.6 we will find that most of these alternatives are of low frequency (appear once or twice in HITS2008). Notice that the choice of 98 percent of the respondents has been captured, the remaining 237 patterns only have a few hundred observations.

By further aggregating the 8 tour purposes into 4 (work, education, shopping, others), we then generate a choice set consisting of 51 aggregated patterns. These patterns will be used as the choice set for predicting the day pattern with 4 tour purposes and 4 intermediate stop purposes.

Availability Every alternative is available for every individual with an exception: It is assumed that education tours are only availability for full-time students because almost all eduction tours observed in the survey are made by them with only a few contradictions.

Model structure and variables The base alternative is the pattern to stay at home. The current model structure is Nested Logit where all non-stay-at-home patterns are grouped together. Variables used in the model include:

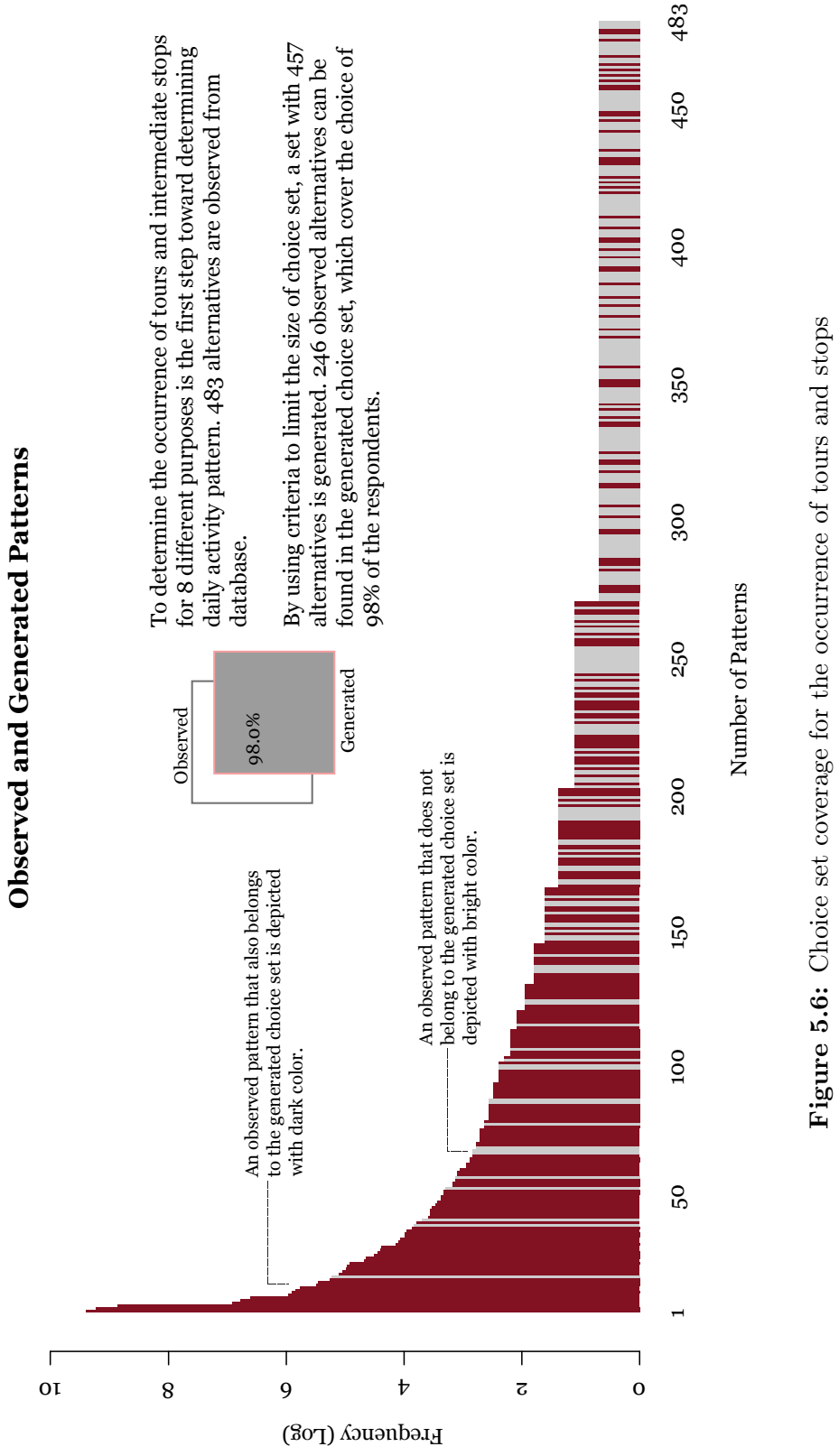


Figure 5.6: Choice set coverage for the occurrence of tours and stops

- Personal socio-demographic characteristics (age, gender, income, person type, etc);
- Household attributes (household composition, residential location type, car availability, etc);
- For education tours, accessibility using a disaggregate logsum of all modes from the tour mode choice model;
- For work, shopping and other tours, accessibility using a disaggregate logsum of all modes and locations from the tour mode/destination choice model for each purpose;
- A tour constant and a stop constant for each purpose;
- Variables related to the composition of day pattern and interactions with socio-demographic characteristics;
- A set of constants related to making tours for exactly T different purposes and stops for exactly S different purposes, for various combinations of T and S;
- Variables related to making a stop of a given purpose in combination with a tour of a given purpose.

Exact number of tours model

The model for exact number of tours predicts details that are not predicted in the day pattern model. For each of the 4 tour purposes with occurrence 1+, the model predicts the exact number of tours to be modeled subsequently. According to the survey data, most tour purposes occur only once during the day with a minor exception. Thus, the choice set for this model is quite small. For education and shopping tours, the choice set is 1 or 2 tours. For work and others, the choice set is 1 to 3 tours. While the model is dominated by alternative specific constants, socio-demographic variables, characteristics of the determined day pattern and accessibility measures are used in the model as well.

5.3.2 Tour level

At tour level, the predictions from day pattern level, as well as household characteristics and personal socio-demographic variables are available. For each work tour, a model is first applied to determine the location type of the work tour between usual (fixed) work location and unusual (non-fixed) work location. For each tour, a mode/destination choice model is then applied (some tour purposes may not need to determine destination, such as education tours). At last, for each tour, the arrival time and departure time of the primary activity are jointly predicted in the time of day model. In the simulator, the work-based sub-tour generation model is applied for each work tour (with known arrival and departure time of tour primary work activity) to generate sub-tours.

The usual work location is a variable in the data indicating whether the person has a usual work location and where it is. If a person has no usual work location, we need to jointly predict the mode and destination of any work tour made by the person. However, if a person indicates that there is a usual work location, then he/she faces a choice on whether or not to go to the usual work location for each of his/her work tours. That is the purpose of usual/unusual work location model. It is a simple binary choice model that includes variables such as person type dummies and number of work tours during the day.

Mode choice model

As indicated in Figure 5.3, for education tours and work tours heading to usual work location, only the mode is to be predicted.

Choice set For mode choice, the choice set consists of 9 alternatives. These alternatives are chosen as they have enough observations in the dataset and they are well adopted in other activity-based modeling frameworks (see for example Cambridge Systematics, Inc., 2010 and DKS Associate et al., 2012). The definitions of these alternatives are:

1. Public bus: the mode is public bus.
2. MRT/LRT: the mode is either MRT or LRT.

3. Private bus: the mode is school bus/company bus/shuttle bus.
4. Drive alone: the mode is car/van/lorry driver and only the driver is on aboard.
5. Shared 2: the mode is car/van/lorry driver or passenger and 2 persons are on board.
6. Shared 3+: the mode is car/van/lorry driver or passenger and 3+ persons are on board.
7. Motorcycle: the mode is either motorcycle rider or passenger.
8. Walk: the mode is walk.
9. Taxi: the mode is taxi.

The structure of the choice set is nested. Instead of selecting the nest structure arbitrarily, McFadden omitted variable test (see Hausman and McFadden, 1984 and McFadden, 1987) is used to test IIA in a subset of all alternatives. Results displayed in Table 5.2 suggest that it is only necessary to group Public bus, MRT/LRT and Private bus into a nest (as shown in Figure 5.7).

Table 5.2: McFadden omitted variable test on subsets of modes (work tour mode choice model)

Model	Final LL	χ^2_1	IIA
Logit	-11603.724	N/A	N/A
{4,5,6}	-11603.347	0.754	Hold
{4,5}	-11603.353	0.742	Hold
{1,2,3}	-11583.998	39.452	Does not Hold
{1,2,8}	-11602.853	1.742	Hold
{1,2,3,8}	-11594.237	18.974	Does not Hold

Availability Transit modes are available to all. Drive alone is available when the person has car/van/lorry driving license and vehicles are available for the person. The rest motorized modes are available to all. The walk mode is available only when the distance from origin to destination is less than 5 km, which is a relatively reasonable distance threshold to consider walk trips (Iacono et al., 2008; Larsen et al., 2010).

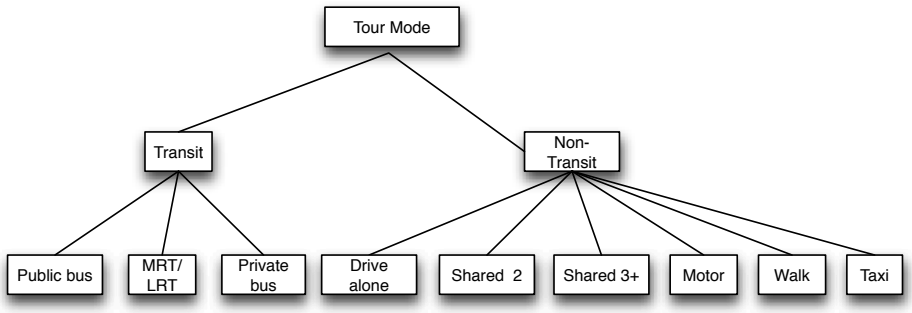


Figure 5.7: Model structure for tour mode choice

Model variables The following variables are included in the model.

- Level of service variables: These variables include in-vehicle time, out-of-vehicle time, and travel costs. Costs will include vehicle operating cost, parking cost and transit fare. For the estimation, the travel time of all modes is taken from network skims. Travel costs of transit modes and walk (zero) are taken directly from skims. For drive alone, shared ride 2, shared ride 3+ and motorcycle, the costs are calculated as the sum of operational cost, parking cost and Electronic Road Pricing (ERP) cost. The formulas for calculating each of the cost components are consistent with the formulas from the Land Transport Authority[4]. Taxi costs are derived from the taxi fare scheme in 2008. Travel costs are divided by income to reflect the fact that individuals with higher income are less sensitive to travel costs.

- Land use variables: The density of residential areas and work locations can determine the popularity of certain modes, such as private bus. Besides, for certain destinations (zones belonging to central Singapore, for example), some modes are more popular than others.

- Personal and household demographic variables, such as person type, gender, income and vehicle availability.

[4]Specifically, for drive alone, the operational cost is 0.147 * zonal distance, the parking cost and ERP cost are taken from the network skims. For shared ride 2 and 3+, the cost is divided by 2 and 3. For motorcycle, its operational cost/ERP cost is 0.5 * car operational cost/ERP cost and its parking cost is 0.65 * car parking cost.

Value of time As the mode choice model includes level of service variables such as travel time and travel costs over income, value of time can be derived from the estimation results. The derived value of time is a measure of model performance at system level. A rational value of time is necessary to ensure the correctness of the mode choice model. For example, Table 5.3 summarizes the value of time derived from the work tour mode choice model. The figures are reasonable when compared to the hourly income.

Table 5.3: Value of time by income, in S$ per hour (derived from work tour mode choice model)

Income (monthly, in S$)	1,000	4,000	8,000
Bus/MRT in-vehicle time	1.75	7.01	14.02
Bus/MRT waiting time	2.61	10.42	20.84
Bus/MRT walking time	6.40	25.59	51.18
Private bus in-vehicle time	1.64	6.57	13.14
Drive alone	3.86	15.44	30.88
Shared ride 2	2.83	11.31	22.63
Shared ride 3+	3.50	14.00	28.01
Motorcycle	4.15	16.58	33.16
Taxi	3.36	13.44	26.88

Mode and destination choice model

For shopping tours, other tours and work tours heading to unusual work location, the destination of tours is to be predicted as well.

Choice set The mode and destination choice model predicts mode and destination zone of a particular tour. The Land Transport Authority in Singapore divides Singapore into 1,092 MTZ for the purpose of transportation planning. By adopting the division, the total number of alternatives is $9 \times 1,092 = 9,828$. The model is formulated as a Nested Logit model, where the first layer predicts tour mode and the second layer predicts destination zone. The availability of each mode-destination pair is determined as follows: the availability

of the modes is the same as in the mode choice model and the availability of destination zones is conditioned on the mode. The second rule is applied only when the tour mode is walk. For other modes, all zones are available.

Model variables There are two types of variables. The first type is related to personal and household demographic variables. The second type is related to destination and the mode to the destination. Purpose-specific size (attraction) variables derived from land use characteristics are included in the model as the attractiveness of each zone.

Tour time of day model

Time of day model is used at both tour level and intermediate stop level. At tour level, it is used to predict jointly primary activity arrival and departure time for tours. To model time of day choice in the context of discrete choice analysis, time is discretized into blocks. The size of the blocks is influenced by the sensitivity of the choice model. For a demand modeling system that aims to capture the effect of time-shifting policies (such as congestion pricing), the size of the blocks should be small. However, blocks that are too small will result in too many alternatives and the marginal difference between nearby blocks will be small. In the pre-day model, the size of blocks is selected to be 30 minutes. It is noted that the way time-dependent travel time is generated from regression models (as introduced in Chapter 3) ensures that this block size can be adjusted if necessary without re-generating the travel time as input.

Choice set 48 time blocks of 30 minutes will be used as alternatives, 3:00 to 3:29 a.m., 3:30 to 3:59 a.m., ... next day 2:00 to 2:29 a.m., next day 2:30 a.m. to 2:59 a.m. Since the departure time from the primary activity should be later than the arrival time, a total of $48 \times 49/2 = 1,176$ alternatives are used. Time window, or availability of time blocks is determined after adjusting for the time periods occupied by all previously simulated tours. The model will be estimated as a Multinomial Logit model.

$$\begin{bmatrix} (300 \sim 329, 300 \sim 329) & (300 \sim 329, 330 \sim 359) & \dots & (300 \sim 329, 2630 \sim 2659) \\ & (330 \sim 359, 330 \sim 359) & \dots & \vdots \\ & & \ddots & \vdots \\ & & & (2630 \sim 2659, 2630 \sim 2659) \end{bmatrix}$$

as

$$\begin{bmatrix} 1 & 2 & \dots & 48 \\ & 49 & \dots & \vdots \\ & & \ddots & \vdots \\ & & & 1176 \end{bmatrix}$$

Model variables Variables used in the model include:

- Time-dependent alternative-specific constants for specific departure and arrival periods and combinations.
- Time window effect or time pressure (people tend to stay at the activity location for a shorter duration if they have other activities and tours to carry out during the day).
- Time-dependent level of service variables. All else being equal, people tend to travel when the destination is most accessible. Level of service variables associated with traveling, such as travel time and costs, are included.

The specification of time of day model adopts the cyclical and continuous indirect utility functions (Ben-Akiva and Abou-Zeid, 2013) to solve the discontinuity issue of utilities due to the discretization of time. The utility function for each alternative in the choice set consists of time-dependent constants and time-dependent constants interacted with covariates (such as primary activity duration). In the time of day model, those constants are specified as continuous functions of time to solve the discontinuity issue. Trigonometric functions

are used to specify these constants and there are a couple of advantages. Firstly, the trigonometric function ensures that the constant is a continuous function of time. Secondly, the trigonometric function is cyclic and there is no discontinuity at the end of day and start of next day.

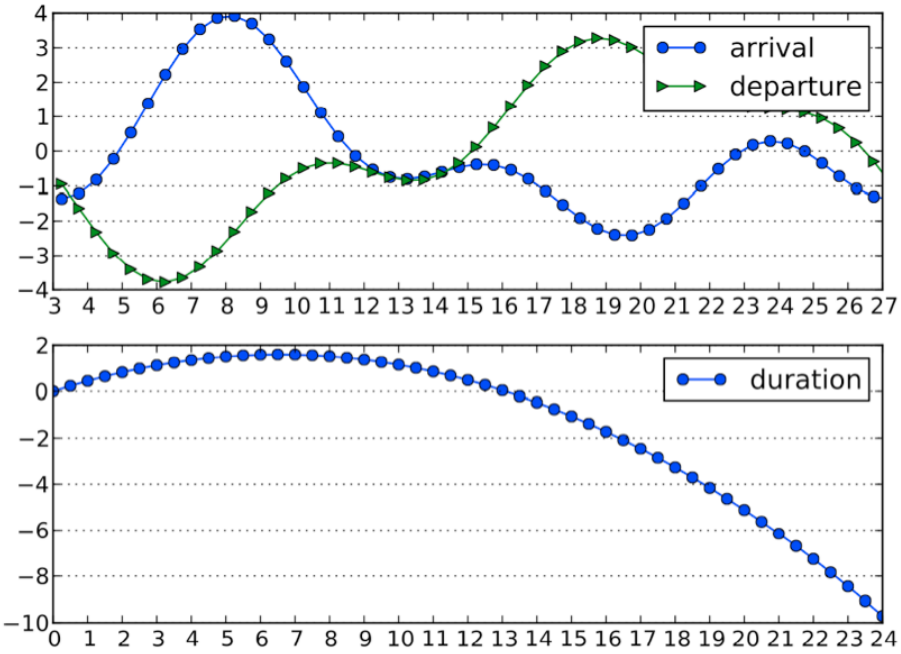


Figure 5.8: Utility from time-dependent constant and duration for work tours

Figure 5.8 shows the time-dependent constants for work tours and contribution to utility of different durations. The vertical axis represents the utils of the utility function, and the horizontal axis is the time scale. It is shown that there are diminishing marginal returns to duration for work activity. In other words, the traveler receives less utility as he/she spends much more time at work. Number of frequencies used in the trigonometric series is carefully selected such that the utility for arrival and departure time is reasonable and the model has a good fit as well. In Figure 5.8, the utility function reveals the fact that travelers prefer to arrive at work activity in the morning (8 am), and depart in the evening (18:30 pm).

Work-based sub-tour generation model

Work-based sub-tours are tours that start and end at the same work location. A stop-and-go process (similar to the one adopted in SMASH) is applied to generate work-based sub-tours. A special "Quit" alternative is added to the choice set. When the model is applied, the choice process is repeated until the quit alternative is chosen or the maximum number of sub-tours has been generated for a work tour. Based on observations from the household survey data in Singapore, the maximum number of sub-tours in a work tour is set to 2 as no observed work tours have more than 2 sub-tours. There are 5 alternatives in the choice set and the model is estimated as a Nested Logit model where all non-quit alternatives are nested together, as shown in Figure 5.9.

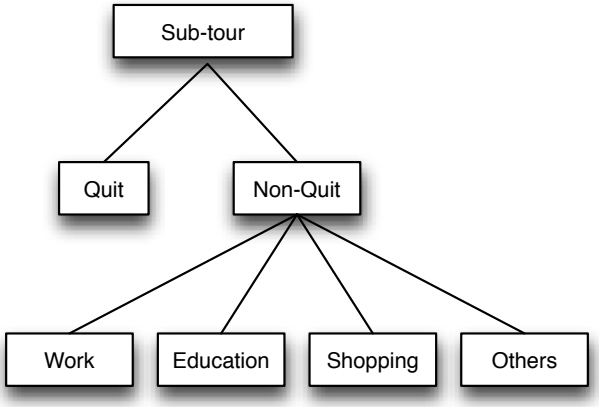


Figure 5.9: Model structure for work-based sub-tour generation

It should be noted that only a few number of sub-tours are detected in household survey data. Thus a model with only alternative-specific constants already fits the data very well (constants in non-quite alternatives are very negative and dominate the prediction). Some extra variables are added to capture the effect of multiple work tours and job categories.

5.3.3 Intermediate stop level

Models at intermediate stop level are constrained by the decisions made at day pattern level and tour level. For each half-tour, intermediate stop level will generate a sequence of stops, predict trip mode and destination, and determine the time of day for each stop.

Intermediate stop generation

This is a Nested Logit model applied to each half-tour in an individual's daily activity pattern. A half-tour is defined as the portion of a tour from home to the primary destination, or the portion from the primary destination back to home, in either case including all intermediate stops. Therefore each tour can be divided into two half-tours exactly. The availability of each of the four stop purposes has been determined in the day pattern model.

The generation of stops is another stop-and-go process. Besides four stop purposes, a "Quit" alternative is added to the choice set, indicating the termination of the process. For stops in the inbound half-tour (from the tour origin, which is home, to the destination of the tour primary activity), intermediate stops are generated backward, from the one closer to the primary activity to the one closer to home. For stops in the outbound half-tour, stops are generated forward. The rationale behind is that for a particular tour, at the time of applying intermediate stop generation model, only the tour primary activity arrival and departure time are determined, and can be used as anchor points for intermediate stop generation.

The utility for "Quit" is a function of household and personal demographic variables, as well as information about the tour/half-tour, including the distance between home and the primary destination, the tour purpose, whether it is inbound or outbound, and the tour mode. Also included in the function are characteristics of the stop itself, including the time of day, whether the stop is the first one simulated in the half-tour. The number of tours remained to be simulated and the time window available after the primary activity of the current tour has been scheduled are included as well.

Variables used in the non-quit alternatives of the intermediate stop generation model include:

- Personal and household demographic variables.
- Day pattern related variables, tour characteristics.
- Time window or time pressure.
- The occurrence of planned stops of the same purpose, both in the current tour and in other planned tours.

- Alternative-specific constants.

Stop mode and destination choice model

Mode and destination choice model at stop level is a single model that jointly predicts the mode and destination for an intermediate stop. It is very similar to the mode and destination choice model at tour level. However, there are two major differences.

Firstly, the availability of modes is constrained by the tour mode of the current tour. To be specific, the availability of modes for stops is determined based on Table 5.4, following the same priority scheme of trip modes in Chapter 4. Secondly, the level of service variables used in the model are the so-called "detoured travel time/costs". One can think of this as the extra disutility on the path from the origin to destination caused by making an intermediate stop located in a specific zone.

Stop time of day model

Time of day model at stop level is a single model, applied to stops in both half-tours, that predicts the arrival or departure time of stops. Different from the time of day model at tour level, this model only predicts one time point (departure time or arrival time, depending on the stop is in the first half-tour or the second half-tour). For stops in the first half-tour, departure time is known (departure time equals arrival time of the next stop or tour primary activity, minus travel time), only arrival time needs to be predicted. For stops in the second half-tour, arrival time is known (arrival time equals departure time of previous stop or tour primary activity, plus travel time), only departure time needs to be predicted.

The choice set consists of 48 30-minute time blocks. For stops in the first half-tour, available alternatives are bounded by time of arrival at home of the previous tour and departure time of the current stop. For stops in the second half-tour, available alternatives are bounded by arrival time of the current stop and the end of day.

Table 5.4: Availability for mode alternatives at intermediate stop level

	Tour mode						
Stop mode	Private bus	Public bus, MRT/LRT	Shared 2, 3+	Drive alone	Taxi	Motorcycle	Walk
Private bus	✓						
Public bus, MRT/LRT	✓	✓					
Shared 2, 3+	✓	✓	✓				
Drive alone	✓	✓	✓	✓			
Taxi	✓	✓	✓	✓	✓		
Motorcycle	✓	✓	✓	✓	✓	✓	
Walk	✓	✓	✓	✓	✓	✓	✓

5.4 Model Calibration/Validation

5.4.1 Overview

After all the models in the pre-day modeling framework are estimated, the model specifications along with the estimation results are updated in corresponding Lua scripts. The demand simulator is now able to generate for each individual in the population a daily plan of activity and travel. However, outputs from the pre-day simulator are not to be fully trusted until rigorous efforts are devoted to validating the results.

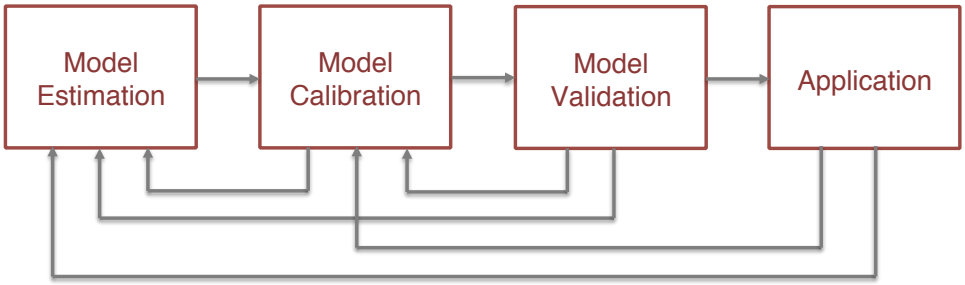


Figure 5.10: General process for the development of travel demand models

Figure 5.10 represents the general process for the development of travel demand models. At the lowest level, the statistical goodness of fit should be well observed during the model estimation process where statistical analysis techniques and observed data are used to develop model parameters or coefficients. Once the models are estimated, small adjustments are made until the overall modeling framework is properly interfaced and that any modeling error is not propagated by chaining the models together such that it can accurately replicate the observed patterns and behavior, which is achieved with model calibration and model validation occurring in an iterative fashion.

Model calibration and validation occurs at three levels. At the first level, there is the need to verify that the chaining of models is capable of replicating the base year travel behavior with good accuracy and precision. The same dataset for estimation is used for calibration and validation. At this level, model validation may reveal the need to return to model estimation or model calibration due to the modeling errors and the propagation of them. At the second level, there is the higher expectation of prediction, which indicates the extent to which the model can be applied to predict travel behavior and assess policy scenarios in a forecasting

setting. A different dataset, usually from future travel surveys and policy scenarios is used for calibration and validation. However, in practice, as mentioned in Roorda et al. (2008), when subsequent data becomes available, forecasting validation is rarely undertaken and the emphasis is usually placed on model re-development and re-estimation. At the third level, there is the need to seek transferability in another region. The model applied in one region can be transferred to another region if it is validated thoroughly. In practice, the case for transferring is even rarer as every region has the need to customize its travel demand model to best suit the local contexts, which makes it more reasonable to start the development process from the beginning.

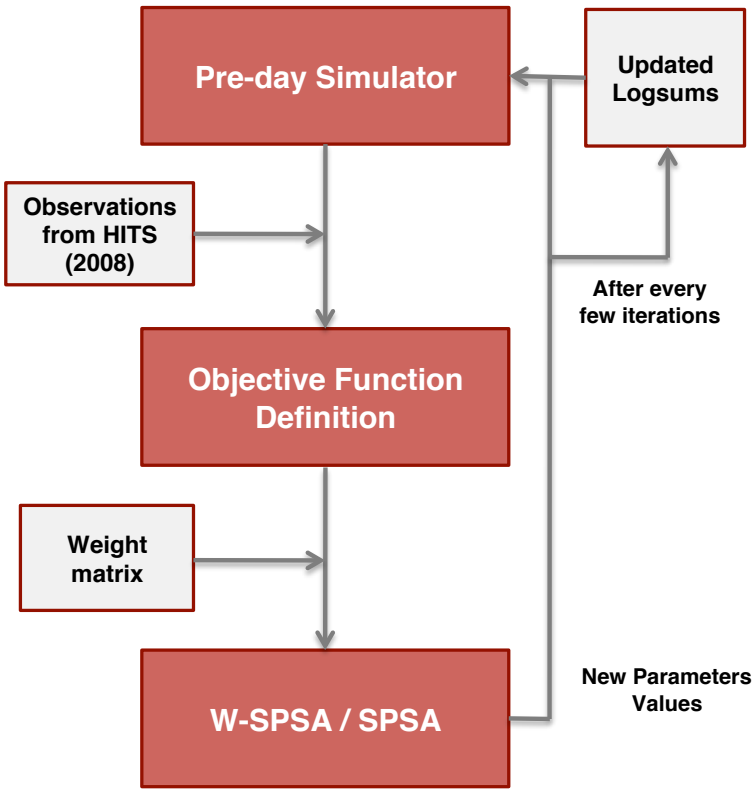


Figure 5.11: Calibration/validation of the pre-day simulator

For the case of the pre-day model, the data from year 2012 is incomplete at the moment for a forecasting-based validation and we are going to only cover the calibration/validation (verification) against the base year observations. Moreover, the pre-day simulator is the first full-fledged module in SimMobilityMT with full functionality. As a result, the aggregate measures of travel used in the calibration/validation will not include the ones related to the supply simulator, such as traffic counts.

Figure 5.11 outlines the calibration/validation process of the pre-day simulator, which also appears in Figure 5.4 when the simulator runs in calibration mode (logsums are only updated every few iterations to speed up the process). Calibration of travel demand models typically involves the refinement and adjustment of model components and parameters. In case of the pre-day simulator, 77 alternative specific constants and 17 scale parameters are used as the optimization variables. To ensure that the simulated demand replicates the base-year (year 2008) travel statistics, an OLS-type objective function is created to measure the difference between simulated and observed statistics. The normalized squared differences between 39 city level observed aggregate statistics (including the distribution for number of tours made by each individual, the distribution for number of intermediate stops in tours, tour mode shares, stop mode shares, distribution for trip distance, number of sub-tours and sub-tour mode shares) and simulated statistics are incorporated in the objective function.

Although the objective function mentioned above is in closed-form, the values taken to calculate the objective come from a simulator, making it impossible to carry out the optimization using direct gradient measurements. One optimization method that has the feature of gradient-free is the Simultaneous Perturbation Stochastic Approximation (SPSA) method, which only relies on objective function measurements. Further, SPSA is especially efficient in high-dimensional problems in terms of providing a good solution for a relatively small number of measurements of the objective function, which is crucial for the calibration of the pre-day simulator since every iteration requires considerable time. Spall (1998) and Spall (2001) introduced the SPSA method and implementation details. SPSA has been used in the calibration of DTA packages, such as DynaMIT (Balakrishna, 2006) and CORSIM (Paz et al., 2012). On top of the basic method, a weighted version of SPSA algorithm (W-SPSA, see Lu, 2013) is implemented for the calibration of the pre-day model. Essentially, the weight matrix in W-SPSA describes the relevances of the measurements to the optimization parameters and W-SPSA is a way to speed up the convergence[5]. The final convergence is determined based on very small toleration value of the change in the objective function in successive

[5]Suppose the objective function can be divided into several parts where for each part, only a distinct group of optimization parameters can influence the objective. Then the problem can be divided in W-SPSA with a partitioned matrix and solved individually to speed up the process. This is an extreme case of how W-SPSA can be useful in the calibration process.

iterations. The next section provides a summary of the calibration/validation against the base-year travel statistics.

5.4.2 Summary of validation results

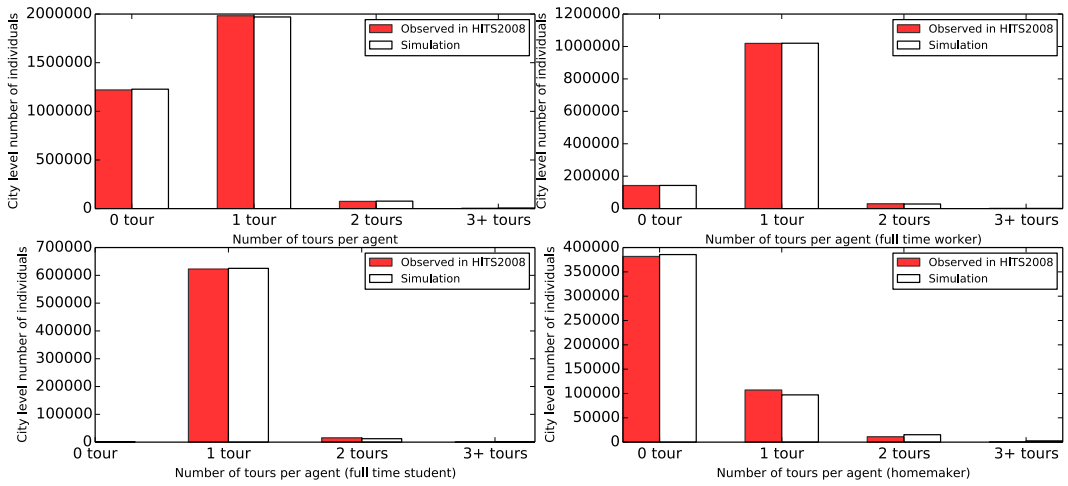


Figure 5.12: Base year validation: Number of tours per individual

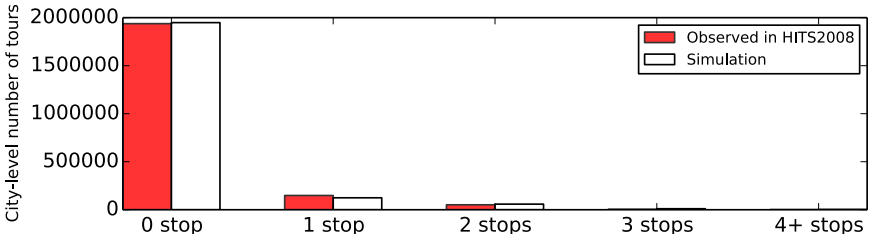


Figure 5.13: Base year validation: Number of intermediate stops per tour

Tour/stop frequency In total, 2,144,037 tours and 10,087 work-based sub-tours are generated in the simulator for the sampled HITS2008 population used for developing the pre-day model, where 2,143,677 tours and 9,993 work-based sub-tours are observed[6]. Generally, the simulator generates precisely the number of tours observed in the base year. Figure 5.12 presents the distribution of number of tours for selected person types, including all the individuals, full-time workers, full-time students and homemakers. It is shown that

[6]While the simulation runs with the sampled HITS2008 population, the number of tours and sub-tours reported here has been expanded based on the household expansion factor for each household in the sampled HITS2008 population used for estimation.

the simulator is able to replicate the distribution precisely. Noted that the distribution is the result of both daily activity pattern model and exact number of tours model, it indicates that the models in the day pattern level of the pre-day modeling framework are functioning properly, which is critical as any discrepancy in this level will be propagated to lower level models. Figure 5.13 shows number of intermediate stops contained in each tour. The simulator is able to capture the fact that the great majority of tours are without intermediate stops.

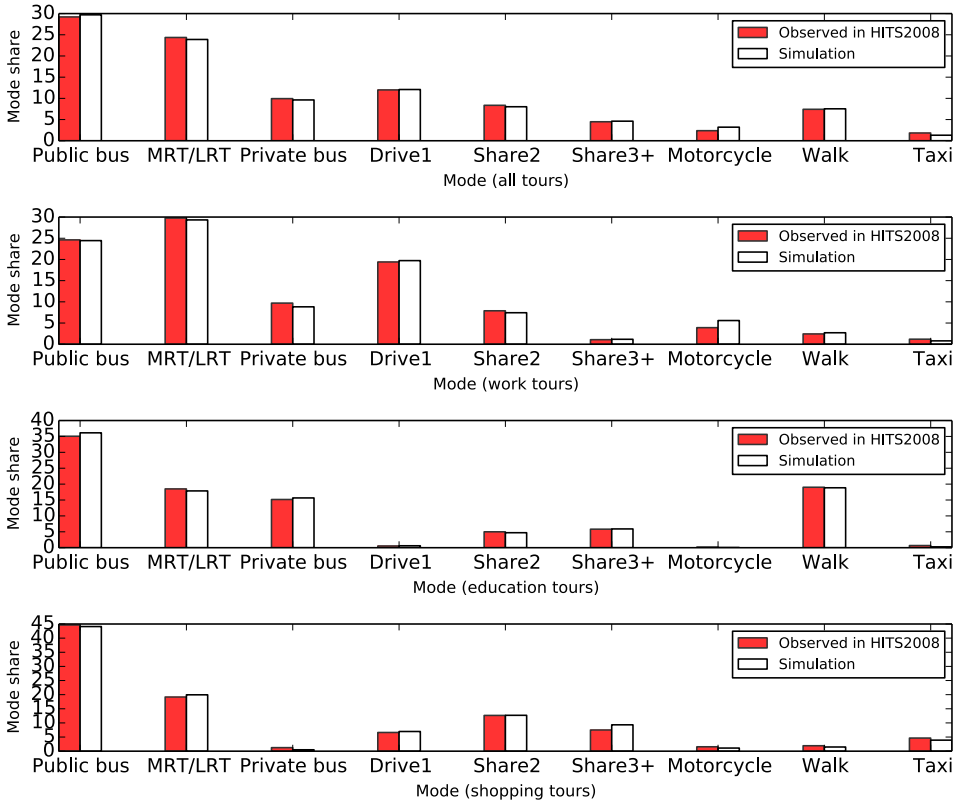


Figure 5.14: Base year validation: Tour mode choice

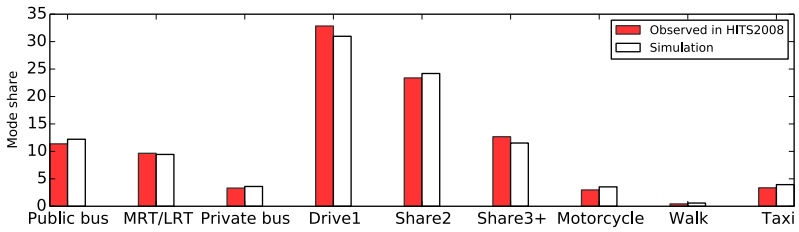


Figure 5.15: Base year validation: Intermediate stop mode choice

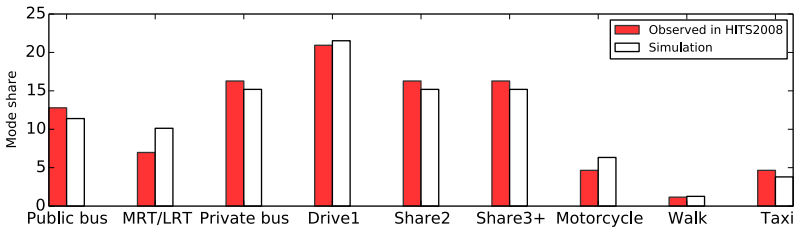


Figure 5.16: Base year validation: Work-based sub-tour mode choice

Tour/stop mode choice Figure 5.14 to Figure 5.16 compare the observed and simulated mode shares for tours, intermediate stops and work-based sub-tours. For all tours, the largest discrepancy occurs at Motorcycle where the mode share is overestimated by 0.83 percent. For work tours, the mode share for Motorcycle is overestimated by 1.68 percent. The largest discrepancy for education tours occurs at Public bus, where the mode share is overestimated by 1.07 percent. For shopping tours, the mode share for Shared ride 3+ is overestimated by 1.81 percent. Generally speaking, the simulated mode shares resemble the observed mode shares for all tours, work tours, education tours and shopping tours.

Some discrepancies are observed in terms of the mode shares for intermediate stops and work-based sub-tours. For intermediate stops, the difference between observed and simulated mode shares is greater than 1 percent for Public bus, Drive alone, Shared ride 2, Shared ride 3+, Motorcycle and Taxi. And the largest discrepancy occurs at Drive alone: 3.75 percent. For sub-tours, the difference between observed and simulated mode shares is greater than 1 percent for Public bus, MRT/LRT, Drive alone, Shared ride 3+ and Motorcycle. And the largest discrepancy occurs at MRT/LRT: 3.14 percent. Mode choice at intermediate stop level is conditioned on the tour level decisions and can be sensitive to the discrepancies at upper levels. For work-based sub-tours, since the validation is based on the sampled HITS2008 population used for estimation where less than 100 work-based sub-tours are included, the mode shares can be very sensitive to the actual number of generated sub-tours in the simulator.

Tour/stop time of day For the timing of a tour, arrival time at the primary activity and departure time from the primary activity are jointly predicted. The resolution for the

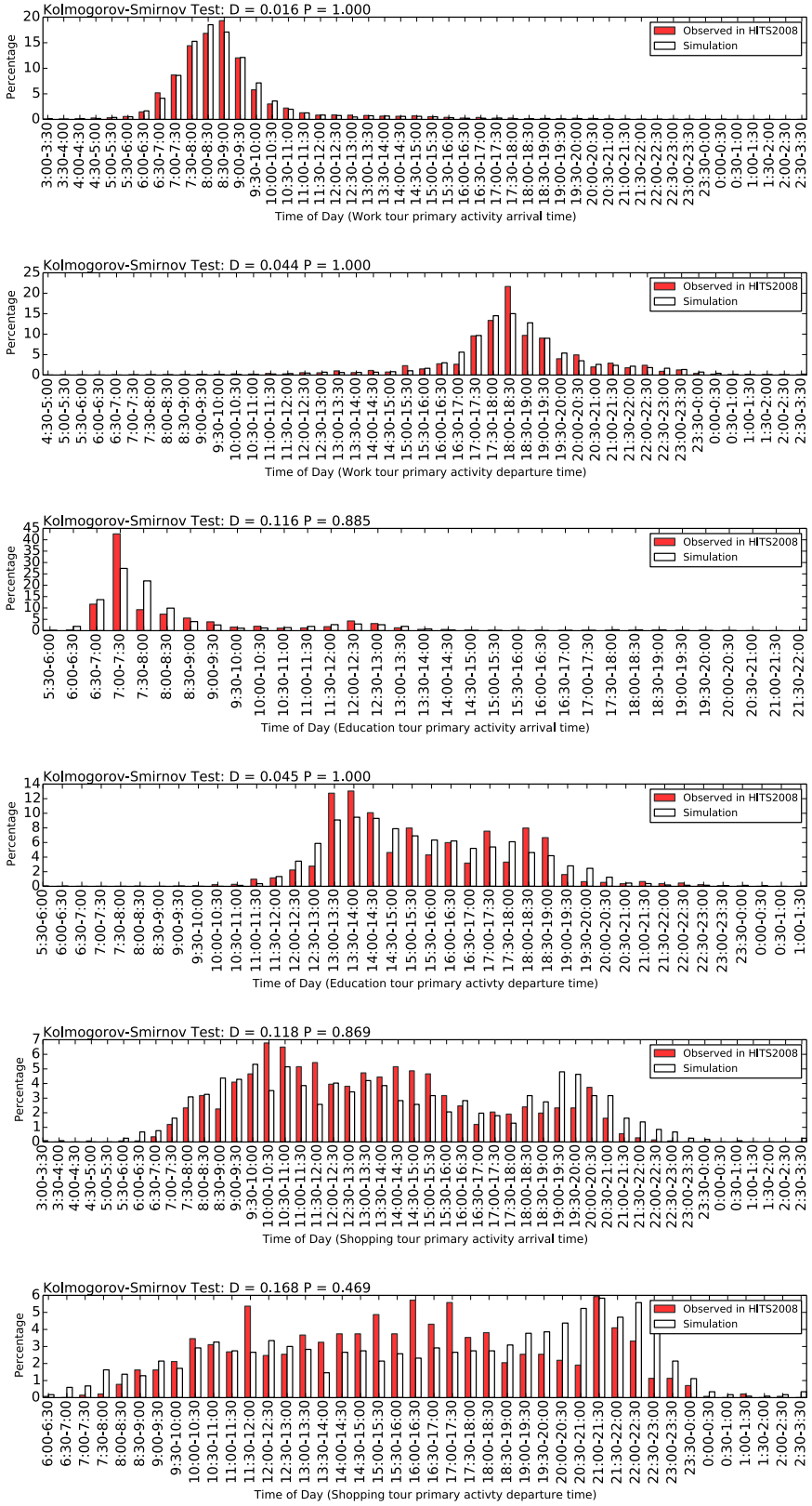


Figure 5.17: Base year validation: Tour time of day

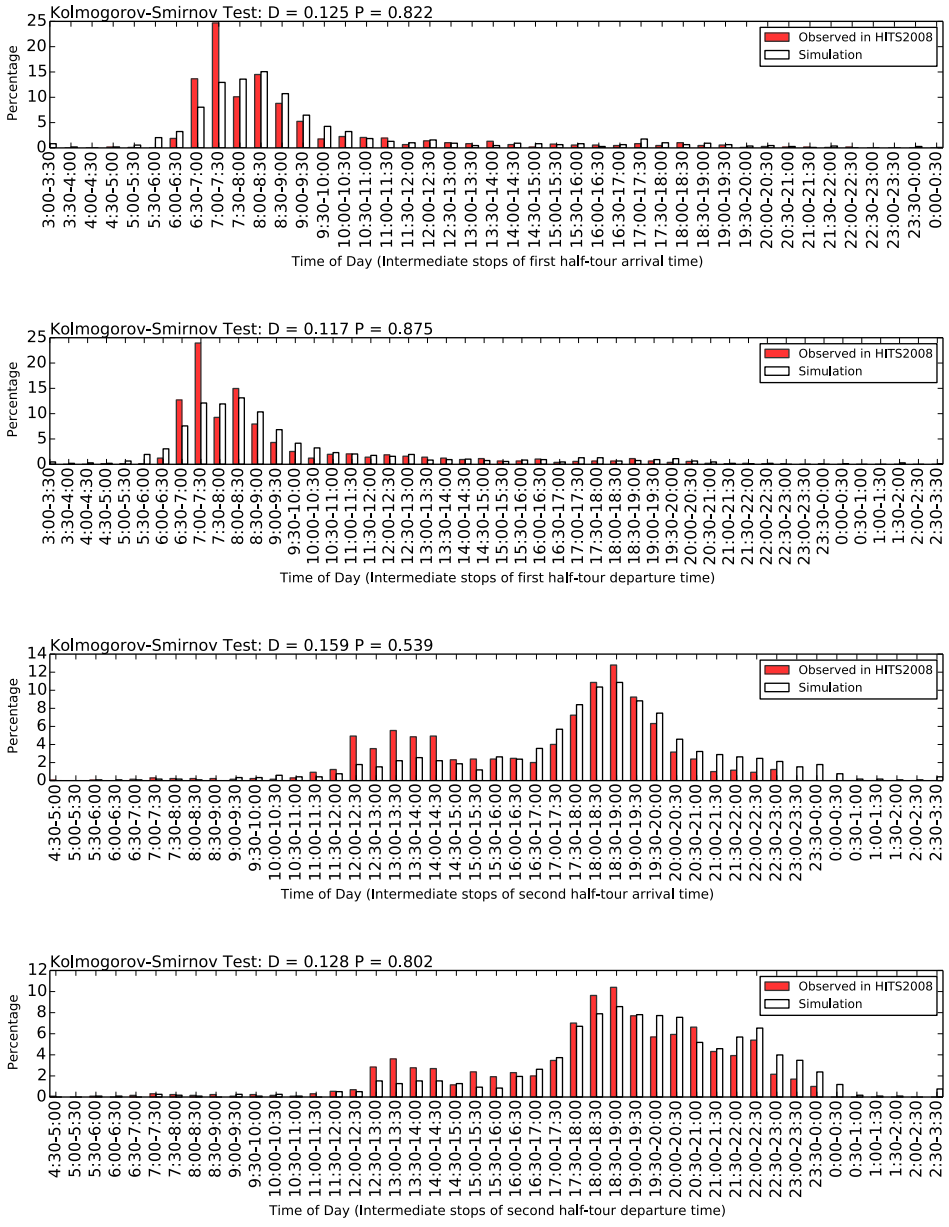


Figure 5.18: Base year validation: Intermediate stop time of day

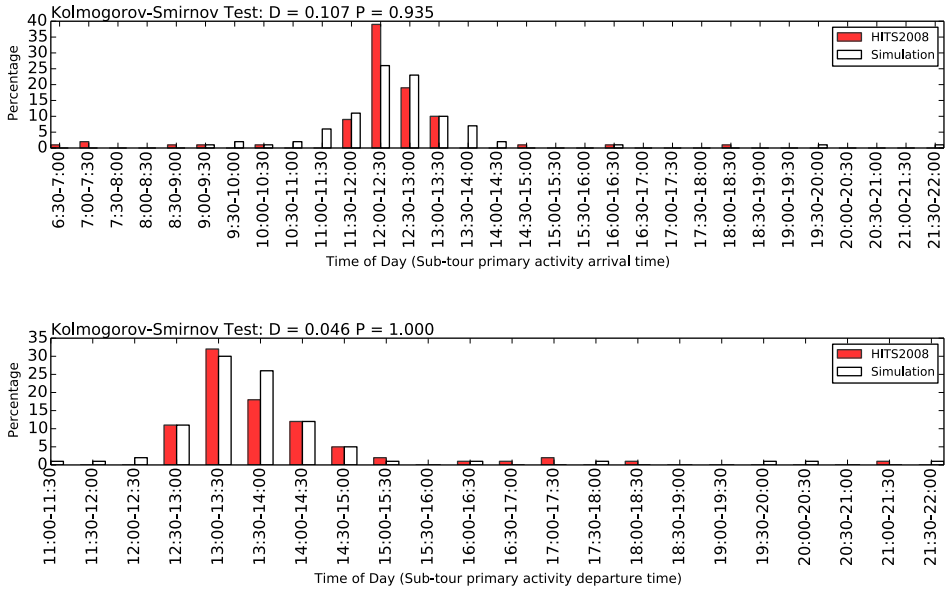


Figure 5.19: Base year validation: Work-based sub-tour time of day

alternatives in the time of day choice model is by 30-minute interval. Figure 5.17 presents six histograms illustrating the distributions of primary activity start time, and primary activity end time across the observed demand (HITS2008) and the simulated demand for three specific types of tours: work, education and shopping. The time of day model successfully replicates the distributions of work tour arrival/departure time. For education tours, some of the peaks are missing in the prediction. However, the model does capture the fact that some education tours start after noon, which is unique for middle schools in Singapore. Furthermore, Kolmogorov-Smirnov test is applied to these distributions, and the P values indicate that the simulated and observed distributions are likely to be the same with high probability.

Tour distance Tour distance is measured as the distance between home and the location of tour primary activity. It partially reflects the soundness of the destination choice model. In the pre-day model, tour destination is predicted for work tours going to unusual work places, shopping tours and other tours. As shown in Figure 5.20, the simulator is able to precisely replicate the decreasing pattern in preference of longer tours observed in HITS2008.

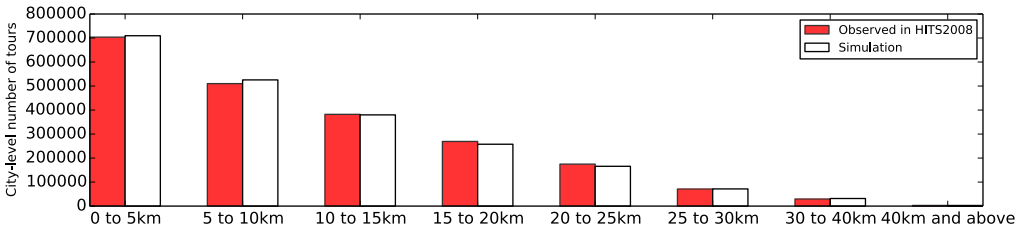


Figure 5.20: Base year validation: Tour distance

5.5 Concluding Remarks

The study in this chapter represents the efforts in developing an activity-based travel demand modeling system for Singapore, and an integrated mid-term simulator SimMobilityMT for the modeling of activity participation and travel decisions occurring on daily basis. The pre-day simulator introduced in this chapter is considered as a benchmark and continuing efforts are being devoted to enhance it. In this section, we would like to provide a summary to the benchmark and a brief discussion on the future works.

5.5.1 Summary of the benchmark

The SimMobilityMT pre-day modeling framework follows Bowman's Day Activity Schedule approach. It is a system of interconnected discrete choice models representing choices at distinct dimensions or facets. This study presents the concept, design and implementation of the pre-day model and its simulator. There are three different hierarchies in the pre-day modeling system: day pattern level, tour level and intermediate stop level. Each level consists of several models. The overall system can be viewed as a hierarchical (or nested) series of choice models. Besides the unique modeling framework, several innovations are observed in the development of the pre-day model as well. Firstly, instead of aggregate logsums, disaggregate accessibility measures (or disaggregate logsums) originated from random utility theory are included in the system. As a result, the sensitivity to potential benefits and costs from lower level models will be reflected at individual level. Secondly, the time of day models in the pre-day framework adopt the cyclical and continuous indirect utility functions, which ensure that the utility function is a continuous function of time and there will be no

discontinuity at any time of day, including the end of day and start of the next day. Thirdly, multiple data sources, such as traditional household travel survey, smart card data and taxi GPS data are fused to generate the time-dependent travel time used in the model estimation.

In terms of the implementation of the pre-day simulator based on the model development efforts, there are several highlights as well. First of all, the pre-day simulator is used to simulate the daily activity schedules for millions of individuals with a reasonable computational time. To address the computational time issue, parallel processing, distributed computing as well as methods for minimizing I/O time are implemented. Besides, the simulator enables the collaboration of software engineers and demand modelers via the carefully designed modularization with "Lua bindings". The separation of software implementation and travel demand models via modularization has shown in practice to be more efficient and less error-prone. Finally, the simulator is designed to be running in several modes, such as logsum computation, simulation and calibration. The chosen mode is controlled by the configuration file. Model calibration is carried out through Simultaneous Perturbation Stochastic Approximation (SPSA), which is usually adopted in the calibration of Dynamic Traffic Assignment (DTA) models.

As a benchmark, the pre-day model has been estimated with HITS2008 and other data sources previously mentioned, the pre-day simulator has been fully implemented using C++, and the whole pre-day modeling system is calibrated and validated to correctly replicate activity participation and travel behavior of the base year. Improvements, of course, can still be made in subsequent phases of the on-going development of the pre-day model and its simulator.

5.5.2 Future development

The pre-day model can be further enhanced in many aspects.

- To validate the pre-day model for the purpose of prediction, HITS2012 and other data sources are being collected and processed.

- Moreover, once the SimMobilityMT within-day module and supply simulator are on

line, the validation can be carried out to compare the actual traffic counts and simulated traffic counts using SimMobilityMT.

- The application of the pre-day model and SimMobilityMT in planning scenarios such as congestion pricing, discount fare for transit, etc.

- Integration with long-term decisions: Accessibility measurements can be generated from the pre-day model and fed into SimMobilityLT framework for the modeling of long term decisions such as household location, fixed work location and vehicle ownership.

Chapter 6

Learning Daily Activity Patterns with Probabilistic Grammars

Daily activity pattern is the reflection and abstraction of actual individual activity participation on daily basis. It carries information on activity type, frequency and sequence. Preference of daily activity patterns varies among population, and thus can be interpreted as personal life styles. This chapter intends to advance studies on human daily activity patterns by providing new perspective and methodology in the modeling and learning of daily activity patterns using probabilistic context-free grammars. In this chapter, similarities between daily activity pattern — which is defined as activity sequence — and language are explored. We developed context-free grammars to parse and generate daily activity patterns. To replicate people's heterogeneity in selecting daily activity patterns, we introduced probabilistic context-free grammars and proposed several formulations to estimate the probability of a context-free grammar with daily activity patterns observed in household travel survey. We conducted experiments on the proposed formulations, finding that under proper context-free grammar and problem formulation, the estimated probabilistic context-free grammar is able to reproduce the observed pattern distribution in household travel survey with satisfactory precision. Practically, the proposed methodology sheds light on the issue of generating customized choice sets for daily activity pattern models in certain activity-based modeling frameworks.

6.1 Introduction

Daily activity pattern is the reflection of each individual's actual activity participation on daily basis. It enables researchers to reveal life styles of population in the interested region (Jiang et al., 2012). As the demand for traveling is rooted in the need for activity participation, daily activity pattern is crucial in travel demand modeling as well.

Daily activity patterns can be extracted from various data sources and used for studies covering a wide range of topics. In Axhausen et al. (2002), researchers studied daily activity patterns extracted from a six-week diary-based travel survey and analyzed the behavioral rhythms of activity participation. Joh et al. (2002) defined daily activity patterns as activity sequences and proposed a multidimensional sequence alignment method to measure the similarities between patterns. Recently, advance in information and communication technology (ICT) has brought various locating techniques — such as GPS — into travel surveys, enabling researchers to take advantage of new travel survey methods using GPS sensors and smartphones. In Allahviranloo and Recker (2013), for example, GPS travel data enabled the application of Support Vector Machine (SVM) in learning the daily traveler activity engagement patterns. Studies covering extracting and modeling of activity/travel patterns from such data include González et al. (2008), Zheng et al. (2009) and Phithakkitnukoon et al. (2010). In another intriguing study, Song et al. (2010) discussed the limits of predictability of human mobility patterns when trajectory data comes from mobile phones.

In the context of activity-based modeling, the modeling of daily activity patterns is the key to distinguish activity-based demand models from tour-based demand models (see Ben-Akiva and Bowman, 1998 and Bowman, 2009). As the first working prototype of a full-day activity-based demand model system proposed in 1995, Bowman and Ben-Akiva (2001) modeled choice of daily activity pattern at the highest level and a choice set for daily activity patterns is pre-specified. Tour and trip decisions are conditioned, or constrained, by the choice of activity pattern and are modeled at second and third level respectively. A similar three-level representation of model structure was developed in Bhat et al. (2004), where daily activity pattern is modeled differently for workers and non-workers. The definition of daily activity pattern varies in different empirical activity-based implementations. Given the concepts

of home-based tour and intermediate stop, which are commonly used in travel demand modeling, a daily activity pattern can be defined as the occurrence of tours and intermediate stops according to diverse activity purposes (see for example, Cambridge Systematics, Inc., 2010 and DKS Associate et al., 2012). The pre-day model introduced in Chapter 5 adopts this definition. Alternatively, some empirical implementations define it as the sequence of participated activities in chronicle order (see for example, Bradley et al., 1998 and Parsons Brinckerhoff, 2010). In principle, daily activity patterns in activity-based models should provide information on type, frequency and sequence of participated activities.

In this study, we adopt activity sequence as the definition of daily activity pattern. Inspired by an analogy between activity sequence and language (or set of synthetic strings), we propose an innovative approach to the modeling and learning of daily activity patterns. By treating activity sequences as a context-free language, the proposed approach is capable of replicating people's heterogeneity in selection of daily activity patterns by learning a probabilistic context-free language from observed activity patterns.

To get firsthand experience, we extracted activity sequences from Household Interview and Travel Survey of Singapore conducted in 2008 (HITS2008). As introduced previously, HITS is a diary-based survey conducted every four years in Singapore. A sample of 1 percent of Singapore population participated in the survey and reported their socio-demographic characteristics, trip-making and activity-participating behaviors in a typical weekday. An activity sequence starts and ends with symbol `h` (home activity) and may involve arbitrary number of `h` in between. Between two adjacent `h`s, there is at least one letter indicating the sequence and purpose of activity. As an illustration of this definition of daily activity patterns, Figure 6.1 presents 15 most frequent daily activity patterns in HITS2008. As shown in the figure, people prefer short patterns and their choices are concentrated on several major patterns. Although more than 80 percent of the surveyed individuals were characterized by three frequent daily activity patterns in HITS: `hwh` (a simple work tour), `h` (stay at home) and `heh` (a simple education tour), it is still surprising that we observed more than 1000 distinct patterns in total. Thus, the question is raised: how to replicate the probability of selecting an individual pattern when modelers try to sample from all the patterns.

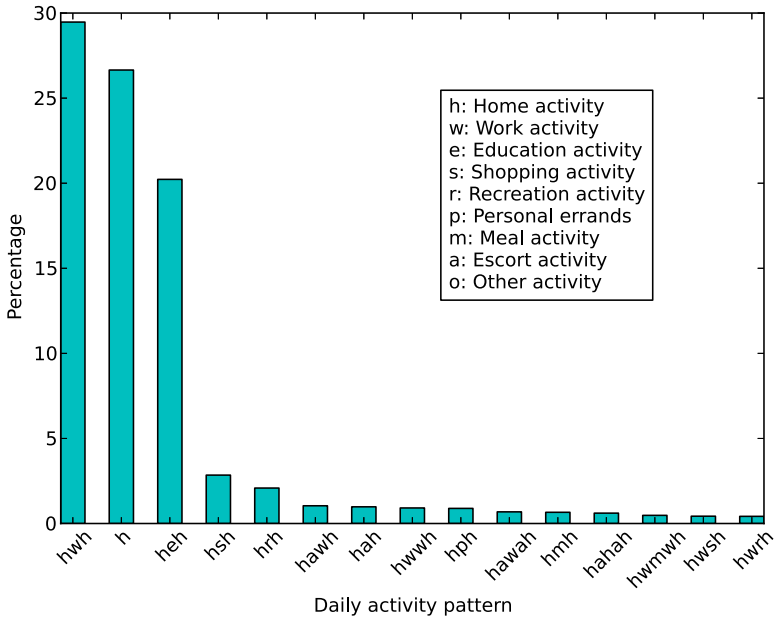


Figure 6.1: Percentage of 15 most frequent daily activity patterns in HITS2008

In this study, we tackle the problem from a new perspective, starting with investigating similarities between daily activity patterns and natural languages. Firstly, both daily activity pattern and language are formulated with a finite set of symbols, or building elements. The building elements are the letters representing activity types in daily activity patterns (e.g., `h` for a home activity, `w` for a work activity), whereas they are words in natural languages such as English. Secondly, both daily activity pattern and language have rules to determine whether a string formulated by a sequence of building elements is in their corresponding sets. For instance, English grammar helps identify whether a sentence is grammatical. Likewise, the definition of daily activity pattern helps identify whether a string consisting of letters of activity types is a well-formed pattern or not. Thirdly, both grammatical sentences and well-formed patterns are characterized by deterministic distributions. For instance, in English, some sentences are more frequently used by speakers. Similarly, different patterns have different occurrence probabilities as implied in Figure 6.1.

Inspired by the similarities between daily activity patterns and natural languages, we further examine that the methodology used to construct and understand languages may be applied

to learn daily activity patterns as well. In early $1950th$, interests in machine translation brought up the idea of text analysis with production rules that identify various components of a sentence and separate them for analysis. In linguistics community, the formal grammar proposed by Noam Chomsky (see Chomsky, 1956; Chomsky, 1957 and Chomsky, 1959) specifies explicit rules for combining elements into sentences and makes it possible to study grammars with mathematics.

The remainder of the chapter is organized as follows. Section 6.2 introduces one particular grammar in formal grammar family—the context-free grammar. We build a simple context-free grammar for daily activity patterns and extend it to a probabilistic context-free grammar (PCFG). In Section 6.3, we propose three problem formulations to estimate a probabilistic context-free grammar with observed daily activity patterns in household travel survey. In Section 6.4, several experiments are conducted to test different context-free grammars and the proposed problem formulations. Finally, application of this methodology in activity-based modeling, limitations and promising improvements are discussed in Section 6.5.

6.2 Methodology

The formal grammar proposed by Chomsky was the first attempt to give a precise characterization of the structure of natural languages. A grammar is a declarative specification of well-formedness, which determines whether a sentence belongs to the language defined by the grammar. A grammar is usually taken as a language generator, but it is also the basis of a recognizer that determines for any given string whether it is grammatical or not. In this section, basic knowledge of formal grammar is introduced first before any discussion on applying it to recognize the structure of daily activity patterns. Following the general definition, we build a simple context-free grammar for daily activity patterns and then introduce the concept of probabilistic grammars.

6.2.1 Basics of formal grammars

Define a formal grammar as $G = (T, N, S, R)$, where

T is a finite set of terminal symbols, which are the building elements of the language (for example, words in English, activity types in daily activity pattern).

N is a finite set of nonterminal symbols which are disjoint from T.

$S \in N$ is the start symbol.

R is a finite subset of $N^+ \times (N \cup T)^*$ (for a set of symbols I, I^* is the set of all strings of symbols in I including the empty string and I^+ is the set of all strings of symbols in I excluding the empty string).

We call R (rewriting) rules. $\alpha \to \beta \in R$, where $\alpha \in N^+$, $\beta \in (N \cup T)^*$ states that α, the left head, can be replaced by β, the right head. $\gamma\alpha\delta$ can be replaced with $\gamma\beta\delta$ if and only if $\alpha \to \beta \in R$ and $\gamma, \delta \in (N \cup T)^*$. Nonterminal symbols are those symbols that can be replaced by applying rules in R.

With grammar G defined, we then define $L(G) = \{w \in T^* | S \Rightarrow^* w\}$ as the language derived, or generated by grammar G. $S \Rightarrow^* w$ is interpreted as that $\forall$ string $w \in L(G)$ can be derived ($\Rightarrow^*$) from the start nonterminal symbol S. As an example, consider the case where G is English grammar. L(G) simply represents the set of all sentences that follow English syntax. Strings in the language $L(G)$ are derived by starting with the start symbol S and repeatedly applying rules in R, replacing nonterminal(s) on the left side of $\to$ with symbol(s) on the right side of $\to$, until no nonterminals are left to be replaced.

Chomsky (1956) defined the hierarchy of formal grammars. Any grammar G defined above is of type-0. For $\forall \alpha \to \beta \in R$

G is a context-sensitive grammar (type-1) if $|\alpha| \leq |\beta|$.

G is a context-free grammar (type-2) if $|\alpha| = 1$.

G is a right-linear grammar (type-3) if $|\alpha| = 1$ and $\beta \in T^*(N \cup \varepsilon)$, where ε is null nonterminal symbol.

Any grammar of type-3 is equivalent to corresponding regular expression or finite automata (see Kleene, 1956; Chomsky and Miller, 1958 and McNaughton and Yamada, 1960 for proof), which is easy to parse but limited in expression power. Although context-sensitive grammars

are powerful enough, no polynomial algorithm is found for parsing (Hopcroft et al., 2006). As a compromise, context-free grammars constrain the length of symbols to be replaced (left head) to be 1. So far it is the most well studied grammar and polynomial algorithms for parsing are available. Besides, context-free grammars are widely used in computer science and learning the structure of DNA/RNA sequences (Dyrka and Nebel, 2009).

Both regular and context-free grammars introduced above may be applied to describe daily activity patterns. Since the set of regular languages is a subset of context-free languages (Chomsky and Miller, 1958), regular languages may not be sufficient for the purpose of describing daily activity patterns. Consider the case of escort activities (`o` in Figure 6.1). The escort activities can be further divided into drop-off and pick-up activities. The modeler may observe that the majority of patterns with escort activities will have a balanced number of drop-off and pick-up activities, which is usually the case for domestic workers accompanying schooling children to school/home. We may observe patterns such as `h()h`, `h((s))h`, where "`(`" stands for a drop-off activity and "`)`" stands for a pick-up activity. Balanced parentheses of arbitrary length is beyond the ability of any finite automata but can be described with context-free grammars (Kozen, 1997). Given the fact that it is quite possible for modelers to come up with grammars that are context-free yet not regular, as the above example has shown, it is reasonable to consider describing daily activity patterns with context-free grammars. The problem formulations that we are about to propose in the next section will work for context-free grammars. And since any regular grammar is also context-free, the formulations work too if the grammar used to describe daily activity patterns is regular.

To have an intuitive understanding of formal grammars, especially context-free grammars, we provide a toy example in Figure 6.2. For convenience, rules with the same left head are written in one line, and different right heads are separated with '|'. The set of terminal symbols T contains limited number of words in English. In the N, each of the symbol represents a class of words or phrases. The rule set R provides rules for how to compose a grammatical English sentence. In this example, the grammar defines a language that contains a very limited subset of English. Only sentences that have noun phrase (NP) in front of verb phrase (VP) are in the language. VP can be replaced with a verb (V) or V followed by a prepositional phrase (PP). The location (L) is either canteen or classroom.

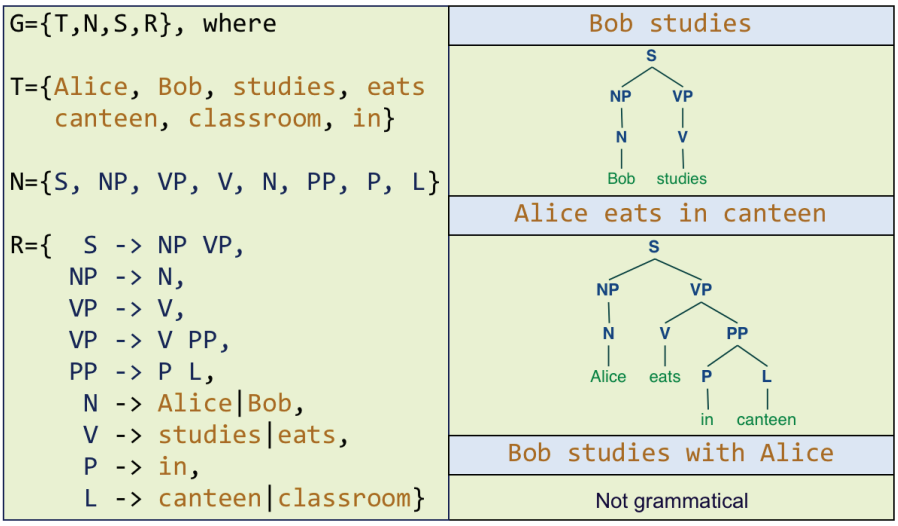


Figure 6.2: An example of context-free grammar

With the context-free grammar G defined, we could derive the language $L(G)$ for G. As a toy example, $L(G)$ contains so few strings (sentences) that we can enumerate and list all the sentences: $L(G)$ ={`Alice studies`, `Alice eats`, `Alice studies in classroom`, `Alice studies in canteen`, `Alice eats in classroom`, `Alice eats in canteen`, `Bob studies`, `Bob eats`, `Bob studies in classroom`, `Bob studies in canteen`, `Bob eats in classroom`, `Bob eats in canteen`}.

Two grammatical sentences in $L(G)$ and corresponding parse trees are also presented in Figure 6.2. Parse trees are used to show the structure of a grammatical sentence. A parse tree generated in respect of context-free grammar G is a finite ordered tree labeled with symbols from $N \cup T$, and

1. the root node is labeled S, indicating the start symbol,
2. for each node n labeled with a nonterminal $\alpha \in N$, there is a rule $\alpha \to \beta \in R$ and n's children are labeled β,
3. nodes labeled with terminals have no children.

ψ_G represents the set of parse trees generated by G and $\psi_G(c)$ is a subset of ψ_G that can generate grammatical sentence $c \in T^*$. The process of determining whether a sentence is a derivation of a grammar or not is called parsing. For context-free grammars, there exist

several well applied parsing algorithms that run in polynomial time, for which Rozenberg and Salomaa (1997) provides a good review. Discussion on those algorithms is beyond the focus of this study.

6.2.2 A grammar for daily activity patterns

With the previous introduction, the context-free grammar is applicable to daily activity patterns as well. In this study, we define a tour as a sequence of activities that starts and ends at home. By utilizing these tour structures, context-free grammars can be developed for activity patterns.

We first provide a simple grammar that is capable of generating well-formed activity sequences. Although the simple grammar is regular, we are going to treat it as a context-free grammar for the rest of the chapter (The treatment is reasonable given the fact that the set of regular languages is a subset of context-free languages). The terminal symbols are $T = \{$`h`, `w`, `e`, `s`, `r`, `p`, `m`, `a`, `o`, $\varepsilon\}$, where ε is a string of length 0. Nonterminal symbols are $N = \{\texttt{DP}, \texttt{H}, \texttt{T}, \texttt{A}, \texttt{E}\}$ and $\texttt{DP} \in N$ represents the start symbol. Rules in R are listed in Figure 6.3.

```
DP → H T
H → "h"
T → ε | A H T
A → E | E A
E → "w" | "e" | "s" | "r" | "p" | "m" | "a" | "o"
```

Figure 6.3: A simple grammar for daily activity sequences

The start symbol `DP` can be replaced with `H` (which leads to a home activity) followed by a possible tour `T`. `T` may either be replaced with the empty string ε or be replaced with a string of activities `A`, home activity and another possible tour `T`. In case of `T` leading to the empty string ε, the generation process is terminated and results in the pattern `h`, indicating that the person stay at home the whole day. The string of activities `A` will generate from left to right in chronicle order at least one activity.

It is self-evident that the simple grammar described above either generates a stay-at-home pattern, or patterns with a sequence of tours. In fact, almost all patterns extracted from HITS2008 belong to the language generated by the above grammar. A few exceptions are

those which start/end at non-home locations. Such exceptions can be taken into account easily by modifying the grammar and allowing daily activity patterns to start/end at non-home locations (add the following rules: `DP` $\rightarrow$ `N T`, `T` $\rightarrow$ `A N` and `N` $\rightarrow$ `"n"` where `"n"` indicates non-home locations).

6.2.3 Probabilistic grammars and learning

With the context-free grammar for daily activity patterns defined, daily activity patterns become a formal language. Given a string that consists of terminal symbols, we can easily verify whether it belongs to the language or not. However, there remains a problem: the language generated by the simple grammar has countably infinite number of sentences, which contradicts the fact that in any household travel survey, only a finite number of patterns, or subset of sentences, can be observed. Moreover, even in the set of observed patterns, the frequency of some patterns are considerably high, while there still exist patterns which are seldomly selected, as shown in Figure 6.1. Recalling the analogy between daily activity patterns and natural languages, similar phenomenon is observed in natural languages such as English. In English, it is often the case that multiple expressions can lead to the same meaning and non-native speakers will make up their own expressions, which are grammatical, yet never used by native speakers.

The grammar we have developed is incapable of explaining this phenomenon and we need to refer to probabilistic grammars. A probabilistic grammar can determine the occurrence probability of a grammatical sentence. Therefore, with probabilistic grammars, it is argued that failing to observe a pattern in a survey might be attributed to the extremely low occurrence probability of that pattern.

Given a context-free grammar $G = (T, N, S, R)$, if each rule $\alpha \rightarrow \beta \in R$ is associated with a probability $p(\alpha \rightarrow \beta)$, such that for each $A \in N$,

$$\sum_{\substack{\beta \in (N \cup T)^* \\ s.t. A \rightarrow \beta \in R}} p(A \rightarrow \beta) = 1, \tag{6.1}$$

the probabilistic context-free grammar (PCFG) can be defined as $G_p = (T, N, S, R, p)$. The

probability of a certain rule reflects that the corresponding trip-making characteristic is frequent or not. For example, in the simple grammar, the probability of `E` $\rightarrow$ `"e"` should be significantly low if only a few people in the sample are students and need to conduct education activities.

With properly assigned probability, the context-free grammar can reproduce the probability of each sentence in the language. In Booth and Thompson (1973), researchers provided the formula to calculate the probability of a parse tree or a sentence of a PCFG. Given a parse tree ω produced by PCFG G_p, the probability of observing ω in respect of G_p is:

$$P(\omega) = \prod_{\alpha \rightarrow \beta \in R} p(\alpha \rightarrow \beta)^{f(\alpha \rightarrow \beta | \omega)}, \tag{6.2}$$

where $f(\alpha \rightarrow \beta|\omega)$ is the frequency of rule $\alpha \rightarrow \beta$ in ω.

Given a grammatical sentence $c \in T^*$ (c is the string generated by of one or many parse trees), since c can be generated by multiple parse trees, the probability of observing the sentence c is defined as:

$$P(c) = \sum_{\omega \in \psi_{G_p}(c)} \prod_{\alpha \rightarrow \beta \in R} p(\alpha \rightarrow \beta)^{f(\alpha \rightarrow \beta | \omega)}. \tag{6.3}$$

The mapping from a sentence c to its parse trees $\psi_{G_p}(c)$ is one-to-many. Therefore, calculation of $P(c)$ requires the sum of probability of every possible parse tree $\omega \in \psi_{G_p}(c)$.

In order to find the probability for each rule that can best replicate the pattern distribution, we next learn those probabilities from data.

6.3 Problem Formulations

In this section, several formulations are proposed to learn the probability of each rule in a context-free grammar from travel survey data. Here, we introduce three formulations, which form a hierarchical structure. Notably, in this structure, the last formulation is a generalization of the first two.

6.3.1 Formulation 1: Estimating a basic PCFG with observed daily activity patterns

Consider the case where the observations are parse trees $\{\omega_1, \omega_2, ..., \omega_n\}$. In fact, for the case of daily activity patterns and the simple previously defined grammar, this is equivalent to corpus of pattern strings $\{c_1, c_2, ..., c_n\}$ as one-to-one mapping is achievable: the simple grammar is defined such that tours in an activity pattern as well as activity sequence in a tour are both generated strictly from left to right and no alternative parse trees can be found for the same daily activity pattern. To maximize the probability of observed daily activity patterns, we define the likelihood function as:

$$\begin{aligned} L &= L(p, \{\omega_1, \omega_2, ..., \omega_n\}) \\ &= \prod_{i=1}^{n} \prod_{\alpha \to \beta \in R} p(\alpha \to \beta)^{f(\alpha \to \beta | \omega_i)}. \end{aligned} \tag{6.4}$$

Take the constraint in Equation 6.1 into consideration and rewrite the likelihood function in Equation 6.4 as log form. Maximum likelihood estimates of probability can be acquired by solving the following maximization problem:

$$\begin{aligned} &\text{Max } LL = \sum_{i=1}^{n} \sum_{\alpha \to \beta \in R} f(\alpha \to \beta | \omega_i) \log^{p(\alpha \to \beta)} \\ &s.t. \\ &\sum_{\substack{\tilde{\beta} \in (N \cup T)^* \\ s.t. A \to \tilde{\beta} \in R}} p(A \to \tilde{\beta}) = 1, \forall A \in N, \end{aligned} \tag{6.5}$$

where $p(\alpha \to \beta), \alpha \to \beta \in R$ are the decision variables. The above problem can be easily solved with Lagrange Multipliers and a closed-form solution is available (see Chi and Geman, 1998):

$$\hat{p}(\alpha \to \beta) = \frac{\sum_{i=1}^{n} f(\alpha \to \beta | \omega_i)}{\sum_{\tilde{\beta} \in (N \cup T)^* s.t. \alpha \to \tilde{\beta} \in R} \sum_{i=1}^{n} f(\alpha \to \tilde{\beta} | \omega_i)}, \forall \alpha \to \beta \in R. \tag{6.6}$$

Probabilistic context-free grammars may suffer from inconsistent distribution where the

sum of probability of every grammatical string in a language is less than 1. As a statistical property of the probabilistic context-free grammar whose probability is estimated from a set of parse trees, Chi and Geman (1998) proved that the estimated probability is tight, suggesting that the estimated probability is consistent and probability of every possible daily activity pattern will sum up to 1.

6.3.2 Formulation 2: Estimating segment-specific probability for PCFG

The above section provides the basic problem formulation for estimating a probabilistic grammar with maximum likelihood estimation, using parse trees of observed daily activity patterns. The estimated probability is tight and can be used to simulate new daily activity patterns and their probability. It is assumed in Formulation 1 that each individual in the survey shares the same probability set. An immediate improvement of the above formulation is to segment the population by socio-demographic characteristics and estimate segment-specific probability sets separately.

One of the promising variables is person type, which is often defined to reflect economic activities and social roles. Apparently, people of different person types are supposed to have different activity preferences (see Hanson, 1982; Pas, 1983 and Sang et al., 2011). For example, it is unlikely for the retired to conduct work activities, and the chance should be considerably high that full-time students will go for education activities.

Suppose that given a variable for segmentation, the population can be divided into m segments and each individual belongs to exactly one segment (exclusive and exhaustive), the problem is formulated as follows:

$$\begin{aligned} &\text{Max } LL = \sum_{k=1}^{m} \sum_{i=1}^{n_m} \sum_{\alpha \to \beta \in R} f(\alpha \to \beta | \omega_{k,i}) \log^{p_k(\alpha \to \beta)} \\ &s.t. \\ &\sum_{\substack{\tilde{\beta} \in (N \cup T)^* \\ s.t. A \to \tilde{\beta} \in R}} p_k(A \to \tilde{\beta}) = 1, \forall A \in N, \forall k. \end{aligned} \tag{6.7}$$

By estimating different probabilities for different segments, for a given context-free grammar, the formulation is supposed to improve the likelihood. However, it remains a question that how much the likelihood will be improved.

The segmentation of population can be based on multiple variables as well, so long as the segmentation is exclusive and exhaustive. For example, the segmentation can be based on person type, gender and income level. A major problem with adding more segmentation variables is that, sooner or later, curse of dimensionality (Bellman, 1961) will occur — there are some segments with no population and the probabilities for those segments cannot be estimated.

6.3.3 Formulation 3: Incorporating latent variables

With the limitation of population segmentation described in the last section, an alternative formulation is proposed in this section to overcome the problem.

For each rule $\alpha \to \beta \in R$ and each individual i, a score function $V_{\alpha\to\beta,i} = V_{\alpha\to\beta}(\mathbf{X}_i)$ is associated where $\mathbf{X}_i$ is a vector of socio-demographic variables of individual i that can be observed directly from survey. The calculated score for each rule is a latent variable inferred from $\mathbf{X}_i$. Logistic function form for probability is defined to meet the constraint that the sum of probability of rules with the same left head is 1, as shown in Equation 6.8: for individual i,

$$p_i(A \to \beta) = \frac{e^{V_{A\to\beta,i}}}{\sum\limits_{\substack{\tilde{\beta}\in(N\cup T)^* \\ s.t. A\to\tilde{\beta}\in R}} e^{V_{A\to\tilde{\beta},i}}}, \forall A \in N. \tag{6.8}$$

Now the problem formulation can be expressed as:

$$\begin{aligned}\text{Max } LL &= \sum_{i=1}^{n} \sum_{\alpha\to\beta\in R} f(\alpha\to\beta|\omega_i)\log^{p_i(\alpha\to\beta)} \\ &= \sum_{A\in N} \sum_{\substack{\tilde{\beta}\in(N\cup T)^* \\ s.t.A\to\tilde{\beta}\in R}} \sum_{i=1}^{n} f(A\to\tilde{\beta}|\omega_i)\left[V_{A\to\tilde{\beta},i} - \log\left(\sum_{\substack{\gamma\in(N\cup T)^* \\ s.t.A\to\gamma\in R}} e^{V_{A\to\gamma,i}}\right)\right].\end{aligned} \tag{6.9}$$

This formulation is complicated and it usually contains much more parameters to be estimated. For implementation, however, it can be separated into several maximization problems. Suppose $V(\mathbf{X})$ is linear in parameters, which is $V_{\alpha\to\beta} = b_{\alpha\to\beta,0} + \sum_{k=1}^{m} b_{\alpha\to\beta,k} \times x_k$. Notice that for given individual, the probability of a rule $p(A\to\beta)$ is only affected by rules with the same left head $\{A\to\tilde{\beta}|\tilde{\beta}\in(N\cup T)^*, s.t. A\to\tilde{\beta}\in R\}$. Thus, two rules with different left head A will contribute independently to the overall likelihood function. Therefore, followed by the second equation mark in Equation 6.9, we change the order of sums and make the likelihood function a sum over all $A\in N$. The part to the right of $\sum_{A\in N}$ can be optimized individually for each $A\in N$.

The third formulation is a generalization of the first two. If we define the score functions with rule-specific constants only, it becomes equivalent with the first formulation. Furthermore, if we define the score functions with rule-specific constants and person type dummy variables, it becomes equivalent with the second formulation. As a generalization, the third formulation can be even more flexible by adding additional variables into the score functions.

The three formulations introduced above can be used to fit a probabilistic context-free grammar with observed daily activity patterns. From basic formulation to population segmentation and finally to incorporation of latent variables, more information from the population is introduced to the model. With additional information, it is expected the model will have a better fit as well as better prediction accuracy.

Beyond prediction accuracy, parameters estimated from the three formulations are interpretable. As mentioned in Section 6.2.3, the probability of a certain rule reflects that the corresponding trip-making characteristic is frequent or not. For the second formulation, the probability of rules can further differentiate trip-making characteristics among people of

different person types. For example, workers should make more tours than the retired. Thus, the probability of $\mathtt{T} \rightarrow \mathtt{A\,H\,T}$ should be higher for workers when compared with the retired. The probability of rules are not directly estimated in the third formulation. However, the sign and scale of the estimated parameters in the third formulation are interpretable and serve the purpose of validating the estimation results.

With the estimated probabilistic context-free grammar, probability of a particular daily activity pattern can be calculated for the whole population, a segment of population, or a particular individual (depending on which formulation is adopted). Context-free grammars ensure that all grammatical daily activity patterns can be generated and the estimated probability determines the probability of occurrence for each pattern. Thus, certain daily activity patterns may have extremely low occurrence probability based on the learned grammar and will not be observed in the data.

6.4 Experiments and Insights

Experiments based on the theory and formulations described in the last section are conducted. Specifically, a PCFG is first estimated with Formulation 1. An alternative grammar for daily activity patterns is then estimated and our results show that the alternative grammar outperforms the simple one. With the alternative grammar, population is segmented by person type and Formulation 2 is adopted to estimate segment-specific probability. It turns out that with this formulation, the log likelihood as well as the prediction accuracy of the PCFG improves dramatically. Finally, additional socio-demographic variables are added to build the score functions introduced in 3.3 and Formulation 3 is applied. It is shown in this experiment that Formulation 3 performs best in terms of likelihood and the ability to predict individual patterns.

6.4.1 Learning daily activity patterns with simple PCFG

Recall the simple context-free grammar for daily activity patterns defined previously. This section starts with estimating its corresponding probability using Formulation 1. Notice that

since DP and H only appear once as left head, the probability for the two rules (DP $\rightarrow$ H T and H $\rightarrow$ "h") is guaranteed to be 1 and does not need to be estimated. Thus, probabilities are to be estimated for the rest of rules. With this context-free grammar, the daily activity patterns observed in HITS2008 are parsed and a total of 34,432 parse trees from population of the same size are generated and used for calculating the probability with Equation 6.6. The log likelihood value of this formulation is $LL = -114,014.63$ for the full sample.

We found that the results from this grammar are poor and inaccurate when used for prediction. One main reason is that the simplified rules in the grammar prevent us from distinguishing tours/stops at different orders. For example, the probability of making the very first tour and the probability of making a second tour after finishing the first one should be different. Nevertheless, the results from this grammar can be used as a benchmark for comparison when advanced grammars are introduced.

Next, we consider an alternative grammar with more details in Figure 6.4 .

```
DP → H T1
H → "h"
T1 → ε | A1 H T2
T2 → ε | A2 H T3
T3 → ε | A3 H T3
A1 → E1_1 | E1_1 A1_2
A1_2 → E1_2 | E1_2 A1_2
A2 → E2_1 | E2_1 A2_2
A2_2 → E2_2 | E2_2 A2_2
A3 → E3_1 | E3_1 A3_2
A3_2 → E3_2 | E3_2 A3_2
E1_1 | E1_2 | E2_1 | E2_2 | E3_1 | E3_2 →
"w" | "e" | "s" | "r" | "p" | "m" | "a" | "o"
```

Figure 6.4: An alternative grammar for daily activity sequences with more rules

As shown, the above alternative context-free grammar has a larger number of rules. In the rule set, the first 3 tours in a daily activity pattern are distinguishable, so are the first 2 activities in the activity sequence of each tour. The 4th and rest of tours have the same reproduction probability. The 3rd and rest of activities in the same tour have the same reproduction probability. HITS2008 has shown consistent results with our formulation: most people will make up to three tours with only a few exceptions and over 80 percent of the

observed tours have only one activity. Therefore, the selected cut-off point guarantees that the probability of this context-free grammar is able to be estimated from data.

So far, two grammars have been generated manually by utilizing tour structures in activity sequences. It is worth noting that modelers could design and fine-tune the grammar such that it can best suit the city or region of interest where the data was collected. The process can be carried out manually, taking advantage of modelers' knowledge of the city (for example, sub-structures such as work-based sub-tours may be common in the observed activity patterns, or people like to eat at food courts before going home. Rules that describe those sub-structures can be added to the grammar). The process can also be carried out with grammar induction algorithms. To customize an activity pattern grammar for a city or region of interest is a good extension of the grammar introduced in this paper and can be an attractive topic for investigation.

It turns out that by using the full sample for estimation, the log likelihood value has increased to $LL = -94,455.09$ with the alternative grammar. So far, two probabilistic context-free grammars have been estimated. By comparing the likelihood only, the alternative grammar outperforms the simple one. The superior performance of the alternative grammar is also reflected in its out-of-sample prediction power. 75 percent of the data are randomly selected and used for estimation, and the rest are used for validation. By using the estimated probability of both grammars, the predicted percentage of the 15 most frequent patterns is shown in Figure 6.5. The prediction performance of the simple grammar is poor; however, the alternative one is able to reproduce daily activity patterns with probabilities that are close to the observed probabilities in the validation group.

Probability estimated with Formulation 1 can fit probability of observed daily activity patterns for the whole population yet is unable to predict the pattern precisely for each individual as no personal information is involved in the estimation. As an improvement, the population is segmented by person type and each segment has its own probability set.

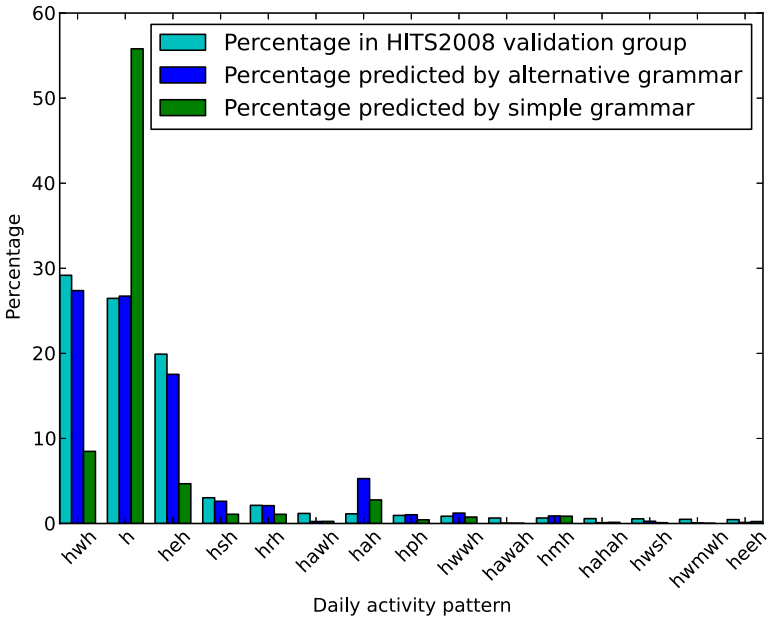


Figure 6.5: Predicted percentage of 15 most frequent patterns in the validation group

6.4.2 Learning daily activity patterns using PCFG with segment-specific probability

As is known, people with different socio-demographic characteristics tend to have different tasks to perform on daily basis, which will result in different preferences on daily activity patterns. The most obvious example of such characteristics is person type, since it is often defined to reflect economic activities and social roles. In HITS2008, person types are defined using economic activities, such as full-time worker, part-time worker, full-time student, retired, homemaker, etc.

The alternative grammar is shown to have better performance in the last experiment and is adopted in this section. With Formulation 2, a probability set is estimated for each segment. The overall log likelihood value has dramatically increased to $LL = -66,387.61$ for the full sample. Same result is achievable by using Formulation 3 and adding person type dummy variables in the score functions. Compared with Formulation 1 (where $LL = -94,455.09$), this formulation can better fit the probability of observed daily activity patterns in the survey. The increased likelihood also suggests that there is a direct and obvious relationship

between person types used for segmentation and the activity participation patterns.

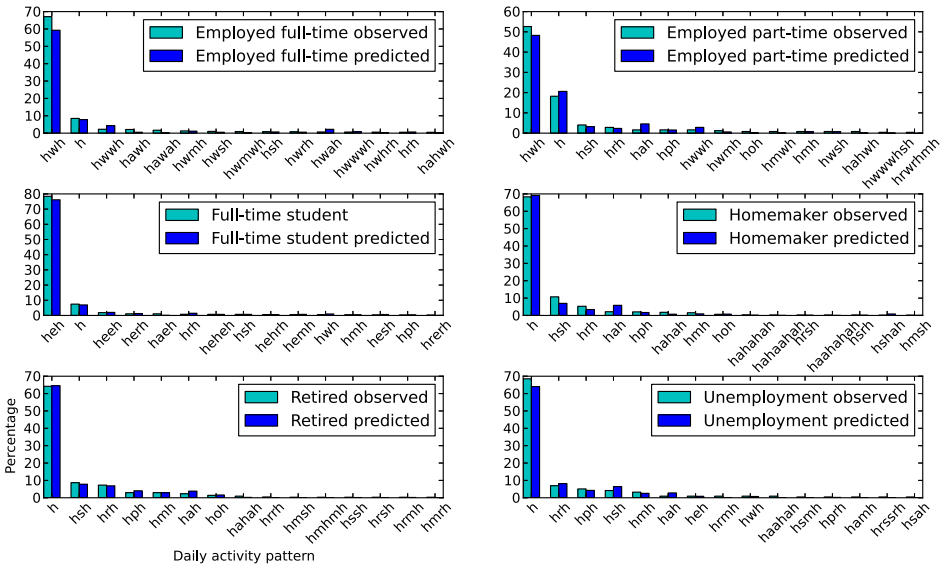


Figure 6.6: Predicted percentage of 15 most frequent patterns by segment

To show the out-of-sample prediction power of this formulation, 75 percent of the parse trees are randomly selected and used for estimation and the rest are used for validation. By using the estimated segment-specific probability, the predicted percentage of the 15 most frequent patterns for selected person types in the validation group is shown in Figure 6.6. For a given person type, people's choices of daily activity patterns are still concentrated on several most frequent patterns. Furthermore, the figure also suggests that it is necessary to estimate a different set of probability for each person type, as people from different person types tend to have different activity participation patterns and daily activity scheduling patterns. With segment-specific probability in respect of the alternative grammar, the PCFG can now reproduce probability of daily activity patterns with much higher precision for all person types.

6.4.3 Learning daily activity patterns using Formulation 3

As a generalization of learning segment-specific probability in Formulation 2, Formulation 3 is much more flexible in terms of adding covariates in the score functions and providing

individual-specific probability sets. In the experiment with Formulation 3, we add the following variables to the score function of each rule: rule-specific constant, dummies for person types, income levels, student types and presence of children in the household. These socio-demographic variables are added based on modeling experience (Pas, 1984). If only rule-specific constants and dummies for person types are used in score functions, the estimation results would become the same as the experiment with segment-specific probability sets.

The optimization of Equation 6.9 is solved after decomposition and the final log likelihood for the full sample is $LL = -65,685.87$. Compared to the likelihood achieved in 6.4.2, or using only rule-specific constants and dummies for person types in the score functions (where $LL = -66,387.61$), the likelihood actually improves. The result is reasonable because more variables are added in the score functions and it is expected that the likelihood will increase. The likelihood ratio test also suggests that the additional covariates (dummies for income levels, student types and presence of children in the household) are significant as a whole.

It is also shown that the prediction accuracy of individual pattern is increased, when compared to the experiment in 6.4.2. We randomly sample 75 percent of the parse trees and use them for estimation. The rest are used for validation. The specification introduced in this section as well as the experiment in 6.4.2 are estimated using the same population. To calculate the prediction accuracy of individual pattern, for each person in the validation data, we would generate the pattern with the largest probability, which is calculated with the probability set estimated from 6.4.2 as well as the score functions estimated from this section. The pattern is correctly predicted for the person if it is the same with the actual choice. For the two experiments in 6.4.2 and this section, we achieve an average coverage of 68.5 percent and 69.7 percent respectively. Based on the measurement, we may conclude that Formulation 3 is able to increase the prediction accuracy of each person. From the results of out-of-sample prediction, there is no sign of overfitting (otherwise a decreased prediction accuracy is expected). In future application of this formulation, estimation can be carefully designed by incorporating cross-validation procedures in order to prevent overfitting.

6.5 Concluding Remarks

6.5.1 Summary

In this chapter, daily activity patterns are formulated as activity sequences. Similarities between activity sequences and languages are first explored. As a result, the idea that the methodology used to understand languages — especially natural languages — is applicable to daily activity patterns arises spontaneously. Different context-free grammars are proposed for daily activity patterns and the flexibility provided by context-free grammars enables modelers to develop grammars that can better reflect the characteristics of the interest region.

To learn daily activity patterns in respect of particular context-free grammar, the concept of probabilistic context-free grammar is introduced and three problem formulations are proposed. Those formulations are able to estimate the probability of a context-free grammar with observed daily activity patterns such that after estimation, the probabilistic context-free grammar can be used to reproduce the probability of observed patterns in household travel survey and weighted random sampling from an infinite set of grammatical activity sequences without a priori limitation of the consideration set becomes possible.

As an ultimate goal, the probabilistic context-free grammar is expected to reproduce the observed probability of patterns precisely. Whether this can be done effectively is shown in the experiments to be affected by both the context-free grammar in use and the problem formulation. In the section of experiments, two context-free grammars are proposed and their probabilities are estimated with Formulation 1 using daily activity patterns observed in HITS2008. It is shown that the alternative grammar outperforms the simple one, which indicates that the definition of the context-free grammar and rules used by the grammar will have an effect on the prediction power of it. Furthermore, by fixing the context-free grammar, the problem formulation will also influence the prediction performance. Formulation 3 outperforms the other two with best results of likelihood and the ability to predict individual patterns.

Both Formulation 2 and 3 are able to reflect heterogeneity in choosing daily activity patterns

to certain extent. Compared to Formulation 2, Formulation 3 is more flexible and able to produce individual-specific rule probability sets, which leads to better fit of data and performance in predicting individual patterns. While closed-form solution is not available for Formulation 3 in general, the problem can be decomposed and solved in parallel, which makes Formulation 3 empirically applicable.

6.5.2 Potentials in activity-based modeling and policy decision-making

Daily activity pattern models are crucial in many activity-based modeling frameworks. The grammar-based representation of daily activity patterns is especially useful in a class of activity-based implementations initiated by Bowman (1998) and Bowman and Ben-Akiva (2001).

In Bowman's approach, the decisions at tour and trip levels are usually conditioned on the selection of daily activity patterns. For many state-of-practice activity-based models that fall in this class, the selection of daily activity patterns is based on a pre-specified choice set. The pre-specified choice set may not be realistic for daily activity patterns as the universal set for daily activity patterns is large if not infinite and the criteria to determine the set of feasible alternatives is not explicitly modeled. While Manski (1977) provided a two-stage theoretical framework where the first stage consists of estimating the probabilities of all possible choice sets of the universal sets, direct application of the framework is computationally intractable. In a setting of mode choice, Swait and Ben-Akiva (1987) first generated the probability that an alternative is part of an individual's choice set or not, then estimated the probability for each possible choice set of the universal set. Cantillo and de Dios Ortúzar (2005) argued that thresholds, which can be expressed as a function of socio-economics characteristics of the individual and the conditions under which the choice process takes place, can be used to determine whether a choice is in the choice set or not. The discussion on generating choice sets with respect of individual characteristics and choice scenarios is also presented in a list of recent studies, such as Kaplan et al. (2009), Kaplan et al. (2011), Castro et al. (2013), and Vij et al. (2013). However, those studies only consider the choice scenarios where the universal set is manageable. The extension of their approaches to daily activity patterns

is questionable as the universal set for daily activity patterns is huge or even infinite (it is infinite if daily activity patterns are defined as activity sequences).

Learning daily activity sequences with probabilistic grammars proposed in this chapter provides a bottom-up approach for generating choice sets of daily activity patterns for target groups, or individuals. Furthermore, characteristics of the groups or individuals are considered in the process. Formulation 3 proposed in the study is able to produce individual-specific rule probability sets. A choice set can be generated for each person by repeated sampling from an infinite set of grammatical activity sequences using weights calculated with individual-specific rule probabilities.

The choice of daily activity patterns is heterogeneous among the population, which is usually influenced by socio-demographic variables associated with the decision makers. However, it is common for accessibility measurements to influence the choice as well. For example, a low accessibility for shopping activities may hinder decision makers from making shopping tours. While socio-demographic variables can be stable for a particular person in a long time period (job, household location, household composition, etc.), accessibility may change due to infrastructure development and upgrades in transportation systems, which usually take place within the planning period. Since accessibility is time-dependent and may influence an individual's perception of consideration set for daily activity patterns, it is reasonable to require that the daily activity pattern models in Bowman's activity-based modeling framework and implementations should be able to generate choice sets that will be influenced by development in infrastructure and transportation systems. Models that follow the requirement should be able to predict patterns that may not be currently well observed in the data that used for estimation if accessibility measurements change in simulation scenarios or in the real world.

It is highlighted in this study that the selection of variables used in Formulation 3 is very flexible. Besides socio-demographic variables, accessibility measurements for different types of activities may be included to influence rule probabilities , which will change if the development in infrastructure and transportation systems changes the current level of accessibility. Therefore, it is possible to generate accessibility-dependent choice sets for each

individual such that different policy scenarios can be applied to test the influence to daily activity patterns. Moreover, no grammatical patterns are ruled out a priori. Whether a pattern can be observed in a choice set of limited size is determined by its weight and the weight is influenced by both socio-demographic variables and accessibility.

6.5.3 Future studies

Formal grammars have been well documented by research communities in linguistics, computational biology and computer science; however, its intrinsic similarity and consistence with human activity patterns have long been ignored. As the first attempt to establish the connection between daily activity patterns and formal languages, this study borrows the concept of probabilistic grammar and applies it in modeling and learning human behaviors. Taking advantage of a large-scale individual-based survey (HITS2008), we examined the applicability of this approach and confirmed that the use of activity sequences (in activity pattern) and words (in language) can be modeled using the same framework. Most important, this connection is a remarkable example to show that new perspective and methodology could be introduced to transportation modeling. To make the motivation of this work more attractive, we hope to expand its scope to broader horizons by further efforts on promising directions that follow:

1. Firstly, it is possible to develop different context-free grammars for different segments of population, which is a directly improved formulation from Formulation 2.

2. Secondly, two context-free grammars are proposed manually in this chapter and based on the experiments, the alternative grammar has better performance. However, it is not proved that no better context-free grammars can be defined. Generating a context-free grammar with the best rule set, formally known as grammar induction, is a hot topic in linguistics and computer science. Grammar induction should utilize modelers' knowledge on the region of interest and the derived grammars should best suit traveling characteristics in the region. There have been works on inducing a context-free grammar using observed sentences from natural language and programming language, see Wyard (1993), Javed et al. (2004) and Dubey et al. (2008). Similar approach can be

adopted to derive a best context-free grammar with the observed daily activity patterns in household travel survey. And this chapter already provides problem formulations that can easily evaluate the performance of any derived grammar.

3. Thirdly, the application of learning daily activity patterns with probabilistic grammars can be explored in the context of daily activity pattern models. A two-stage model can be built for the choice of daily activity patterns. In the first stage, a choice set of patterns is customized for each individual. The choice of patterns will be carried out in the second stage.

Chapter 7

A Two-Stage Choice Model for Daily Activity Patterns

A grammar-based representation of daily activity patterns is explored in the last chapter with several problem formulations and experiments. As it unfolds in this chapter, we are going to show the implication of the new perspective and how the modeling of daily activity patterns can take advantage of the grammar-based representation. In this work, we explore a probabilistic context-free grammar (PCFG) based representation of daily activity patterns within the framework of discrete choice modeling and propose a two-stage choice model for daily activity patterns where a choice set is customized for each individual based on socio-demographic variables, followed by a Multinomial Logit model to determine the chosen pattern. Moreover, in sprite of the experience in route choice modeling, several additive terms are proposed and able to capture the correlations among alternatives because of overlapping. Then, with promising results of replicating the activity participation behavior in the base year, the two-stage model can be incorporated into the pre-day activity-based modeling framework introduced in Chapter 5.

7.1 Introduction

In the context of activity-based modeling, the concept of daily activity pattern is crucial to distinguish between activity-based and tour-based travel demand models. However, unlike the modeling of travel mode or destination, there exist no unambiguous labels for daily activity patterns and researchers always need to define daily activity patterns in the first place. Long before any operational activity-based models exist, there have been studies on the subject. In Adler and Ben-Akiva (1979), the researchers defined daily travel patterns as the ordered sequences of trips with detailed trip information and specified a utility maximization model of it. However, it is criticized that the set of such defined activity-travel patterns is virtually infinite and Adler and Ben-Akiva (1979) provided no means for distinguishing the activity-travel patterns actually considered nor did it address the statistical and computational issues associated with enumerating large set of alternatives and incorporation of interrelations (Horowitz, 1980). Moreover, the defined sequence contains detailed activity-travel decision-makings, such as mode, destination and time of day choice, making it a detailed plan rather than a skeleton of a plan. Several rule-based activity-based models accept the detailed representation and intend to generate full-day activity-travel plans in an activity scheduling process, see for example STARCHILD (Recker et al., 1986a and Recker et al., 1986b), SMASH (Ettema et al., 1993 and Ettema et al., 1996) and AMOS (Kitamura et al., 1996).

While a number of operational activity-based travel demand modeling frameworks still feature the activity scheduling process and treat daily activity patterns as travel plans with details, such as ALBATROSS and CEMDAP, there have been many modeling frameworks that define daily activity patterns alternatively as the abstraction and skeleton of the daily activity-travel plan with limited information. In a classic and well-adopted activity-based modeling framework that falls in this class, Bowman and Ben-Akiva (2001) developed a three-level representation of the model structure such that the choice of daily activity patterns is modeled first at the top level, followed by the decision-makings at tour and trip levels. In this hierarchical structure, although wrapped in the same framework with other activity participation and travel decisions, such as mode, destination and time of day, there is a

fundamental difference between the choice of daily activity patterns and the others: it is definition-specific. For mode choice and destination choice, the alternatives are self-evident and easy to identify and envision. For route choice, although routes are mostly unknown and hidden in the network from which they need to be explicitly extracted (Prato, 2009), the definition of a route or path between any OD pairs is self-evident. However, the definitions for daily activity patterns vary and the modeling of them largely depends on how they are defined in the first place.

The pre-day activity-based travel demand model for Singapore introduced in Chapter 5 adopts Bowman's approach and defines daily activity patterns as the occurrence of tours and intermediate stops of different activity purposes. As mentioned in Chapter 6, another popular definition of daily activity patterns for activity-based modeling frameworks with Bowman's approach is daily activity sequences. Regardless of the definitions, if the selection of daily activity patterns is modeled with discrete choice analysis, a choice set for daily activity patterns shall be generated.

As we mentioned in Chapter 6, the pre-specified choice set may not be realistic for daily activity patterns as the universal set for daily activity patterns is large if not infinite and the criteria to determine the set of feasible alternatives is not explicitly modeled. In this study, we define a daily activity pattern as a sequence of activities. By taking advantage of the probabilistic context-free grammar (PCFG) based representation of daily activity patterns, we propose a two-stage choice model for daily activity patterns. In the first stage, PCFG is used to characterize the structure and deterministic distribution of daily activity sequences. The choice set is then generated by repeated weighted sampling from a universal set with infinite number of activity sequences. The weights are generated for each individual based on the estimated probabilistic context-free grammar and are influenced by socio-demographic variables. In the second stage, the customized choice sets generated are applied in discrete choice models. While the customized choice sets require a flexible choice structure for the choice model, we are able to take correlations among alternatives into consideration with commonality factors, activity or rule size variables developed for daily activity patterns.

7.2 Grammar-based Representation of Daily Activity Patterns

Recall the definition of daily activity patterns in the pre-day model. Each pattern only determines if tours or intermediate stops will occur for each considered activity purpose. The exact number of tours, sequence of tours, generation of intermediate stops and sub-tours are modeled afterward. The logic of the modeling framework becomes complicated and prone to error as more individual models are to be specified and estimated. In contrast, For this study, a daily activity pattern is defined as a sequence of activities. The advantage of adopting this definition in an activity-based modeling framework is that it will simplify the modeling framework as more information is embedded in daily activity sequences. However, there are two major challenges to apply daily activity sequences in an activity-based modeling framework like the pre-day model.

Firstly, an immediate effect of adopting this definition is that the number of daily activity sequences will explode if no constraints are issued on the choice set. In fact, even though we can limit the length of the pattern, indicating that only a maximum number of activities can be conducted in the real world due to time constraint, the number of included activity patterns is still considerably large. We may assume that there are 8 activity types. Suppose the maximum length of an activity sequence (including home activities) is 3, only 9 patterns can be defined. Once the threshold increases to 6, over 5,000 patterns can be defined. It becomes a critical problem that how the infinite number of possible combinations can be represented with a limited number of features.

Secondly, to understand the second challenge, we could take a closer look at such defined daily activity patterns in household travel survey data. From HITS2008, a total number of 1,001 unique patterns are observed and Table 7.1 provides some information on the distribution of patterns among the population. As can be seen from the most frequent patterns, for a group of people characterized by economic activity, they tend to have a relatively unique set of daily activity patterns, which is shaped by their person types and other socio-demographic characteristics. Similarly, the diversity of the patterns in the choice set is also influenced

by these characteristics. For example, student is the second largest group in the survey yet only 189 unique patterns are observed. In terms of the trip chaining characteristics, while people without mandatory activities tend to travel less, their trip chaining behavior within a home-based tour is more complicated than students but less complicated than workers probably due to lack of mobility convenience.

Table 7.1: Distribution of daily activity patterns among the population

	All	Full-time worker	Student	Homemaker	Retired
Sample size	34,432	13,105	8,843	4,849	2,482
Number of unique patterns	1,001	532	189	203	115
Trip chaining characteristics (std)					
Avg. length of patterns	2.79 (1.43)	3.29 (1.23)	3.03 (0.79)	1.91 (1.68)	1.89 (1.39)
Avg. number of activities	1.00 (0.96)	1.31 (0.98)	1.07 (0.53)	0.50 (1.00)	0.49 (0.82)
Avg. number of tours	0.79 (0.56)	0.98 (0.39)	0.96 (0.32)	0.41 (0.72)	0.40 (0.55)
Avg. number of activities per tour	1.26 (0.71)	1.35 (0.83)	1.11 (0.40)	1.22 (0.60)	1.19 (0.56)
Most frequent patterns[1](percentage)					
1	hwh (29.49%)	hwh (67.26%)	heh (78.33%)	h (68.94%)	h (64.46%)
2	h (26.67%)	h (7.94%)	h (7.07%)	hsh (9.73%)	hsh (8.78%)
3	heh (20.24%)	hawh (2.48%)	heeh (1.46%)	hrh (4.78%)	hrh (7.41%)
4	hsh (2.85%)	hwwh (1.89%)	herh (1.18%)	hah (2.23%)	hph (3.75%)
5	hrh (2.09%)	hawah (1.69%)	hrh (1.06%)	hahah (2.17%)	hmh (2.82%)

[1] Same notations are used as in Figure 6.1

The analysis above should provide some insights for modeling daily activity patterns as activity sequences in an activity-based modeling framework. It was argued in Bovy (2009) that the choice set formation and choice from considered options are distinct mental processes in the context of route choice. However, it is the case for the modeling of daily activity patterns as well. As the second challenge faced by modeling daily activity patterns as activity sequences, the great number of alternatives suggests that a decision maker will have to form a choice set for daily activity patterns consciously or not, in order to make successive decisions. The choice set of patterns being considered by a decision maker is not explicitly acquirable from household travel surveys containing trip information for a single day. Rather, the choice set is inferred from a group of people with similar characteristics under the assumption that the constraints considered by the decision maker are correlated with the measurement of social characteristics and performance of transportation network. We have observed in Table

7.1 that how grouping people by economic activity can result in distinct sets of patterns for different groups.

Fortunately, the two challenges may be addressed by the grammar-based representation proposed in Chapter 6. The grammar proposed for daily activity patterns puts all grammatical sequences in a set where each alternative in the set can be matched to a parse tree generated by the grammar with finite rules. We call it deterministic grammar-based representation since it solves the problem of determining whether a sequence is an activity pattern or not. However, the deterministic representation is unable to reflect the heterogeneous choice of daily activity patterns among the population, not to mention forming a customized choice set from an infinite number of activity patterns. To address this problem, we could then use the probabilistic grammars and problem formulations proposed in Chapter 6.

Table 7.2: Estimation results for rules starting with `T1` (score for `T1` $\rightarrow$ `A1 H T2` is fixed to be 0)

Variable	Parameter (t-stat)
Constant	-0.3691 (-2.573)
Dummy: University student	0.8573 (8.634)
Dummy: Full-time worker	-2.046 (-13.96)
Dummy: Part-time worker	-1.009 (-6.211)
Dummy: Self-employed	-0.4541 (-2.967)
Dummy: Student	-2.361 (-15.60)
Dummy: Homemaker	1.172 (8.015)
Dummy: Unemployed	0.9672 (6.494)
Dummy: Retired	1.000 (6.230)
Dummy: Domestic worker	2.391 (14.49)
Dummy: Household with children under 15	-0.01007 (-0.2982)
Dummy: Personal income >8000 SGD	-0.7793 (-5.258)
Sample size	34,432
Number of estimated parameters	12
Init Log Likelihood	-23,867.137
Final Log Likelihood	-13,170.527
Rho Ratio	0.448

It is highlighted that Formulation 3 could be used to tackle the second challenge as adding additional socio-demographic variables and even accessibility measurements makes it very flexible and powerful to depict the heterogeneity among the population. For example, the rule `T1` $\rightarrow \epsilon$ | `A1 H T2` in Figure 6.4 determines whether `T1` leads to not making any tours at all or making the first tour. Since workers should make more tours than the retired, the probability of `T1` $\rightarrow \epsilon$ should be lower for workers when compared with the retired, which is reflected in Table 7.2.

A probabilistic grammar estimated using Formulation 3 with sufficient socio-demographic variables gives a probabilistic grammar-based representation to the daily activity patterns observed in the population. As each individual will have a customized set of rule probabilities, it enables a customized weight for each daily activity pattern. With the weights generated by the alternative grammar estimated in Section 6.4.3, we next explore how those weights can be used to generate a customized choice set of daily activity patterns for each individual.

7.3 Customized Choice Set Generation

The set of all daily activity patterns is infinite in theory if daily activity patterns are defined as activity sequences. The grammar-based representation of daily activity patterns allows us to describe the set using a finite number of features (rules). What's more, the probabilistic grammar-based representation takes one step further by its ability to depict the heterogeneity in the choice of daily activity sequences among the population, which has been achieved in the last chapter by estimating the probabilistic grammar from observed activity patterns.

Nevertheless, it is not self-evident so far that how such representation can benefit the modeling of daily activity patterns in an activity-based modeling framework. While the set of feasible options is extremely large, if not infinite, enumeration in the modeling of of daily activity sequence is impossible. Thus, explicit choice set generation is necessary. Moreover, from a policy point of view, it is advantageous to have an explicit distinction between choice set generation and choice in a two-stage modeling process (Başar and Bhat, 2004) because the modelers can link the impact of accessibility measurement to the changing attractiveness

of particular alternatives.

The attractiveness of particular alternatives can be linked to the weight assigned by the estimated probabilistic grammar. For each individual, we could assign a weight for each daily activity pattern by calculating the probability of the parse tree that generates the pattern using the individual-specific set of rule probabilities. And the sum of all the weights is 1. Since the selected daily activity sequence will always be in the set, the weight of it actually captures the attractiveness of this pattern to the individual. It is worth noting that the attractiveness of a particular pattern is subjective as the set of rule probabilities that derive the attractiveness, or weight, is customized and subjective.

Given the insights above, a customized choice set for each individual can be generated by repeated weighted sampling from a universal set with infinite number of activity patterns. To understand the repeated weighted sampling approach, we first consider the case of drawing one sample from the set according to the weight attached to each alternative. The sampling process is to generate a parse tree with Monte Carlo simulation and replace all non-terminal symbols into terminals. The probability of generating this particular tree (activity sequence) is exactly the weight that assigned to the activity sequence. A repeated sampling approach repeats the process of generating parse trees with Monte Carlo simulation and removes duplicate patterns until N (desired choice set size) distinct daily activity patterns are generated.

For each individual, a choice set of size N is generated. The individual is covered if the observed pattern is in the choice set. The coverage is calculated as the ratio of covered individuals over all individuals in the sample. Figure 7.1 provides the average coverage of choice set size per reproduction N and number of reproductions M, $\bar{C}(N, M)$, where N varies from 5 to 45 and M is from 1 to 10. The simulation for each combination of N and M repeats 20 times for stabilized coverage. When $M = 1$, $\bar{C}(N, 1)$ is the average coverage of a choice set with size N. With more reproductions M, M choice sets of size N are generated, then the choice sets generated are joined with duplicate patterns removed to form a larger choice set. The coverage is calculated with this larger choice set. The rationale behind the reproduction is that repeated weighted sampling from the infinite set of activity sequences

will inevitably generate repeated patterns with large probability. As the choice set size goes up, the time to generate the choice set goes up dramatically since it takes more draws to generate patterns of smaller probability. Figure 7.2 displays the average joined choice set size for various (N, M) combinations.

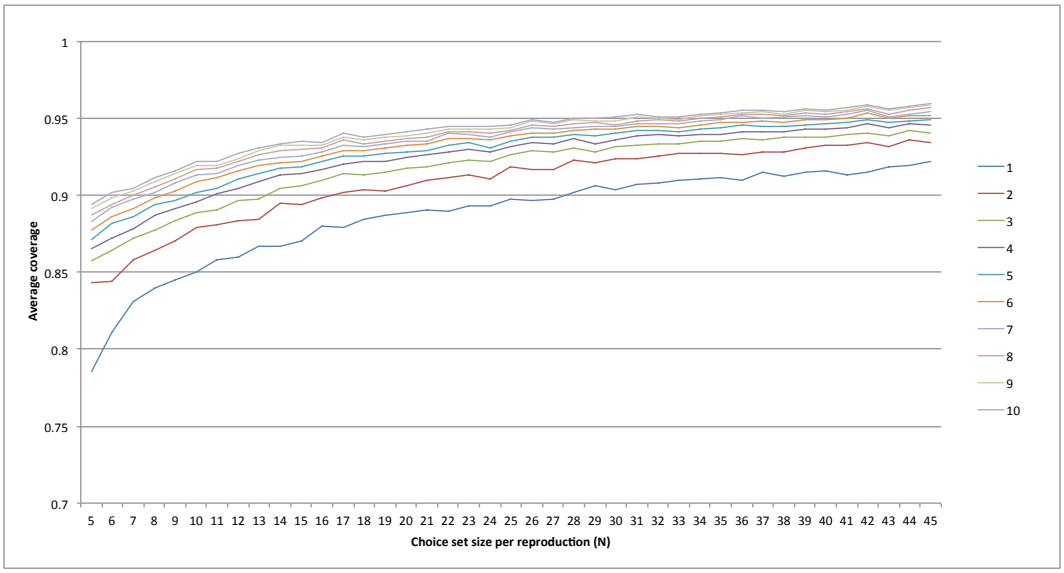


Figure 7.1: Coverage of choice sets generated by repeated sampling

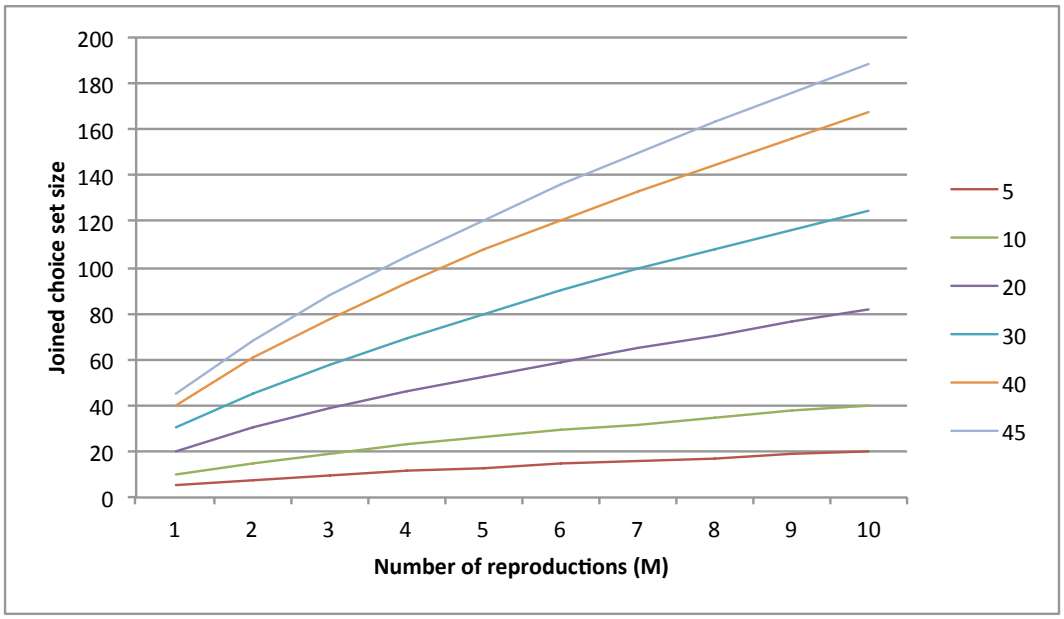


Figure 7.2: Average joined choice set size generated by repeated sampling

Figure 7.1 shows that given the choice set size per reproduction N, the coverage will increase with larger number of reproductions M. For a choice set of size 45, the coverage can reach

92 percent. With 10 reproductions, the coverage is 96 percent. However, the marginal increase of the coverage is decreasing, and the explanation to which can be found in Figure7.2: the increasing speed of the size of the joined choice set is decreasing. Generally speaking, choice set generation with repeated weighted sampling is able to produce a choice set with reasonable size and satisfying coverage. However, as the size of the choice set goes up, it takes significantly longer time to generate the choice set, which is unacceptable if a customized choice set is to be generated for each individual.

Alternatively, we could generate a choice set that features the first N patterns with the largest probabilities. To be specific, we expend the start symbol of the grammar to enumerate all parse trees to certain depth. Given the fact that context-free grammars are in favor of smaller parse trees, the set of pre-enumerated parse tress will inevitably include the possible candidates of the first N patterns with the largest probabilities. The choice set is then generated by ordering the alternatives in the set with a given set of rule probabilities and choosing the first N patterns in the ordered set. No simulation is involved in the process of choice set generation and the time it takes to generate the choice set is constant in spite of choice set size N.

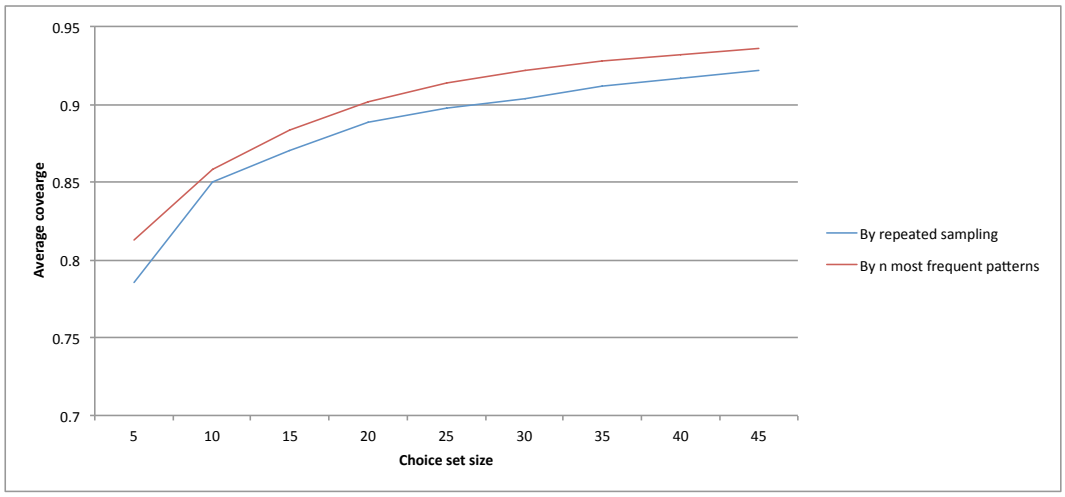


Figure 7.3: Comparison of the two choice set generation processes

Figure 7.3 reveals the difference of the two choice set generation processes in terms of the coverage. Obviously, the choice set generated by choosing the N most frequent patterns has a better coverage. In fact, it is the best coverage one can ever get when running the repeated weighted sampling process for once.

The second choice set generation process also allows us to generate much larger choice sets in much shorter time. Figure 7.4 shows the coverage of such choice sets. Overall, the coverage can reach 98 percent when the size of the choice set is 500. Figure 7.4 also presents the coverage for different person types, specifically full time workers (or workers for short) and non-workers. The coverage of choice sets generated for non-workers is always higher that that of workers. However, the gap is getting smaller with larger choice set size. It suggests that workers are living with more diverse life styles in terms of activity participation, which results in more diverse and complicated choice sets.

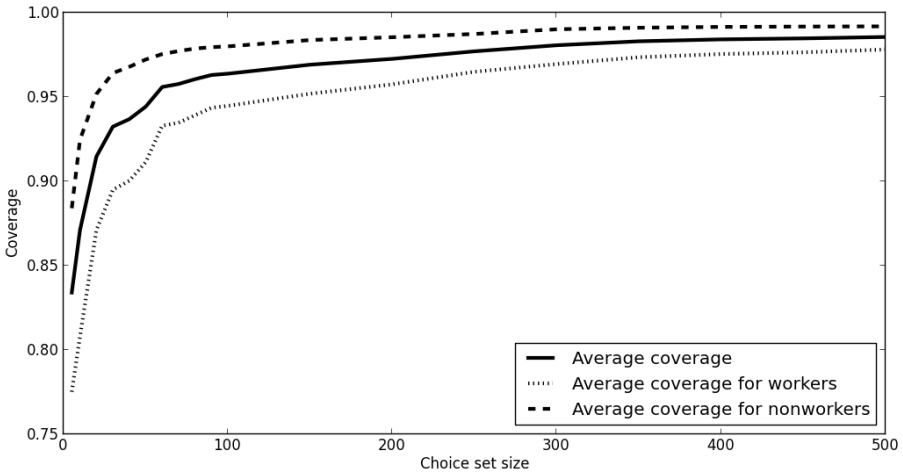


Figure 7.4: Coverage of choice sets generated with the N most frequent patterns by person type

Throughout this section, we have been discussing how to generate choice sets of daily activity sequences by taking advantage of the probabilistic grammar-based representation of activity sequences. The choice set can only be generated when the set of rule probabilities is given. In future studies, suppose both the socio-demographic characteristics and accessibility measurements can affect the rule probabilities, the choice sets generated in the first stage will evolve dynamically with the change of social-demographic characteristics of an individual as well as the development of transportation systems and network performance. Thus, the proposed model can be applied to policy scenarios where new patterns are expected to emerge from the development and upgrade of infrastructure and transportation systems.

7.4 Modeling the Choice of Daily Activity Patterns

Choice set generation is the first step in the two-stage choice model for daily activity sequences. Once a choice set is customized for each individual, discrete choice models are adopted in the second stage to model the choice of daily activity patterns. Followed by a Multinomial Logit model as the base case, several formulations are proposed in the subsequent sections to account for the correlations among alternative patterns in the choice set.

7.4.1 Base case

Discrete choice models are adopted in the second stage to model the choice of daily activity patterns. Different from ordinary discrete choice models, the choice model does not have a fixed choice structure, which is similar to route choice models. In a route choice model, the choice set will be affected by the actual network performance (see for example, Bovy and Fiorenzo-Catalano, 2007; Bovy, 2009 and Prato, 2009). In the base case, a Multinomial Logit model is applied and each individual has a customized choice set of size 50 generated in the first stage (with first N patterns). Constants listed below are interacted with socio-demographic variables in the utility functions.

- **Work, Edu, Shop, Other** Presences of work, education, shopping and other activities
- **WorkT, EduT, ShopT, OtherT** Exact number of tours in the day pattern by tour purpose
- **WorkI, EduI, ShopI, OtherI** Presence of stops in the day pattern by stop purpose
- **Mtour_Kstop** Dummy for patterns containing exactly M tour purposes and K stop purposes
- Combination of activities (e.g., if both work tours and shopping stops appear in a pattern, workshop_ts = 1)

It is noted that daily activity sequences are not characterized by unambiguous observable labels. Instead, they are described by generic attributes, such as the attributes listed above. Since the choice set is not fixed, the evaluation of each daily activity pattern should be based

Table 7.3: Estimation results of the base case model

Variable		Est. Val. (t-stat)	Socio-demographic dummy * Activity dummy			
				Work	Edu	Shop	Other
Tour constants	WorkT	-19.48(-58.72)	Part-time worker	1.95(15.65)	0(fixed)	1.93(9.94)	0.93(7.16)
	EduT	-6.31(-40.22)	Self-employed	0.73(6.54)	0(fixed)	1.25(6.44)	0.78(7.75)
	ShopT	-9.50(-23.71)	University student	0(fixed)	0(fixed)	-3.32(-6.96)	-3.12(-11.00)
	OtherT	-12.41(-28.52)	Homemaker	0(fixed)	0(fixed)	4.76(19.83)	3.88(18.68)
Stop constants	WorkI	-0.98(-4.68)	Retired	0(fixed)	0(fixed)	4.88(17.98)	4.17(17.31)
	EduI	0(fixed)	Unemployed	0(fixed)	0(fixed)	4.75(15.81)	4.53(16.25)
	ShopI	2.06(7.84)	Domestic workder	0(fixed)	0(fixed)	0.37(1.35)	1.13(8.94)
	OtherI	2.68(17.47)	Student between 16 - 19	0(fixed)	0(fixed)	-3.36(-17.07)	-4.21(-33.68)
No activity constants	at home	-2.71(-14.36)	Student between 5 - 15	0(fixed)	0(fixed)	-4.93(-14.75)	-4.42(-39.66)
	at home * full time worker	-2.97(-42.18)	Age betweeen 20 - 25	1.01(9.59)	1.57(15.36)	-0.16(-0.72)	-0.77(-6.28)
	at home * homemaker	2.08(10.67)	Age between 26 - 35	0.35(5.73)	-1.69(-12.74)	0.17(1.65)	-0.30(-4.79)
	at home * university student	-20(fixed)	Age between 51 - 65	0.702(10.39)	0(fixed)	-0.051(-0.67)	-0.123(-2.41)
	at home * student between 16 - 19	-20(fixed)	Male with children of age 0 - 4	1.72(16.42)	-1.06(-12.34)	-0.34(-1.64)	-0.48(-4.81)
	at home * student between 5 - 15	-20(fixed)	Male with children of age 5 - 15	1.59(17.81)	0.081(1.11)	-0.33(-1.84)	0.038(0.48)
	at home * retired	2.03(8.59)	Female with no child	-1.11(-14.41)	0(fixed)	0.24(2.40)	-0.38(-5.76)
	at home * unemployed	2.19(8.00)	Female with children of age 0 - 4	-0.18(-1.83)	0.093(1.10)	0.099(0.70)	-0.29(-3.33)
Sub-tour constants		0.61(3.30)	Female with children of age 5 - 15	0.17(2.02)	0.95(15.11)	0.30(2.53)	-0.12(-1.72)
Activity combination	workshop_tt	1.99(7.86)	Household with only adults	1.52(18.77)	0(fixed)	0.13(1.01)	-0.0069(-0.10)
	workshop_ts	-2.28(-7.74)	Household with only workers	0.35(4.69)	0(fixed)	0.34(2.46)	-0.19(-2.44)
	workothers_tt	2.36(11.47)	Work at home	-3.63(-14.80)	0(fixed)	1.04(4.35)	1.10(6.56)
	workothers_ts	-1.72(-14.33)	With car(s) available	-0.79(-13.03)	-0.054(-1.02)	0.34(5.30)	0.37(6.69)
	edushop_tt	6.06(19.67)	With motorcycle(s) available	0.51(5.41)	0.26(2.82)	0.093(0.78)	-0.25(-2.93)
	edushop_ts	0.72(1.99)					
	eduothers_tt	5.02(19.09)					
	eduothers_ts	0.80(5.18)					
	shopothers_tt	-0.36(-1.05)					
	shopothers_ts	-5.10(-17.04)					
Logsums	WorkT * work logsum	1.27(59.47)					
	ShopT * shop logsum	0.17(4.24)	Sample Size	28,656			
	OtherT * other logsum	0.53(9.19)	Number of estimated parameters	96			
Additional constants	one tour one stop dummy	-5.61(-44.15)	Init. Log-Likelihood	-111,209.8			
	one tour two stop dummy	-7.62(-28.59)	Final Log-Likelihood	-44,968.2			
	two tour one stop dummy	-1.59(-4.90)	Rho-square	0.596			

on generic attributes, rather than alternative specific attributes (such as the alternative specific constants in mode choice).

For a sample of 28,656 individuals used for the estimation[1], 96 parameters are estimated for the Logit model and the final log likelihood is $-44,968.2$, with a Rho ratio of 0.596. Table 7.3 summarizes the estimation results of the model with estimated parameters. In the model, the logsums for different tour purposes are generated using the pre-day simulator (in logsum calculation mode). They measure the expected utility from the mode/destination choice model when performing tours of different purposes. As expected, the parameters for logsums are all positive and significant.

The base case model suffers from the its inability to deal with correlations among alternatives in the choice set. It is intuitive that such correlations indeed exist. For example, `hwh` is a pattern with a simple work tour. `hwhsh` is a pattern with a simple work tour followed by a shopping tour. The latter is an extension based on the first one and they have a work tour in common. In the next section, several formulations are proposed to take such correlations into consideration.

7.4.2 Accounting for correlations among alternatives

The customized choice sets generated in the first stage require a flexible choice structure. In order to capture the correlations among alternatives in the customized choice sets, the proposed modeling structure should be able to assign a variance and covariance for a new pattern. The requirement rules out (Cross) Nested Logit, which requires a fixed choice structure. Inspired by the commonality factors and path size variables used in route choice models to account for correlations and preserve the advantage of estimating a Logit model in the meantime, we propose several commonality factor and size variable formulations for daily activity patterns in this section.

[1]The analysis starts with a sample of 29,207 individuals, which is exactly the same sample used to develop the pre-day model. After the choice set generation process, only those with the chosen pattern covered in the generated customized choice set are kept for the estimation.

Commonality factor for daily activity patterns

Cascetta et al. (1996) developed the C-logit approach that includes a commonality factor in the utility function for a route choice model. The commonality factor of a path is directly proportional to the degree of overlapping (similarity) of the path with other paths in the choice set. It is expected that heavily overlapped paths have larger commonality factors and thus a smaller systematic utility. Recall the specification of the commonality factor proposed in Cascetta et al. (1996):

$$CF_k = \beta_0 \ln \sum_{h \in I_{rs}} \left(\frac{L_{hk}}{L_h^{1/2} L_k^{1/2}} \right)^{\gamma}, \tag{7.1}$$

where L_{hk} is the length of the links that are common to path h and k, while L_h and L_k are the length of path h and k respectively. In fact, $\frac{L_{hk}}{L_h^{1/2} L_k^{1/2}}$ ranges from 0 to 1 and describes the similarity of path h and k. It equals 0 if the two paths do not overlap at all. On the contrary, it equals 1 if the two paths are identical.

The concept of overlapping is borrowed for this study and we develop a more generalized version of commonality factor for daily activity patterns as follows. For individual n with choice set C_n generated in the first stage, for each alternative $k \in C_n$, define

$$CF_k = \beta_{cf} \ln \sum_{j \in C_n} Sim_{k,j}^{\gamma} \tag{7.2}$$

such that

$$p(k|C_n) = \frac{exp[V_k^n + CF_k]}{\sum_{i \in C_n} exp[V_i^n + CF_i]} \tag{7.3}$$

For alternative pattern k and j, $Sim_{k,j}$ is a measurement to represent the similarity between the two patterns. We define $Sim_{k,j}$ as follows:

$$Sim_{k,j} = 1 - LevenshteinDistance(k,j)/max(|k|,|j|) \in [0,1], \tag{7.4}$$

where Levenshtein distance was introduced in Levenshtein (1966) to quantify the dis-similarity of two strings. Levenshtein distance measures the dis-similarity of two strings by the minimum number of single-character edits such as insertions, deletions or substitutions, required to

Table 7.4: Overview of commonality factor estimation results

	Base Case	CF ($\gamma = 1$)	CF ($\gamma = 2$)
Sample Size	28,656	28,656	28,656
No. Parameters	96	97	97
Init. Log Likelihood	-111,209.8	-111,209.8	-111,209.8
Final Log Likelihood	-44,968.2	-42,685.5	-43,031.0
Rho	0.596	0.616	0.613
beta for CF (t-stat)	-	-0.388 (-81.44)	-0.441 (-72.33)

change one string into another. As an example, it takes two insertions to change `hwh` into `hwhsh`. The normalized Levenshtein distance is the Levenshtein distance divided by the length of the longer string. The length of the longer string is actually the trivial distance between any two strings since the shorter string can always be achieved by deleting additional characters in the longer string and replacing the rest of the characters with those in the shorter one. The similarity measurement is then achieved by 1 - normalized Levenshtein distance. For the two daily activity sequences, `hwh` and `hwhsh`, the calculated similarity will be 0.6.

In Equation 7.2, the parameter γ may be estimated or constrained to a convenient value, often 1 and 2 (Ben-Akiva and Bierlaire, 1999). β_{cf} should be estimated from data and it is expected to be negative, indicating that heavily similar patterns have larger commonality factors and thus a smaller systematic utility with respect to similar but relatively independent patterns. The results of incorporating the commonality factor in the choice model for daily activity patterns are presented in Table 7.4. By incorporating the commonality factor with $\gamma = 1$, the model outperforms the other two, which indicates that (1) the correlations among activity patterns are not ignorable and (2) the proposed similarity measurement is able to capture the correlations.

Rule size and activity size variables

In route choice problems, C-Logit models adopting commonality factors have been criticized for a long time of being lack of theory foundation. A related concept that finds its root in the theory of aggregation of alternatives (Ben-Akiva and Lerman, 1985), the path size, is invented for route choice, in a hope to account for correlations and be consistent with the utility maximization theory (see Ben-Akiva and Bierlaire, 1999; Ramming, 2002; Hoogendoorn-Lanser et al., 2005 and Bovy et al., 2008).

Essentially, the path size variable in route choice models is a deterministic additive term in the utility function to capture the correlation due to similarity. The idea that size variables can be applied to our model arises spontaneously as it is convenient to define and quantify the similarity of two daily activity sequences. We propose the following two size variables for daily activity patterns: activity size (AS) and rule size (RS), by borrowing the so-called exponential path size formulation in Ramming (2002). For individual n, C_n is the customized choice set generated in the first stage. For each alternative $i \in C_n$, define activity/rule size variable as:

$$AS_i(RS_i) = \sum_{a \in \Omega_i} \frac{f_a}{L_i} \frac{1}{\sum_{j \in C_n} \left(\frac{L_i}{L_j}\right)^{\gamma} \delta_{aj}} \tag{7.5}$$

The notations used in the equation are summarized in Table 7.5. Generally speaking, both paths and daily activity sequences are the summation of more basic components. For paths, those components are links or legs, while for activity sequences, those components are individual activities, or rules. It is quite intuitive to understand that a path is the summation of its individual links. However, it is not quite straightforward to understand the basic components in daily activity sequences. In fact, the key lies in the representation of daily activity sequences. If daily activity sequences are represented as strings, it is reasonable to consider individual characters as the basic elements. Alternatively, if daily activity sequences are defined by a grammar thus represented as parse trees in respect to that grammar, it is then reasonable to consider individual rules that form the parse trees as the basic elements.

The hypothesis is that each component of a daily activity pattern contributes to the size variable. The contribution is proportional to the frequency of observing each component.

Table 7.5: Notations in activity/rule size variables

Symbol	Activity size (AS)	Rule size (RS)
Ω_i	set of activities in pattern i	set of grammar rules used to derive pattern i
f_a	frequency of activity a in i	frequency of applying rule a in i
L_i	total number of activities in i (with repetition)	total number of rules in i (with repetition)
δ_{aj}	if activity a is in pattern j	if rule a is used to derive pattern j

Each component in a pattern has a size 1 unless it is observed in other patterns in the choice set as well. The extent to which a component contributes to the size variable depends on the number of alternatives in which the component is observed.

In the simplest form of Equation 7.5 with $\gamma = 0$, if a component a is observed in N alternatives in C_n, then the contribution of a to size variable is $f_a/(L_i \times N)$. $\gamma > 0$ is there to penalize long patterns in favor of shorter ones. The value of γ is not a problem if the patterns in the choice set have more or less equal length. For the case of daily activity sequences, the alternatives in the choice set have various lengths and an educated guess with respect to γ should be made. We may refer to route choice literatures where path size variables are adopted for the value of γ. For example, in Ramming (2002), the author suggested that a value of γ as high as 20 is enough to reflect the large influence of the differences in length. However, since we are dealing with size variables for daily activity sequences, experiments should be done to select proper value for γ.

Size variable of an alternative enters the utility function in its logarithm form and should be multiplied by a weighting parameter β_{rs} or β_{as}. According to Ben-Akiva and Lerman (1985) and Hoogendoorn-Lanser et al. (2005), the value of the beta for a size variable should be positive. Moreover, if it equals 1, the size variable accounts for only the positive correlation because of overlapping; otherwise, it may incorporate some unobserved effects of the similar

alternatives. As a result, the weighting parameter should be estimated from the data to capture how exactly the size variable can approximate the correlation brought by overlapping between activity sequences.

Table 7.6 summarizes the estimation results by using the rule size variable. Overall, the inclusion of rule size variable improves the likelihood. The models with $\gamma = 20 \sim 40$ show the best Logit estimation performance with a relative 10 percent improvement from the base case. Moreover, these models have a β_{rs} close to 1, indicating a significant effect of the defined rule size variable in terms of approximating the overlapping between activity sequences. It is noted that by taking different values for γ, the performance of the models has considerable differences. Such differences indicate that γ plays an essential role in characterizing the correlations and decreasing the overlapping among alternatives. A high value of γ implies that when comparing short and long patterns, the shorter patterns in terms of number of rules used to construct the parse tree are considered to be more elemental, which is correct as the rules used to construct shorter patterns are almost inevitably used to construct longer patterns. As a result, applying high value of γ improves the estimation results.

Table 7.6: Overview of rule size estimation results

Model	No. Parameter	Beta for RS (t-stat)	Log Likelihood	Rho	Relative improvement
Base case	96	-	-44,968.2	0.596	reference
RS(gamma=0)	97	-0.079 (-9.68)	-44,903.1	0.596	0.1%
RS(gamma=1)	97	-0.007 (-0.80)	-44,967.8	0.596	0.0%
RS(gamma=5)	97	0.274 (30.53)	-44,496.1	0.600	1.0%
RS(gamma=10)	97	0.726 (81.36)	-42,409.0	0.619	5.7%
RS(gamma=15)	97	0.995 (100.24)	-40,713.4	0.634	9.5%
RS(gamma=20)	97	1.070 (97.70)	-40,164.0	0.639	10.7%
RS(gamma=30)	97	1.090 (89.06)	-40,050.6	0.640	10.9%
RS(gamma=40)	97	1.080 (82.11)	-40,116.4	0.639	10.8%

Alternatively, we can use the definition of activity size and the estimation results are summarized in Table 7.7. As shown in the table, the model with best performance is the one with $\gamma = 10$. Compared to the results we get from models with rule size variable, models with activity size variable perform much better. The better performance may be

Table 7.7: Overview of activity size estimation results

Model	No. Parameter	Beta for AS (t-stat)	Log Likelihood	Rho	Relative improvement
Base case	96	-	-44,968.2	0.596	reference
AS(gamma=0)	97	-0.074 (-7.59)	-44,931.3	0.596	0.1%
AS(gamma=1)	97	0.143 (13.75)	-44,857.2	0.597	0.2%
AS(gamma=3)	97	0.643 (49.62)	-43,055.0	0.613	4.3%
AS(gamma=5)	97	1.170 (92.09)	-39,460.9	0.645	12.2%
AS(gamma=10)	97	1.410 (71.82)	-33,634.5	0.698	25.2%
AS(gamma=15)	97	0.890 (57.96)	-34,280.4	0.692	23.8%
AS(gamma=20)	97	0.621 (55.77)	-34,984.0	0.685	22.2%

attributed to the fact that the activity size variable is able to incorporate correlations due to additional unobserved effects other than overlapping (since its best model, AS10, has a larger estimated beta for the size variable, in comparison with the best model with rule size variable). However, it is worth noting that activity size variables only rely on the characters in the activity strings while rule size variables rely on the parse trees of activity strings, which are generated in respect of a particular grammar. In other words, the rule size variables are grammar-specific. Thus, the results from models using rule size variable may improve if better grammars for synthesizing daily activity sequences are proposed.

7.5 Application in the Pre-day Modeling Framework

As reviewed in Chapter 2, certain econometric-based modeling frameworks (such as the ones with individual or coordinated daily activity patterns) define daily activity patterns as the abstraction of detailed activity plans and schedules on daily basis, and model them at the top level in a hierarchical modeling structure. Starting from Chapter 6, we define daily activity patterns as activity sequences and there are several reviewed frameworks that adopt this definition (such as the Portland Metro Model, Oregon and Ohio Model). However, few of them explicitly consider the choice set generation process, not to mention generating a customized choice set for each individual.

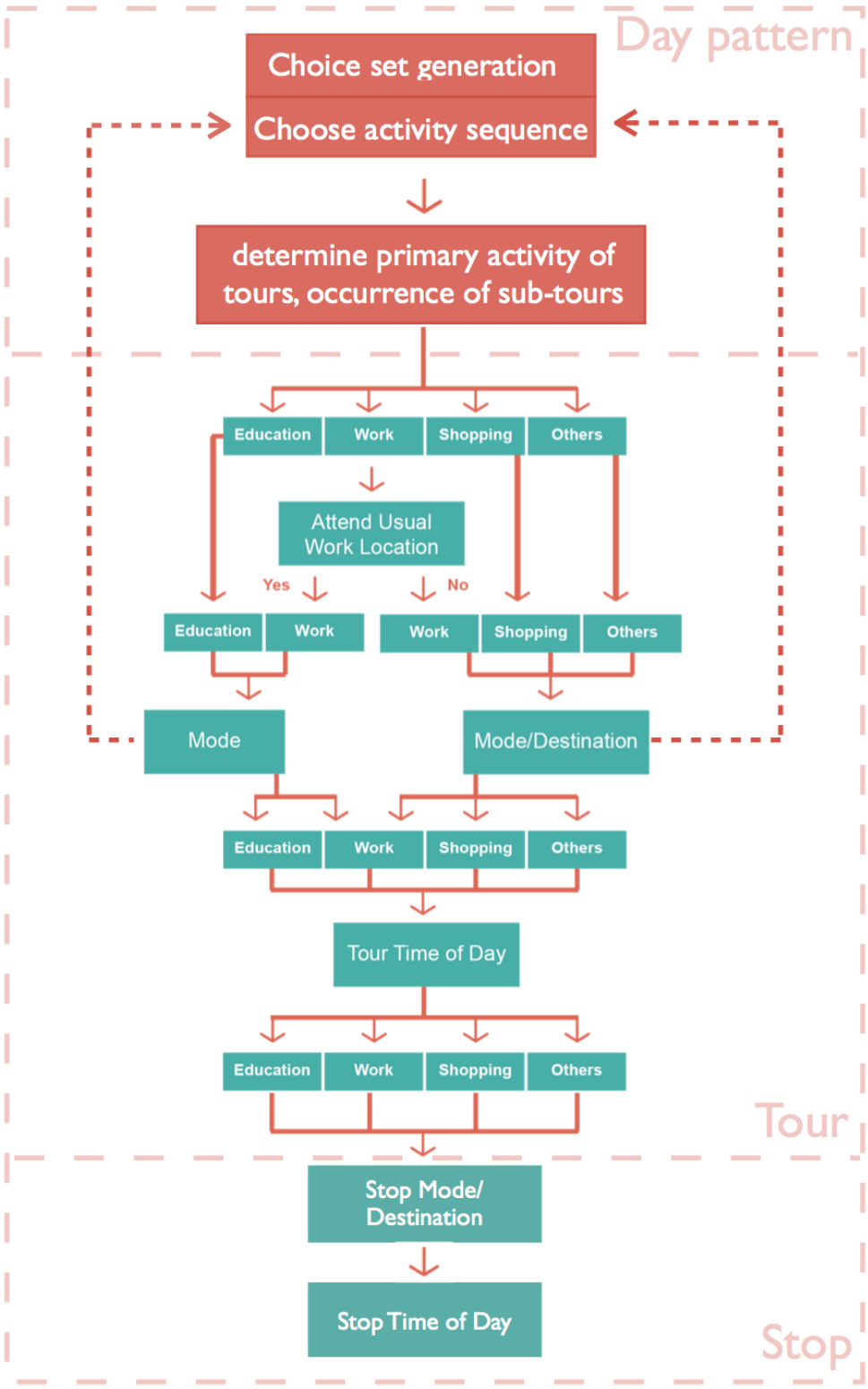


Figure 7.5: A revised process flow for the pre-day model to incorporate the two-stage model

We have shown in this chapter that by taking advantage of the probabilistic grammar-based representation of daily activity sequences, a two-stage choice model can be developed with a customized choice set for each individual. As it unfolds, it is now reasonable to consider the application of this two-stage model in an activity-based modeling framework, such as the pre-day model developed in Chapter 5. Compared to the daily activity patterns in the pre-day model (defined as the occurrence of tours and intermediate stops of different purposes), activity sequences provide more information, such as activity sequence, sequence of tours, intermediate stops and appearance of work-based sub-tours. As a result of the more detailed representation of daily activity patterns, the process flow shown in Figure 5.3 can be simplified. Figure 7.5 provides the revised process flow. After daily activity sequence is determined by the two-stage model (choice set generation and choice making), the primary activity of each tour and occurrence of sub-tours can be determined: the primary activity of each tour is determined by the priority of activity (as introduced in Chapter 4); the occurrence of work-based sub-tours is determined by observing the subsequence in a tour that starts and ends with work activity. Compared to the original process flow, there are no work-based sub-tour generation model, intermediate stop generation model and exact number of tours model as those facets have been embedded in the outcome of the two-stage model. The new process flow uses the same accessibility measurements in the two-stage model as they have been used in the day pattern model and exact number of tours model of the original framework.

In order to be incorporated into the revised pre-day framework, the two-stage model needs to be able to replicate the activity participation behavior in the facets it covers. Specifically, Figure 7.6 summarizes the results of the base year validation for the two-stage model (with activity size variable AS10, where $\gamma = 10$). Overall, the model is able to replicate the base year statistics in terms of tour count for the tour purposes considered in the pre-day model, number of tours for each individual and number of stops in each tour, with minor discrepancy that can be ignored. Compared to the original day pattern model, the two-stage one performs just as good as it if not better in terms of replicating the base year behavior. Talking about the influence to the rest of the pre-day framework, the two-stage model is able to embed more information by modeling activity sequences (rather than occurrence of tours

Tour Count	HITS2008	Simulated
Work Tour	1,129,207	1,131,785
Edu Tour	643,245	642,478
Shopping Tour	122,509	120,971
Other Tour	248,716	245,266
Sub-Tour	9,993	9,544

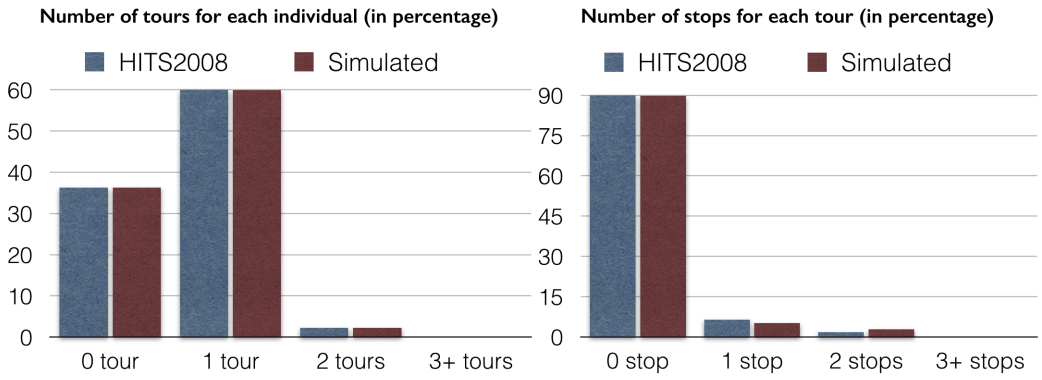


Figure 7.6: Base year validation for the two-stage model

and intermediate stops of different purposes) as daily activity patterns and thus simplify the modeling framework. Moreover, embedding more facets in daily activity patterns usually suggests a larger choice set (as in the case of the Oregon State Model, a choice set of over 3,000 alternative activity sequences is pre-defined). However, the two-stage model solves the issue by taking advantage of the grammar-based representation of activity sequences and generating a customized and much smaller choice set for each individual.

7.6 Summary

This study is an extension of Chapter 6. It explores the application of the grammar-based representation of daily activity patterns (defined as activity sequences) in certain hierarchical econometric-based activity-based modeling frameworks. In those activity-based modeling frameworks, daily activity patterns are defined as the abstraction of activity participation on daily basis and are modeled at the top of the hierarchy. Nevertheless, the definition of daily activity patterns is not self-evident and the modeling of them largely depends on how they are defined.

Daily activity patterns are defined as activity sequences in this study. Compared with the definition applied in the pre-day model, activity sequences embed more information and may simplify the activity-based modeling frameworks. However, the potential advantage comes with the cost of the explosion of alternatives. The study solves the problem of (1) how to represent the vast number of alternative activity sequences and (2) how to reproduce the heterogeneity among the population in the process of developing a choice set for daily activity sequences. The first problem has been addressed in Chapter 6. The answer to the second problem has also found its root in the (probabilistic) grammar-based representation and is elaborated in this chapter. By taking advantage of the representation, the choice set customized for each individual has satisfying coverage with relatively small size.

The choice process is carried out with Logit model. Moreover, we are able to capture the correlations among alternatives through several proposed additive terms to the Logit model. Commonality factor and two size variables are created in sprite of their applications in route choice models.

Finally, the two-stage model is incorporated into the pre-day activity-based modeling framework in Chapter 5. Compared to the original day pattern model in the framework, the two-stage model simplifies the modeling framework as well as being capable of replicating the activity participation behavior in the facets it covers. The application of the two-stage model in the pre-day modeling framework is an example to show how new methodologies and models can be integrated into the benchmark.

In the future, as mentioned in Chapter 6, more advanced grammars can be developed to replace the current definition of daily activity sequences, which may be more detailed and realistic, embedding more information in the sequence itself. The consecutive studies presented in Chapter 6 and Chapter 7 have provided a clear flow on how to apply this new definition of daily activity patterns all the way into an operational activity-based modeling framework, which may be the largest contribution of these two chapters to the application and innovation of activity-based travel demand models.

Chapter 8

Conclusions

8.1 Concluding Remarks

Activity-based travel demand models have been receiving more and more attention in the research community as well as in practice. With the transformation from supply-oriented planning to inventory-based planning continuously taking place worldwide, the focus of urban transportation planning in developed regions has shifted from massive development of transportation infrastructure with intensive capital investment to assessment of transportation policies, environmental impacts and management of travel demand. To better serve the transformation, the weakness and inadequacy of the traditional four-step method have been well observed and the need for commitment to a more disaggregate and policy-sensitive modeling approach is therefore appreciated. Activity-based models stem from the principle that the need for travel is derived from the need for activity participation. They are built based on the seminal works and prototypes emerged in the research community. After the continuous development of theoretical foundations and empirical implementations for over two decades, several groups of activity-based models, such as rule based ones and econometric-based ones, have been in the place to be put into operation. Nevertheless, the methodologies and features of activity-based models will keep evolving in attempts to be consistent with emerging theoretical achievements and tailored to specific regions with distinct policy focus and practical considerations.

In view of the trend, this thesis has developed and implemented a comprehensive econometric-based activity-based modeling framework with individual daily activity patterns for Singapore. Innovations in terms of data, modeling approach, framework implementation and alternative

model for daily activity patterns are highlighted.

Chapter 3 and 4 focus on the two essential data sources required for the development of activity-based models.

Chapter 3 addresses the issue of providing better travel time input for the development of time of day choice models that require travel time estimates at a finer time resolution. While travel time acquired from network skims for a small number of time windows is inadequate to model travelers' response to congestion mitigation strategies, we are able to fuse travel time collected via multiple sources and develop regression models to relate travel time collected from taxi GPS data to the network travel time and compare the results to a similar model estimated with household travel survey data. The rationale behind this procedure is to develop a formula that allows the calculation of travel time for any origin-destination pair and for any time of the day, given the network travel time for three time periods (AM peak, PM peak, and off-peak). As shown in the results, although there are significant differences in the estimated coefficients, which do not vary across time of day, the two data sources exhibit comparable profiles for time-of-day variation of speed up to certain scales. Similar regression models are developed for public transportation modes where household travel survey data and smart card data are utilized. Travel time generated in this study serves as one of the prerequisites for developing activity-based travel demand models in Singapore.

Chapter 4 investigates the compatibility of existing trip-based household travel surveys conducted in Singapore with the development of activity-based models. Sufficient efforts, such as data checks, trip-to-tour conversion, and work-based sub-tour detection are carried out before reaching the positive conclusion that the same household travel surveys once used for the development of four-step models are sufficient for encoding tours with a trip-to-tour conversion process and detection of work-based sub-tours. It is an immediate suggestion that the processed household travel survey data can be used to support the implementation of activity-based models in Singapore. However, the trip-based surveys also impose limitations that should be taken into consideration when developing the activity-based modeling framework for Singapore, such as data inconsistency and ignorance of household interactions in designing the surveys.

Chapter 5 represents the major efforts in developing an activity-based travel demand modeling framework for Singapore and an integrated mid-term simulator SimMobilityMT for the modeling of activity participation and travel decisions occurring on daily basis. The pre-day activity-based model behind the simulator is formulated through a system of interconnected discrete choice models representing choices at distinct dimensions of daily activity schedule. This study presents the concept, design and implementation of the pre-day model and its simulator. Key features in terms of the implementation of the pre-day simulator, such as modularization, parallelization and multiple running modes are highlighted. With the pre-day model estimated and the simulator implemented, model calibration/validation process is carried out against the activity participation and travel behavior of the base year. It is shown that the pre-day simulator is capable of correctly replicating the behavior of the base year. As a benchmark, the pre-day model has been fully implemented and operational. Nevertheless, improvements can still be made in subsequent phases of development, such as application in planning scenarios to assess the performance of various TDM strategies and integration with long-term models for the modeling of land use policies.

Chapter 6 and 7 explore a possible alternative daily activity pattern model that can be used to replace the one implemented in the pre-day model introduced in Chapter 5.

In Chapter 6, similarities between daily activity pattern — which is defined as activity sequence — and language are first explored. Context-free grammars are then developed to parse and generate daily activity patterns. To replicate people's heterogeneity in selecting daily activity patterns, we introduce probabilistic context-free grammar and propose several formulations to estimate the probability of a context-free grammar with daily activity patterns observed in a household travel survey. With experiments, the estimated probabilistic context-free grammar is able to reproduce the observed pattern distribution in household travel survey with satisfactory precision. This chapter intends to advance studies on human daily activity patterns by providing new perspective and methodology in the modeling and learning of daily activity patterns. Specifically, as the first attempt to establish the connection between daily activity patterns and formal languages, this study borrows the concepts in probabilistic grammar and applies them in modeling and learning activity participation behavior. While there are many promising directions initiated by the study, such as grammar induction

for daily activity patterns, the most relevant one in connection with the development of activity-based models is that the proposed methodology sheds light on the issue of generating customized choice sets for daily activity pattern models in certain activity-based modeling frameworks, such as the pre-day model developed in Chapter 5.

Chapter 7 takes advantage of the grammar-based representation of daily activity patterns within the framework of discrete choice modeling and proposes a two-stage choice model for daily activity patterns, where a choice set is customized for each individual based on socio-demographic variables, followed by a Multinomial Logit model to determine the chosen pattern. Several commonality factor and size variable formulations are proposed to account for the correlations among alternatives in the customized choice set. The two-stage model is implemented in the pre-day model with a slightly modified modeling framework (by taking advantage of the modularization feature of the pre-day simulator), where models at daily activity pattern level are replaced by the two-stage model. Besides, as daily activity patterns are modeled as activity sequences and more information about trip chaining is thus inherited in the outcome of the two-stage model, the pre-day framework can be simplified by removing work-based sub-tour generation and intermediate stop generation process. As the results suggest, the two-stage model is able to replicate the base year activity participation behavior at daily activity pattern level.

8.2 Future Works

In general, the proposed activity-based modeling framework has been developed and implemented successfully. The thesis has introduced several topics that are directly connected to the model development process. However, in case of the operation of activity-based models, improvement and adjustment that aim to suit specific application scenarios, to be consistent with more advanced theories and modeling approaches, and to better utilize new data, are always appreciated. Several promising directions of future works are provided as follows.

First, although the development and implementation of the pre-day activity-based simulator have come to an end, it only represents the demand-side tool for demand microsimulation in

SimMobilityMT, while the supply-side assignment tool is still in development. SimMobilityMT is fully developed only when the supply simulator is online and integrated with the demand simulator. Then, a model calibration and validation process should be carried out to asses the performance of SimMobilityMT in forecasting scenarios represented in future household travel surveys (such as HITS2012). To further extend the capability of the daily activity-travel simulator, SimMobilityMT can be integrated into an integrated land use and transportation demand model.

Second, a fully developed SimMobilityMT is a promising tool to asses planning scenarios involving various TDM strategies, such as congestion pricing, discount fare for transit travel and vehicle emission-based registration fee and tax.

Third, for the purpose of model maintenance and enhancement, refinements to individual models are often desired. The modularization feature of the pre-day simulator makes it simpler for modelers to update or adjust individual models due to new data, emerging theories and customized needs in a variety of policy evaluations. The thesis has demonstrated that the models at day pattern level of the pre-day modeling framework can be replaced by the two-stage model introduced in Chapter 6 and 7. A few more examples may include the incorporation of intra-household interactions, hazard-based duration for the primary activity of tours and more realistic choice set generation procedures for location choice, etc.

Bibliography

Abdulazim, T., Abdelgawad, H., Habib, K. M. N., Abdulhai, B., 2013. Using smartphones and sensor technologies to automate collection of travel data. Transportation Research Record: Journal of the Transportation Research Board 2383 (1), 44–52.

Abou-Zeid, M., Rossi, T., Gardner, B., 2006. Modeling time-of-day choice in context of tour- and activity-based models. Transportation Research Record: Journal of the Transportation Research Board 1981 (1), 42–49.

Adler, T., Ben-Akiva, M., 1979. A theoretical and empirical model of trip chaining behavior. Transportation Research Part B: Methodological 13 (3), 243–257.

Algers, S., Daly, A., Kjellman, P., Widlert, S., 1995. Stockholm Model System (SIMS): Application. Volume 2: Modelling transport systems. In: Proceedings of the 7th World Conference on Transport Research, Sydney, Australia. pp. 345–361.

Allahviranloo, M., Recker, W., 2013. Daily activity pattern recognition by using support vector machines with multiple classes. Transportation Research Part B: Methodological 58, 16–43.

Anderson, R., Jiang, Z., 2013. Agency experience using activity-based models: Experience in the State of Ohio. TMIP Webinar Series, available at http://media.tmiponline.org/webinars/2013/TMIP_ABM_Webinars/ODOT/MORPC_and_ODOT_ABM_webinar.pdf, accessed on June 9, 2015.

Anggraini, R., Arentze, T., Timmermans, H., 2007. Refining ALBATROSS: Modeling household activity generation and allocation decisions using decision tree induction. In: Proceedings of the 11th World Conference on Transportation Research, Berkeley, USA. p. 21p.

Bibliography

Arentze, T., Hofman, F., Joh, C., Timmermans, H., 1999. The development of ALBATROSS: Some key issues. In: Brilon, W., Huber, F., Schreckenberg, M., Wallentowitz, H. (Eds.), Traffic and Mobility. Berlin: Springer, pp. 57–72.

Arentze, T., Timmermans, H., 2000. ALBATROSS: A learning based transportation oriented simulation system. Eindhoven: EIRASS.

Arentze, T., Timmermans, H., 2004. A learning-based transportation oriented simulation system. Transportation Research Part B: Methodological 38 (7), 613–633.

Arentze, T., Timmermans, H., 2007. Robust approach to modeling choice of locations in daily activity sequences. Transportation Research Record: Journal of the Transportation Research Board 2003 (1), 59–63.

Auld, J., Mohammadian, A., 2009. Framework for the development of the Agent-based Dynamic Activity Planning and Travel Scheduling (ADAPTS) model. Transportation Letters 1 (3), 245–255.

Auld, J., Mohammadian, A., 2012. Activity planning processes in the Agent-based Dynamic Activity Planning and Travel Scheduling (ADAPTS) model. Transportation Research Part A: Policy and Practice 46 (8), 1386–1403.

Axhausen, K. W., Zimmermann, A., Schönfelder, S., Rindsfüser, G., Haupt, T., 2002. Observing the rhythms of daily life: A six-week travel diary. Transportation 29 (2), 95–124.

Balakrishna, R., 2006. Off-line calibration of dynamic traffic assignment models. Ph.D. thesis, Massachusetts Institute of Technology.

Balmer, M., 2007. Travel demand modeling for multi-agent transport simulations: Algorithms and systems. Ph.D. thesis, ETH Zurich.

Balmer, M., Axhausen, K. W., Nagel, K., 2006. Agent-based demand-modeling framework for large-scale microsimulations. Transportation Research Record: Journal of the Transportation Research Board 1985 (1), 125–134.

Barceló, J., Ferrer, J., Grau, R., 1994. AIMSUN2 and the GETRAM simulation environment. In: Proceedings of the 13th EURO Conference, Glasgow, United Kingdom.

Bibliography

Başar, G., Bhat, C. R., 2004. A parameterized consideration set model for airport choice: An application to the San Francisco Bay Area. Transportation Research Part B: Methodological 38 (10), 889–904.

Beckx, C., Panis, L. I., Arentze, T., Janssens, D., Torfs, R., Broekx, S., Wets, G., 2009. A dynamic activity-based population modelling approach to evaluate exposure to air pollution: Methods and application to a Dutch urban area. Environmental Impact Assessment Review 29 (3), 179–185.

Bellemans, T., Kochan, B., Janssens, D., Wets, G., Arentze, T., Timmermans, H., 2010. Implementation framework and development trajectory of FEATHERS activity-based simulation platform. Transportation Research Record: Journal of the Transportation Research Board 2175 (1), 111–119.

Bellemans, T., Kochan, B., Janssens, D., Wets, G., Timmermans, H., 2008. Field evaluation of personal digital assistant enabled by global positioning system: Impact on quality of activity and diary data. Transportation Research Record: Journal of the Transportation Research Board 2049 (1), 136–143.

Bellman, R., 1961. Adaptive control processes: A guided tour. Princeton: Princeton University Press.

Ben-Akiva, M., Abou-Zeid, M., 2013. Methodological issues in modelling time-of-travel preferences. Transportmetrica A: Transport Science 9 (9), 846–859.

Ben-Akiva, M., Bierlaire, M., 1999. Discrete choice methods and their applications to short term travel decisions. In: Hall, R. (Ed.), Handbook of Transportation Science. Vol. 23 of International Series in Operations Research and Management Science. New York: Springer, pp. 5–33.

Ben-Akiva, M., Bierlaire, M., Koutsopoulos, H., Mishalani, R., 1998. DynaMIT: A simulation-based system for traffic prediction. In: Proceedings of the DACCORS short term forecasting workshop, The Netherlands.

Ben-Akiva, M., Bowman, J. L., 1995. Activity-based disaggregate travel demand model system with daily activity schedules. In: Proceedings of the EIRASS Conference on

Activity-Based Approaches: Activity Scheduling and the Analysis of Activity Patterns, Eindhoven, The Netherlands.

Ben-Akiva, M., Bowman, J. L., 1998. Activity-based travel demand model systems. In: Marcotte, P., Nguyen, S. (Eds.), Equilibrium and Advanced Transportation Modeling. Boston: Kluwer Academic Publishers, pp. 27–46.

Ben-Akiva, M., Bowman, J. L., Gopinath, D., 1996. Travel demand model system for the information era. Transportation 23 (3), 241–266.

Ben-Akiva, M., Lerman, S. R., 1985. Discrete choice analysis: Theory and application to travel demand. Cambridge: MIT Press.

Bhat, C. R., Guo, J. Y., Srinivasan, S., Sivakumar, A., 2004. Comprehensive econometric microsimulator of daily activity-travel patterns. Transportation Research Record: Journal of the Transportation Research Board 1894 (1), 57–66.

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

Booth, T., Thompson, R., 1973. Applying probability measures to abstract languages. IEEE Transactions on Computers C-22 (5), 442–450.

Bovy, P. H. L., 2009. On modelling route choice sets in transportation networks: A synthesis. Transport Reviews 29 (1), 43–68.

Bovy, P. H. L., Bekhor, S., Prato, C. G., 2008. The factor of revisited path size: Alternative derivation. Transportation Research Record: Journal of the Transportation Research Board 2076 (1), 132–140.

Bovy, P. H. L., Fiorenzo-Catalano, S., 2007. Stochastic route choice set generation: Behavioral and probabilistic foundations. Transportmetrica 3 (3), 173–189.

Bowman, J. L., 1998. The day activity schedule approach to travel demand analysis. Ph.D. thesis, Massachusetts Institute of Technology.

Bowman, J. L., 2009. Historical development of activity-based model theory and practice. Traffic Engineering and Control 50 (2), 314–318.

Bibliography

Bowman, J. L., Ben-Akiva, M., 2001. Activity-based disaggregate travel demand model system with activity schedules. Transportation Research Part A: Policy and Practice 35 (1), 1–28.

Bowman, J. L., Bradley, M., Shiftan, Y., Lawton, T. K., Ben-Akiva, M., 1998. Demonstration of an activity based model system for Portland. In: Proceedings of the 8th World Conference on Transport Research, Antwerp, Belgium.

Bradley, M., Bowman, J. L., Griesenbeck, B., 2007. Development and application of the SACSIM activity-based model system. In: Proceedings of the 11th World Conference on Transport Research, Berkeley, USA.

Bradley, M., Bowman, J. L., Griesenbeck, B., 2010. SACSIM: An applied activity-based model system with fine-level spatial and temporal resolution. Journal of Choice Modelling 3 (1), 5–31.

Bradley, M., Outwinter, M., Jonnalagadda, N., Ruiter, E., 2001. Estimation of an activity-based micro simulation model for San Francisco. In: Proceedings of the 80th Annual Meeting of the Transportation Research Board, Washington, D.C., USA.

Bradley, M., Portland Metro, Bowman, J. L., Cambridge Systematics, Inc., 1998. A system of activity-based models for Portland, Oregon. Tech. Rep. USDOT Report DOT-T-99-02, Prepared for the Federal Highway Administration Travel Model Improvement Program of the USDOT and EPA, Washington, D.C.

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

Bricka, S., Bhat, C. R., 2006. Using GPS data to inform travel survey methods. In: Proceedings of Innovations in Travel Demand Modeling Conference, Transportation Research Board, Austin, USA. pp. 89–93.

California EPA Air Resources Board, 2010. Regional greenhouse gas emission reduction targets for automobiles and light trucks pursuant to Senate Bill 375. available at http://www.arb.ca.gov/cc/sb375/sb375.htm, accessed on June 9, 2015.

Bibliography

Cambridge Systematics, Inc., 2002. San Francisco travel model development: Final report. Tech. rep., Prepared for the San Francisco County Transportation Authority, available at http://www.sfcta.org/images/stories/Planning/DataMart/SFModelDocumentation.zip, accessed on June 9, 2015.

Cambridge Systematics, Inc., 2010. DRCOG model design plan. Tech. rep., Prepared for the Denver Regional Council of Governments, available at https://drcog.org/services-and-resources/data-maps-and-modeling/travel-modeling/focus-travel-model, accessed on June 9, 2015.

Cantillo, V., de Dios Ortúzar, J., 2005. A semi-compensatory discrete choice model with explicit attribute thresholds of perception. Transportation Research Part B: Methodological 39 (7), 641–657.

Cascetta, E., Biggiero, L., 1997. Integrated models for simulating the Italian passenger transport system. In: Transportation Systems 1997: A proceedings volume from the 8th IFAC/IFIP/IFORS Symposium, Chania, Greece. Vol. 1.

Cascetta, E., Nuzzolo, A., Russo, F., Vitetta, A., 1996. A modified logit route choice model overcoming path overlapping problems: Specification and some calibration results for interurban networks. In: Proceedings of the 13th International Symposium on Transportation and Traffic Theory, Pergamon, France. pp. 697–711.

Castiglione, J., Bradley, M., Gliebe, J., 2015. Second Strategic Highway Research Program (SHRP 2) Report S2-C46-RR-1: Activity-Based Travel Demand Models: A Primer. Transportation Research Board.

Castiglione, J., Hiatt, R., Chang, T., Charlton, B., 2006. Application of travel demand microsimulation model for equity analysis. Transportation Research Record: Journal of the Transportation Research Board 1977 (1), 35–42.

Castro, M., Martínez, F., Munizaga, M., 2013. Estimation of a constrained multinomial logit model. Transportation 40 (3), 563–581.

Chapin, F., 1974. Human Activity Patterns in the City: Things People Do in Time and in Space. Volume 13 of Wiley Series in Urban Research. New York: John Wiley & Sons.

Chi, Z., Geman, S., 1998. Estimation of probabilistic context-free grammars. Computational Linguistics 24 (2), 299–305.

Chiao, K.-A., Mohseni, A., Bhowmick, S., 2006. Lessons learned from the implementation of New York activity-based travel model. In: Proceedings of the Innovations in Travel Demand Modeling Conference, Austin, USA. pp. 173–176.

Childress, S., Nichols, B., Charlton, B., Coe, S., 2015. Using an activity-based model to explore possible impacts of automated vehicles. In: Proceedings of the 94th Annual Meeting of the Transportation Research Board, Washington, D.C., USA.

Chomsky, N., 1956. Three models for the description of language. IRE Transactions on Information Theory 2 (3), 113–124.

Chomsky, N., 1957. Syntactic Structures. Berlin: de Gruyter Mouton.

Chomsky, N., 1959. On certain formal properties of grammars. Information and Control 2 (2), 137–167.

Chomsky, N., Miller, G. A., 1958. Finite state languages. Information and Control 1 (2), 91–112.

City of Chicago, 1959. Chicago area transportation study: Final report. Tech. rep., Prepared for the Bureau of Public Roads, U.S. Department of Commerce, available at https://datahub.cmap.illinois.gov/dataset/historic-cats-plans/resource/137e9c4f-a2ac-4fe7-a1a9-7463feefb0d9, accessed on June 9, 2015.

Cullen, I., Godson, V., 1975. Urban networks: The structure of activity patterns. Progress in Planning 4, Part 1, 1–96.

Daly, A., 1982. Estimating choice models containing attraction variables. Transportation Research Part B: Methodological 16 (1), 5–15.

Davidson, W., Vovsha, P., Freedman, J., Donnelly, R., 2010. CT-RAMP family of activity-based models. In: Proceedings of the 33rd Australasian Transport Research Forum (ATRF), Canberra, Australia. Vol. 29.

Bibliography

DKS Associate, Bradley Research and Consulting, Transportation Systems and Design Sciences, 2012. Sacramento activity-based travel simulation model (SACSIM11): Model reference report. Tech. rep., Prepared for Sacramento Area Council of Governments, available at http://sacog.org/mtpscs/files/MTP-SCS/appendices/C-4%20SACSIM%20Documentation.pdf, accessed on June 9, 2015.

Doherty, S. T., Nemeth, E., Roorda, M., Miller, E. J., 2004. Computerized household activity-scheduling survey for Toronto Canada area: Design and assessment. Transportation Research Record: Journal of the Transportation Research Board 1894 (1), 140–149.

Dubey, A., Jalote, P., Aggarwal, S., 2008. Learning context-free grammar rules from a set of program. Software, IET 2 (3), 223–240.

Dyrka, W., Nebel, J. C., 2009. A stochastic context free grammar based framework for analysis of protein sequences. BMC Bioinformatics 10 (1), 323.

Ettema, D., Borgers, A., Timmermans, H., 1993. Simulation model of activity scheduling behavior. Transportation Research Record: Journal of the Transportation Research Board 1413 (1), 1–11.

Ettema, D., Borgers, A., Timmermans, H., 1996. Simulation model of activity scheduling heuristics (SMASH): Some simulations. Transportation Research Record: Journal of the Transportation Research Board 1551 (1), 88–94.

Feigenbaum, E. A., Buchanan, B. G., Lederberg, J., 1970. On generality and problem solving: A case study using the DENDRAL program. Tech. rep., Stanford University, USA, available at http://profiles.nlm.nih.gov/ps/access/BBABKV.pdf, accessed on June 9, 2015.

Freedman, J., Castiglione, J., Charlton, B., 2006. Analysis of New Starts Project by using tour-based model of San Francisco, California. Transportation Research Record: Journal of the Transportation Research Board 1981 (1), 24–33.

Fried, M., Havens, J., Thall, M., 1977. Travel behavior – A synthesized theory. Tech. rep., Prepared for the National Cooperative Highway Research Program, Transportation Research Board, National Research Council.

Fu, L., Rilett, L. R., 2000. Estimation of timeâĂŘdependent, stochastic route travel times using artificial neural networks. Transportation Planning and Technology 24 (1), 25–48.

Gärling, T., Brännäs, K., Garvill, J., Golledge, R. G., Gopal, S., Holm, E., Lindberg, E., 1989. Household activity scheduling. In: Transport Policy, Management and Technology Towards 2001. Vol. IV. Ventura, CA: Western Periodicals, pp. 235–248.

Gärling, T., Kwan, M.-P., Golledge, R. G., 1994. Computational-process modelling of household activity scheduling. Transportation Research Part B: Methodological 28 (5), 355–364.

Giaimo, G., Anderson, R., Wargelin, L., 2009. Large-scale deployment of a GPS-based household travel survey in Cincinnati. In: Proceedings of the 12th Transportation Research Board National Transportation Planning Applications Conference, Houston, USA.

Gliebe, J., 2006. Linking tour-based models with microsimulation: The TRANSIMS experience in Portland. In: Proceedings of Innovations in Travel Demand Modeling Conference, Transportation Research Board, Austin, USA.

Golledge, R. G., Kwan, M.-P., Gärling, T., 1994. Computational process modeling of household travel decisions using a geographical information system. Papers in Regional Science 73 (2), 99–117.

Golob, T., Kim, S., Ren, W., 1996. How households use different types of vehicles: A structural driver allocation and usage model. Transportation Research Part A: Policy and Practice 30 (2), 103–118.

González, M. C., Hidalgo, C. A., Barabási, A.-L., 2008. Understanding individual human mobility patterns. Nature 453 (7196), 779–782.

Goulias, K. G., Bhat, C. R., Pendyala, R. M., Chen, Y., Paleti, R., Konduri, K. C., Lei, T., Tang, D., Youn, S., Huang, G., et al., 2012. Simulator of activities, greenhouse emissions, networks, and travel (SimAGENT) in Southern California. In: Proceedings of the 91st Annual Meeting of the Transportation Research Board, Washington, D.C., USA.

Hägerstrand, T., 1970. What about people in regional science? Papers of the Regional Science Association 24 (1), 6–21.

Hanson, S., 1982. The determinants of daily travel-activity patterns: Relative location and sociodemographic factors. Urban Geography 3 (3), 179–202.

Hausman, J., McFadden, D., 1984. Specification tests for the multinomial logit model. Econometrica 52 (5), 1219–1240.

Hensher, D., Button, K., 2000. Handbook of Transport Modeling. Oxford: Pergamon.

Hess, S., Daly, A., Rohr, C., Hyman, G., 2007. On the development of time period and mode choice models for use in large scale modelling forecasting systems. Transportation Research Part A: Policy and Practice 41 (9), 802–826.

Hood, J., Sall, E., Charlton, B., 2011. A GPS-based bicycle route choice model for San Francisco, California. Transportation Letters 3 (1), 63–75.

Hoogendoorn-Lanser, S., van Nes, R., Bovy, P. H. L., 2005. Path size and overlap in multi-modal transport networks: A new interpretation. In: Proceedings of the 16th International Symposium on Transportation and Traffic Theory, College Park, USA.

Hopcroft, J. E., Motwani, R., Ullman, J. D., 2006. Introduction to Automata Theory, Languages, and Computation, 3rd Edition. Boston: Addison-Wesley Longman.

Horowitz, J., 1980. A utility maximizing model of the demand for multi-destination non-work travel. Transportation Research Part B: Methodological 14 (4), 369–386.

Hunt, K., Petersen, E., 2004. The role of gender, work status and income in auto allocation decisions. In: Proceedings of the 83rd Annual Meeting of the Transportation Research Board, Washington, D.C., USA.

Iacono, M., Krizek, K., El-Geneidy, A., 2008. Access to destinations: How close is close enough? Estimating accurate distance decay functions for multiple modes and different purposes. Tech. rep., Prepared for the Minnesota Department of Transportation, available at http://www.lrrb.org/PDF/200811.pdf, accessed on June 9, 2015.

Isbell, N. A., Goulias, K. G., 2014. Modeling second-by-second traffic emissions in a mega-region. In: Proceedings of the 83rd Annual Meeting of the Transportation Research Board, Washington, D.C., USA.

Javed, F., Bryant, B. R., Črepinšek, M., Mernik, M., Sprague, A., 2004. Context-free grammar induction using genetic programming. In: Proceedings of the 42nd Annual Southeast Regional Conference. ACM-SE 42. Association for Computing Machinery, New York, USA, pp. 404–405.

Jiang, S., Ferreira, J., González, M. C., 2012. Clustering daily patterns of human activities in the city. Data Mining and Knowledge Discovery 25 (3), 478–510.

Joh, C.-H., Arentze, T., Hofman, F., Timmermans, H., 2002. Activity pattern similarity: A multidimensional sequence alignment method. Transportation Research Part B: Methodological 36 (5), 385–403.

Jones, P. M., 1977. New Approaches to Understanding Travel Behavior: The Human Activity Approach. Oxford University.

Jones, P. M., Dix, M. C., Clarke, M. I., Heggie, I. G., 1983. Understanding Travel Behaviour. Aldershot: Gower.

Kaplan, S., Bekhor, S., Shiftan, Y., 2009. Two-stage model for jointly revealing determinants of noncompensatory conjunctive choice set formation and compensatory choice. Transportation Research Record: Journal of the Transportation Research Board 2134 (1), 153–163.

Kaplan, S., Bekhor, S., Shiftan, Y., 2011. Development and estimation of a semi-compensatory residential choice model based on explicit choice protocols. The Annals of Regional Science 47 (1), 51–80.

Kitamura, R., Fujii, S., 1998. Two computational process models of activity-travel behavior. In: T. Gärling, T. L., Westin, K. (Eds.), Theoretical Foundations of Travel Choice Modeling. Oxford: Elsevier, pp. 251–279.

Kitamura, R., Fujii, S., Kikuchi, A., Yamamato, T., 1998. An application of micro-simulator of daily travel and dynamic network flow to evaluate the effectiveness of selected TDM measures for CO2 emission reduction. In: Proceedings of the 77th Annual Meeting of the Transportation Research Board, Washington, D.C., USA.

Kitamura, R., Lula, C., Pass, E., 1993. AMOS: An activity-based flexible and behavioural tool for evaluation of TDM measures. In: Proceedings of 21st PTRC Summer Annual Meeting, University of Manchester, United Kingdom.

Kitamura, R., Nilles, J. M., Conroy, P., Fleming, D. M., 1990. Telecommuting as a transportation planning measure: Initial results of California pilot project. Transportation Research Record: Journal of the Transportation Research Board 1285 (1), 98–104.

Kitamura, R., Pas, E., Lula, C., Lawton, T., Benson, P., 1996. The sequenced activity mobility simulator (SAMS): An integrated approach to modeling transportation, land use and air quality. Transportation 23 (3), 267–291.

Kleene, S., 1956. Representations of events in nerve nets and finite automata. In: Shannon, C., McCarthy, J. (Eds.), Automata Studies. Princeton: Princeton University Press, pp. 3–41.

Knudson, B., Weidner, T., 2010. Using the Oregon statewide integrated model for the Oregon freight plan analysis. In: Proceedings of the TRB SHRP2 Symposium: Innovations in Freight Demand Modeling and Data, Washington, D.C., USA.

Kozen, D. C., 1997. Automata and Computability, 1st Edition. New York: Springer.

Kristoffersson, I., Engelson, L., 2008. Estimating preferred departure times of road users in a real-life network. In: Proceedings of the European Transport Conference 2008, the Netherlands.

Larsen, J., El-Geneidy, A., Yasmin, F., 2010. Beyond the quarter mile: Examining travel distances by walking and cycling, Montréal, Canada. Canadian Journal of Urban Research 19, 70–88.

Lawe, S., 2010. DaySim activity-based model implementation for Jacksonville, FL. Presented to the Panel on Activity-based Models, Florida Department of Transportation, available at http://www.fsutmsonline.net/images/uploads/mtf-files/DaySim_Activity_Based_Model_Implementation_Steve_Lawe.pdf, accessed on June 9, 2015.

Lefevre, B., Leipziger, D., Raifman, M., 2014. The trillion dollar question: Tracking public and private investment in transport. Working Paper, available at http://www.wri.org/publication/trillion-dollar-question, accessed on June 9, 2015.

Lenntorp, B., 1977. Paths in space-time environments: A time-geographic study of movement possibilities of individuals. Environment and Planning A 9 (8), 961–972.

Leonard, D., Gower, P., Taylor, N., 1989. CONTRAM: Structure of the model. Tech. rep., Transport and Road Research Laboratory, Berkshire, England.

Levenshtein, V. I., 1966. Binary codes capable of correcting deletions, insertions, and reversals. Soviet Physics Doklady 10 (8), 707–710.

Lu, L., 2013. W-SPSA: An efficient stochastic approximation algorithm for the off-line calibration of dynamic traffic assignment models. Master's thesis, Massachusetts Institute of Technology.

Lu, Y., Muhammad, A., Basak, K., Pereira, F., Carrion, C., Saber, V., Loganathan, H., 2015. SimMobility mid-term simulator: A state of the art integrated agent based demand and supply model. In: Proceedings of the 93rd Annual Meeting of the Transportation Research Board, Washington, D.C., USA.

Mahmassani, H., Hu, T., Jayakrishnan, R., 1992. Dynamic traffic assignment and simulation for advanced network informatics (DYNASMART). In: Proceedings of the 2nd International CAPRI Seminar on Urban Traffic Networks, Capri, Italy.

Manski, C., 1977. The structure of random utility models. Theory and Decision 8 (3), 229–254.

McFadden, D., 1973. Conditional logit analysis of qualitative choice behavior. In: Zarembka, P. (Ed.), Frontiers of Econometrics. New York: Academic Press, pp. 105–142.

McFadden, D., 1987. Regression-based specification tests for the multinomial logit model. Journal of Econometrics 34 (1), 63–82.

McNaughton, R., Yamada, H., 1960. Regular expressions and state graphs for automata. IEEE Transactions on Electronic Computers 9 (1), 39–47.

Metropolitan Transportation Commission, 2011a. Bay Area express lanes public partnership application for high occupancy toll lanes. Tech. rep., Prepared for the California Transportation Commission, available at http://mtc.ca.gov/projects/express_lanes/pdfs/FINAL_CTC_Application_092811b.pdf, accessed on June 9, 2015.

Metropolitan Transportation Commission, 2011b. Plan/Bay Area: Technical summary of predicted traveler responses to first round scenarios. Tech. rep., Prepared by Metropolitan Transportation Commission, San Francisco, available at http://planbayarea.org/pdf/First_Round_Travel_Model_Technical_Summary.pdf, accessed on June 9, 2015.

Meyer, M. D., 2000. Refocusing transportation planning for the 21st century. In: Refocusing Transportation Planning for the 21st Century: Proceedings of Two Conferences.

Miller, E. J., 2014. Operational implementation of the TASHA agent-based microsimulation travel model system in the Greater Toronto-Hamilton area. In: Proceedings of the 5th Transportation Research Board Conference on Innovations in Travel Modelling, Baltimore, USA.

Miller, E. J., 2015. First we take Toronto ... Lessons from the TASHA implementation and next steps. In: Proceedings of the Behavioural Detail and Computational Demands in Agent-Based Models Workshop, Future Cities Laboratory, Singapore.

Miller, E. J., Roorda, M. J., 2003. A prototype model of 24-hour household activity scheduling for the Toronto area. Transportation Research Record: Journal of the Transportation Research Board 1831 (1), 114–121.

Mirzaei, A., Eluru, N., 2007. A comparison of CEMDAP activity-based model with DFWRTM 4-step model. In: Proceedings of the 11th Transportation Research Board National Transportation Planning Application Conference, Daytona Beach, USA.

Mitchell, R., Rapkin, C., 1954. Urban Traffic: A Function of Land Use. New York: Columbia University Press.

Miwa, T., Sakai, T., Morikawa, T., 2008. Route identification and travel time prediction using probe-car data. International Journal of ITS Research 2 (1), 21–28.

Newell, A., Simon, H. A., 1972. Human Problem Solving. Englewood Cliffs: Prentice-Hall.

Nichols, B., Childress, S., Coe, S., 2014. SoundCasting at PSRC: Activity-based model development with Emme. In: Proceedings of the INRO Conference, Seattle, USA.

Ohmori, N., Nakazato, M., Harata, N., 2005. GPS mobile phone-based activity diary survey. In: Proceedings of the Eastern Asia Society for Transportation Studies. Vol. 5. pp. 1104–1115.

Outwater, M., Charlton, B., 2006. The San Francisco model in practice: Validation, testing, and application. In: Proceedings of Innovations in Travel Demand Modeling Conference, Transportation Research Board, Austin, USA. pp. 24–29.

Owen, L. E., Zhang, Y., Rao, L., McHale, G., 2000. Street and traffic simulation: Traffic flow simulation using CORSIM. In: Proceedings of the 32nd Conference on Winter Simulation. Society for Computer Simulation International, pp. 1143–1147.

Parsons Brinckerhoff, 2005a. The MORPC travel demand model validation and final report. Tech. rep., Prepared for the Mid-Ohio Regional Planning Commission.

Parsons Brinckerhoff, 2005b. Transportation models and data initiative general final report: New York best practice model (NYBPM). Tech. rep., Prepared for the New York Metropolitan Transportation Council, available at http://www.nymtc.org/project/bpm/model/bpm_finalrpt.pdf, accessed on June 9, 2015.

Parsons Brinckerhoff, 2006. Progress report for the year 2005, regional transportation plan major update project for the Atlanta Regional Commission, general modeling task 13 (activity/tour-based models). Tech. rep., Prepared for the Atlanta Regional Commission.

Parsons Brinckerhoff, 2010. Ohio statewide model. Tech. rep., Prepared for the Ohio Department of Transportation, available at https://www.dot.state.oh.us/Divisions/Planning/SPR/ModelForecastingUnit/Documents/osmp.pdf, accessed on June 9, 2015.

Parsons Brinckerhoff, HBA Specto Incorporated, EcoNorthwest, 2010. Oregon statewide integrated model (SWIM2): Model description. Tech. rep., Prepared for the Oregon Department of Transportation, available at http://www.oregon.gov/ODOT/TD/TP/docs/statewide/swim2.pdf, accessed on June 9, 2015.

Pas, E. I., 1983. A flexible and integrated methodology for analytical classification of daily travel-activity behavior. Transportation Science 17 (4), 405–429.

Pas, E. I., 1984. The effect of selected sociodemographic characteristics on daily travel-activity behavior. Environment and Planning A 16 (5), 571–581.

Paz, A., Molano, V., Gaviria, C., 2012. Calibration of CORSIM models considering all model parameters simultaneously. In: Proceedings of the 15th International IEEE Conference on Intelligent Transportation Systems, Anchorage, USA. pp. 1417–1422.

Pendyala, R. M., Chiu, Y.-C., Waddell, P., Hickman, M., Konduri, K. C., Sana, B., 2010. The design of an integrated model of the urban continuum - Location choices, activity–travel behavior, and dynamic traffic patterns. In: Proceedings of the 12th World Conference on Transport Research, Lisbon, Portugal.

Pendyala, R. M., Kitamura, R., Kikuchi, A., Yamamoto, T., Fujii, S., 2005. Florida activity mobility simulator: Overview and preliminary validation results. Transportation Research Record: Journal of the Transportation Research Board 1921 (1), 123–130.

Pendyala, R. M., Kitamura, R., Reddy, D. V. G. P., 1995. A rule-based activity-travel algorithm integrating neural network of behavioural adaptation. In: Proceedings of the EIRASS Conference on Activity-Based Approaches: Activity Scheduling and the Analysis of Activity Patterns, Eindhoven, The Netherlands.

Pendyala, R. M., Kitamura, R., Reddy, D. V. G. P., 1998. Application of an activity-based travel-demand model incorporating a rule-based algorithm. Environment and Planning B: Planning and Design 25 (5), 753–772.

Petersen, E., Vovsha, P., 2005. Auto allocation modelling in activity-based models. In: Proceedings of the 10th Transportation Research Board National Transportation Planning Application Conference, Portland, USA.

Petersen, E., Vovsha, P., 2006a. Directions for coordinated improvement of travel surveys and models. In: Proceedings of Innovations in Travel Demand Modeling Conference, Transportation Research Board, Austin, USA. pp. 85–88.

Petersen, E., Vovsha, P., 2006b. Intra-household car type choice for different travel needs. In: Proceedings of the 85th Annual Meeting of the Transportation Research Board, Washington, D.C., USA.

Phithakkitnukoon, S., Horanont, T., Lorenzo, G., Shibasaki, R., Ratti, C., 2010. Activity-aware map: Identifying human daily activity pattern using mobile phone data. In: Salah, A., Gevers, T., Sebe, N., Vinciarelli, A. (Eds.), Human Behavior Understanding. Berlin: Springer, pp. 14–25.

Pinjari, A. R., Eluru, N., Copperman, R. B., Sener, I. N., Guo, J. Y., Srinivasan, S., Bhat, C. R., 2006. Activity-based travel-demand analysis for metropolitan areas in Texas: CEMDAP models, framework, software architecture and application results. Tech. Rep. FHWA/TX-07/0-4080-8, Prepared for the Texas Department of Transportation.

Popuri, Y., Ben-Akiva, M., Proussaloglou, K., 2008. Time-of-day modeling in a tour-based context: Tel Aviv experience. Transportation Research Record: Journal of the Transportation Research Board 2076 (1), 88–96.

Post, E. L., 1943. Formal reductions of the general combinatorial decision problem. American Journal of Mathematics 65 (2), 197–215.

Prato, C. G., 2009. Route choice modeling: Past, present and future research directions. Journal of Choice Modelling 2 (1), 65–100.

Rahmani, M., Koutsopoulos, H., Ranganathan, A., 2010. Requirements and potential of GPS-based floating car data for traffic management: Stockholm case study. In: Proceedings of the 13th International IEEE Conference on Intelligent Transportation Systems, Madeira Island, Portugal. pp. 730–735.

Ramming, M., 2002. Network knowledge and route choice. Ph.D. thesis, Massachusetts Institute of Technology.

Rasouli, S., Timmermans, H., 2014. Activity-based models of travel demand: Promises, progress and prospects. International Journal of Urban Sciences 18 (1), 31–60.

RDC Inc., 1995. Activity-based modeling system for travel deamnd forecasting. Tech. rep., Prepared for the Metropolitan Washington Council of Governments, available at http://www.tongji.edu.cn/~yangdy/amos/amospapr.htm, accessed on June 9, 2015.

Recker, W., 1995. The household activity pattern problem: General formulation and solution. Transportation Research Part B: Methodological 29 (1), 61–77.

Recker, W., McNally, M., Root, G., 1986a. A model of complex travel behavior: Part I — Theoretical development. Transportation Research Part A: General 20 (4), 307–318.

Recker, W., McNally, M., Root, G., 1986b. A model of complex travel behavior: Part II — An operational model. Transportation Research Part A: General 20 (4), 319–330.

Ronald, N., Arentze, T., Timmermans, H., 2012. Modeling social interactions between individuals for joint activity scheduling. Transportation Research Part B: Methodological 46 (2), 246–290.

Roorda, M. J., Doherty, S. T., Miller, E. J., 2005. Operationalising household activity scheduling models: Addressing assumptions and the use of new sources of behavioral data. In: Integrated Land Use and Transportation Models, Behavioral Foundations. Oxford: Elsevier, pp. 61–85.

Roorda, M. J., Miller, E. J., 2005. Strategies for resolving activity scheduling conflicts: An empirical analysis. In: Timmermans, H. J. P. (Ed.), Progress in Activity-Based Analysis. Oxford: Elsevier, pp. 203–222.

Roorda, M. J., Miller, E. J., Habib, K. M. N., 2008. Validation of TASHA: A 24-h activity scheduling microsimulation model. Transportation Research Part A: Policy and Practice 42 (2), 360–375.

Roorda, M. J., Miller, E. J., Kruchten, N., 2006. Incorporating within-household interactions into a mode choice model using a genetic algorithm for parameter estimation. Transportation Research Record: Journal of the Transportation Research Board 1985 (1), 171–179.

Rossi, T., 2012. TourCast: Taking advantage of experience from around the U.S. in implementing an activity-based model. Presented to New York Metropolitan Transportation Council, available at http://www.nymtc.org/project/BPM/BPM_UserGrpMtgs/2012%20Meetings/NYBPM%20Users%20Grp%20Mtg%20091212/NYMTC%20pres%20Rossi%209-12-12.pdf, accessed on June 9, 2015.

Rossi, T., Lemp, J., Komaduri, A., Ehrlich, J., 2013. Comparison of activity-based model parameters between two cities. In: Proceedings of the 14th Transportation Research Board National Transportation Planning Applications Conference, Columbus, USA.

Rousseau, G., 2012. ARC's experience using its CT-RAMP activity-based model. In: TMIP Webinar Series, available at http://media.tmiponline.org/webinars/2012/TMIP_ABM_Webinars/ARC_ABM/ARC_ABM_Webinar_Oct_22_2012.pdf, accessed on June 9, 2015.

Rozenberg, G., Salomaa, A., 1997. The Handbook of Formal Language. Vol. 2. Berlin: Springer.

Sabina, E. E., Erhardt, G. D., Consult, P., Rossi, T., Coil, J., 2006. Processing the Denver travel survey to support tour-based modeling: Methods, data and lessons learned. In: Proceedings of Innovations in Travel Demand Modeling Conference, Transportation Research Board, Austin, USA. pp. 49–53.

Sabina, E. E., Rossi, T., 2006. Using activity-based models for policy decision making. In: Proceedings of Innovations in Travel Demand Modeling Conference, Transportation Research Board, Austin, USA. pp. 177–180.

SACOG, 2007. Comments on the Placer Vineyards specific plan. Prepared by the Sacramento Area Council of Governments, available at http://www.placer.ca.gov/~/media/cdr/ECS/EIR/PVSP/SFEIRPgs95to143.pdf, accessed on June 9, 2015.

Bibliography

SACOG, 2008. Sacramento Region 2035 Metropolitan Transportation Plan. Prepared by the Sacramento Area Council of Governments, available at http://sacog.org/mtp/2035/final-mtp/, accessed on June 9, 2015.

San Francisco County Transportation Authority, 2010. San Francisco mobility, access, and pricing study. Tech. rep., Prepared by the San Francisco County Transportation Authority, available at http://www.sfcta.org/sites/default/files/content/Planning/CongestionPricingFeasibilityStudy/PDFs/MAPS_study_final_lo_res.pdf, accessed on June 9, 2015.

Sang, S., O'Kelly, M., Kwan, M.-P., 2011. Examining commuting patterns: Results from a journey-to-work model disaggregated by gender and occupation. Urban Studies 48 (5), 891–909.

Schmitt, D., 2007. Application of the MORPC micro-simulation model: New Starts review. In: Proceedings of the 11th Transportation Research Board National Transportation Planning Application Conference, Daytona Beach, USA.

Shen, L., Stopher, P. R., 2014. Review of GPS travel survey and GPS data-processing methods. Transport Reviews 34 (3), 316–334.

Song, C., Qu, Z., Blumm, N., Barabási, A.-L., 2010. Limits of predictability in human mobility. Science 327 (5968), 1018–1021.

Spall, J. C., 1998. Implementation of the simultaneous perturbation algorithm for stochastic optimization. IEEE Transactions on Aerospace and Electronic Systems 34 (3), 817–823.

Spall, J. C., 2001. Stochastic Optimization, Stochastic Approximation and Simulated Annealing. New York: John Wiley & Sons.

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

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

Stopher, P., FitzGerald, C., Xu, M., 2007. Assessing the accuracy of the Sydney household travel survey with GPS. Transportation 34 (6), 723–741.

Swait, J., Ben-Akiva, M., 1987. Empirical test of a constrained choice discrete model: Mode choice in São Paulo, Brazil. Transportation Research Part B: Methodological 21 (2), 103–115.

Vanasse Hangen Brustlin, Inc., 2006. Review of current use of activity-based modeling. Tech. rep., Prepared for the Metropolitan Washington Council of Governments, available at http://www.ampo.org/assets/669_fy06activitybasedmodels.pdf, accessed on June 9, 2015.

Vij, A., Carrel, A., Walker, J. L., 2013. Incorporating the influence of latent modal preferences on travel mode choice behavior. Transportation Research Part A: Policy and Practice 54 (1), 164–178.

Vovsha, P., 2009. Integration of AB models and microsimulation models. Traffic Engineering and Control 50 (2), 85–86.

Vovsha, P., Freedman, J., Livshits, V., Sun, W., 2011. Design features of activity-based models in practice: CT-RAMP experience. Transportation Research Record: Journal of the Transportation Research Board 2254 (1), 19–27.

Vovsha, P., Gliebe, J., Petersen, E., Koppelman, F., 2004. Comparative analysis of sequential and simultaneous choice structures for modeling intra-household interactions. In: Timmermans, H. (Ed.), Progress in Activity-Based Analysis. Oxford: Elsevier, pp. 223–258.

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

Walker, J., Ehlersa, E., Banerjeea, I., Dugundjib, E. R., 2011. Correcting for endogeneity in behavioral choice models with social influence variables. Transportation Research Part A: Policy and Practice 45 (4), 362–374.

Walsh, D., Su, M., Luk, J., 2008. New performance indicators for network operations using real-time traffic data. Road and Transport Research 17 (3), 47 – 54.

Weiner, E., 1997. Urban transportation planning in the United States: A historical overview (5th edition). Tech. Rep. DOT-T-97-24, US Department of Transportation, Washington, D.C.

Wen, C.-H., Koppelman, F., 2000. A conceptual and methodological framework for the generation of activity-travel patterns. Transportation 27 (1), 5–23.

Wyard, P., 1993. Context-free grammar induction using genetic algorithms. In: IEE Colloquium on Grammatical Inference: Theory, Applications and Alternatives. pp. P11/1–P11/5.

Yang, D., Arentze, T., Timmermans, H., 2010. Primary and secondary effects of teleworking policies on household energy consumption. In: Proceedings of the 12th World Conference on Transport Research, Lisbon, Portugal.

Yang, Q., 1997. A simulation laboratory for evaluation of dynamic traffic management systems. Ph.D. thesis, Massachusetts Institute of Technology.

Zheng, Y., Zhang, L., Xie, X., Ma, W.-Y., 2009. Mining interesting locations and travel sequences from GPS trajectories. In: Proceedings of the 18th International Conference on World Wide Web, Madrid, Spain. pp. 791–800.

Zorn, L., Sall, E., Wu, D., 2012. Incorporating crowding into the San Francisco activity-based travel model. Transportation 39 (4), 755–771.

Appendices

Appendix A

Recent Research Accomplishments

Journal Papers

[1] Li, S. and Lee, D.-H., 2015. Learning Daily Activity Patterns with Probabilistic Grammars. *Transportation*. DOI: 10.1007/s11116-015-9622-1

Conference Proceedings

[1] Li, S., Enam A., Abou-Zeid, M. and Ben-Akiva, M., 2013. Travel Time Modeling with GPS and Household Survey Data. In *Proceedings of the 92nd Annual Meeting of the Transportation Research Board, January 2013, Washington, D.C., USA*.

[2] Li, S., Carrion, C., Abou-Zeid, M. and Ben-Akiva, M., 2013. Activity-based Travel Demand Models for Singapore. In *Proceedings of the 18th HKSTS International Conference, December 2013, Hong Kong*.

[3] Li, S. and Lee, D.-H., 2014. Learning Daily Activity Pattern with Probabilistic Grammar. In *Proceedings of the 93rd Annual Meeting of the Transportation Research Board, January 2014, Washington, D.C., USA*.

[4] Lu, Y. and Li, S., 2014. Empirical Study of Within-Day O-D Prediction Using Taxi GPS Data in Singapore. In *Proceedings of the 93rd Annual Meeting of the Transportation Research Board, January 2014, Washington, D.C., USA*.

[5] Li, S., Pereira, F., Lee, D.-H. and Ben-Akiva, M., 2015. A Two-Stage Choice Model for Daily Activity Patterns with Customized Choice Sets. In *Proceedings of the 4th International Choice Modeling Conference, May 2015, Austin, Texas, USA*.