MODELING ACTIVITY SELECTION AND SCHEDULING BEHAVIOR OF POPULATION COHORTS WITHIN AN ACTIVITY-BASED TRAVEL DEMAND MODEL SYSTEM

by

Mohammad Hesam Hafezi

Submitted in partial fulfilment of the requirements
for the degree of Doctor of Philosophy

at

Dalhousie University
Halifax, Nova Scotia
March 2018

I dedicate this dissertation to my parents,
who taught me the meaning of unconditional love,
and
to my brother.

Table of Contents

List of Tables

List of Figures

Abstract

Understanding the time-use activity patterns of population cohorts in the region will contribute greatly to modeling spatio-temporal urban transportation demand models. The research detailed in this dissertation focuses on the development of the Scheduler for Activities, Locations, and Travel (SALT) disaggregated travel demand microsimulation model. The SALT modeling framework comprises a series of micro-behavioral modules that employ behaviorally realistic econometric, advanced machine learning, and data mining techniques to construct the 24-hour activity schedule and the corresponding travel linked with activities accomplished by individuals. A state-of-art three-dimensional, four-stage pattern recognition model is developed to identify population clusters with homogeneous time-use daily activity patterns, and to derive a representative set of activity patterns in each cluster. Each identified population cluster provides essential information related to temporal, spatial, and socio-demographic characteristics of individuals and activities, which are crucial for modeling the successive micro-behavioral modules of the SALT model. The representative behavior within each cluster is then used as an information guide for agent-based modeling.

A new agent-based inference model is developed to predict various facets of the daily activity agenda, such as stop number, activity type, and activity sequential arrangement. In the next phase, temporal attributes of each activity in the agenda are predicted and the 24-hour activity schedule of all individuals is formed through a heuristic decision rule-based algorithm. Finally, a population synthesizer procedure is developed in order to implement the SALT system for the entire region. In addition, this study models the daily time-use activity patterns and estimated emission factors for university commuters, considered as a special trip generator in regional travel demand models. The data used for the analysis is from the large Halifax Space-Time Activity Research (STAR) household survey, which provides GPS-validated time-diary data for 2,778 person-days. Results show that the SALT scheduling model is able to assemble the traveler's schedule with an average 82% accuracy in the 24-hour period. The proposed simulation modeling framework is useful for urban and transport modelers to advance transportation demand management for different segments of the urban population, as well as to analyze environmental mitigation and transport policy scenarios.

List of Abbreviations and Symbols Used

ADAPTS	Agent-Based Dynamic Activity Planning and Travel Scheduling Model
ALBATROSS	A Learning-BAsed TRansportation Oriented Simulation System
AMOS	Activity Mobility Simulator
AMOS	Activity Mobility Simulator
BDA	Big Data Analytics
BPNN	Back-Propagation Neural Network
CARLA	Combinatorial Algorithm for Rescheduling Lists and Activities
CART	Classification and Regression Tree
CATI	Computer-Assisted Telephone Interview
CDF	Cumulative distribution function
CEMDAP	Comprehensive Econometric Micro-simulator for Daily Activity-travel Patterns
CHI2	Minimum Chi-Squared
CMAs	Census Metropolitan Areas
CO	Carbon Monoxide
CO	Combinatorial Optimization
CO_2	Carbon Dioxide
CT	Census Tract
DA	Dissemination Area
DAP	Daily Activity Pattern
EnACT	Environmentally Aware Commuter Travel Diary Survey
FBS	Fitness-Based Synthesis
FCM	Fuzzy C-Means
GHG	Greenhouse Gases

GISICAS	GIS-Interfaced Computational process model for Activity Scheduling
GPS	Global Positioning System
GSS	General Social Survey
HIPF	Hierarchical Iterative Proportional Fitting
HMM	Hidden Markov Models
ICB	Inner Commuter Belt
ICT	Information and Communications Technology
INC	Inner City
IPF	Iterative Proportional Fitting
IPU	Iterative Proportional Updating
KS	Kolmogorov-Smirnov
MCS	Monte Carlo Simulation
MDA	Mean Decrease Accuracy
MDCEMV	Multiple Discrete Continuous Extreme Value
MDI	Mean Decrease Impurity
MIPFP	Multidimensional Iterative Proportional Fitting Procedure
ML	Maximum Likelihood
MOVES	Motor Vehicle Emission Simulator
M-SAM	Multiple Sequence Alignment Method
NO_x	Nitrogen Oxide
NRN	National Road Network
OCB	Outer Commuter Belt
ONC	On-Campus Zone
OOB	Out-Of-Bag
PCA	Principal Component Analysis

PDF	Probability Distribution Function
PM_{10}	Particulate Matter Under 10 Micron Diameter
$PM_{2.5}$	Particulate Matter Under 2.5 Microns Diameter
PUMF	Public Use Microdata File
RF	Random Forest
RL	Regional Level
SALT	Scheduler for Activities, Locations, and Travel
SAM	Sequence Alignment Method
SMASH	Simulation Model of Activity Scheduling Heuristics
STAR	Space-Time Activity Research
STARCHILD	Simulation of Travel/Activity Responses to Complex Household Interactive Logistic Decisions
SUB	Suburban Area
SVM	Support Vector Machine
TASHA	Travel and Activity Scheduler for Household Agents
TDM	Travel Demand Management
TESS	Transportation and Environmental Simulation Studies
THC	Total-Hydrocarbon
TURP	Time Use Research Program
VOC	Volatile-Organic-Compounds
WLSQ	Weighted Least Squares
AH_{qw}	Amount of cell w in the count Table q
AH_{qw}^{y-1}	Value of cell w in the count Table q
B_{lit}	Best split at each node in decision tree
D_r	Train data

EP_q	Error-percentage for control Table q
E_{min}	Shortest distance between two clusters
$E_n(C)$	Subset of training data
$F^{(t)}$	Fuzzy centroid
F_{1Q}^{xy}	Fitness value type 1 for control Table Q
F_{2Q}^{xy}	Fitness value type 2 for control Table Q
G_{qw}^{y-1}	Difference value between control and count Tables
H_{qw}	Amount of cell w in control Table q
$I_{m,n}$	Out-of-bag data set of the m-tree
$I_{m,n}^{u}$	Permuted data for variable u
MH_{qw}^{x}	Contribution of the x^{th} household in the seed data to the w^{th} cell
O_n	Out-of-bag error
$O_n(.;\Theta_m)$	Estimation for the m-th tree
$S_{clas,n}(u,w)$	Split criterion
T_C	Set of all feasible cuts in C
T_e	Test data
$T_{i,j}$	Alignment with maximum score
T_i	Density of data point
X_i	Response variable
Y_n	Set of predictor variables
$\bar{Y}_C$	Average of the Y_i such that X_i belongs to C
$\bar{x}^{(j)}$	Sample class mean
h_l	Fuzzy membership
$\hat{P}$	Final prediction result

$\hat{d}$	Predicted classification
$\hat{d}(n\|m)$	Posterior probability of class n for observation m
f_i	Relative frequency of activity j in the cluster g
mm_y	Selected households for adding into the count Tables
m^q	Selected household type 1 or 2 according to the fitness value
n_q	Average distribution of control Table q in the seed data
p_i	Data point represents a transformed person-day activity pattern
rm_y^{qw}	Selected household for the cell w in the count Table q
r_n	Random observations drawn from the sampled data points
t_n	New arrival data at the testing stage
u_n^*, w_n^*	Best cut point
u_r	Cluster radius
v_u	Variance for the uth class
w^*	Maximum density
β_{ih}	Membership degree of data points in each cluster
$\bar{\partial}$	Accept ratio
$\underline{\partial}$	Reject ratio
MDA($X^{(u)}$)	Mean Decrease Accuracy of the variable $X^{(u)}$
$I(G)$	Gini index
Q	Predictor variables
SC	Distance score between two members of cluster
T	Kolmogorov-Smirnov test
$T(d\|n)$	Classification cost of an observation as d when its true class is n
W	Index representing the various cells in the control Table
m	Number of trees

n	Number of activity categories
q	Index representing for the both count and control Tables
r	Number of corresponding strings g_i and g_j in a bit string
s	Number of corresponding lengths of d_{g_i} and d_{g_j} in a bit string
u	Randomly selected predictor variable
w	Index representing the various cells in the count Table
x	Selected household
y	Iteration number
z	Sample size of person-days in the dataset
$-\rho$	Penalty for two mismatched strings
$-\sigma$	Gap penalty
τ	Fuzzy parameter
φ	Specified minimum threshold in the algorithm
ϑ	Squash factor

Acknowledgements

First and foremost, I would like to express my sincere gratitude to my supervisor, Professor Lei Liu, for his intellectual and scientific support throughout my PhD program. A reliable, intelligent, caring, and patient supervisor, who created such an excellent research atmosphere that transformed my PhD to be one of the best, and most productive periods of my life. My discussions with Prof. Liu deeply enlarged my vision towards my research aims and provided me with several innovative ideas to explore. I am sincerely thankful to him for all his contributions of time, energy and dedication to my success. Without his understanding and positive attitude this journey wouldn't have been as enjoyable as it was.

I would also like to extend a special thank you to Professor Hugh Millward. His experience in transportation modeling, time-use research and modeling of activity travel patterns, challenging questions, and invaluable feedback gave me guidance as well as inspiration that I needed to remain confident throughout this endeavor. I have been extremely fortunate to work with Prof. Millward and have him as an example, both academically and personally, in my life. This thesis would not have been possible without his patient, guidance and kind advice. Also, I would like to extend my gratitude to all my advisory committee members, Professor Brian Baetz, Professor Yonggan Zhao and Dr. Haibo Niu, for their helpful technical comments on my thesis.

I gratefully acknowledge the financial support provided to this project by the National Science and Engineering Research Council (NSERC), the Government of Nova Scotia for the Nova Scotia Research and Innovation Graduate Scholarship, Faculty of Engineering for Excellence Award and the Faculty of Graduate Studies for Excellence in Performance Award.

The primary data source for this research was provided by the Halifax STAR Project at Saint Mary's University, supported through the Atlantic Innovation Fund from the Atlantic Canada Opportunities Agency, Project No.181930. The data provided by the STAR project was pivotal in achieving the aim of the research. Their support was greatly appreciated. I would also like to thank Dalhousie research ethics board and Dalhousie

university administration for cooperation in promoting and implementing the EnACT survey.

I also owe a great debt of gratitude to my friend and colleague Dr. Naznin Sultana Daisy, who helped me during the different stages of my research and developed other components of the SALT model system towards forming this project an innovative and exciting approach to the field. My gratitude extends as well to my friends and colleagues at TESS, including but not limited to Dr. Zhi Wang, Li Lingyu, Wenwen Pei and Hao Wen who also helped me during the different stages of my research at Dalhousie and their advice, support and suggestions when requested. It was an amazing experience to be part of this united team. I would also like to acknowledge all of the reviewers, anonymous and otherwise, who took the time to provide comments and advice on all aspects of this work at various points.

Last but not least, I would like to express my sincere and heartful gratitude to my dear parents and my brother for encouraging me to fulfill my dreams and cheering for me through my success. Their love and support made this journey easier to me. I could never have achieved this milestone without you.

Chapter 1 Introduction

1.1 Background

Transportation can be considered as one of the main and essential human activities, that involves nearly everyone on a daily basis. Complexities in individual travel behavior have increased with continued urban development and rapid technological progress. Trip chaining and multimode transport, flexible working hours, self-employment, and online shopping have become far more common in recent years (Goran 2001). As travel behavior becomes more complex, travel demand forecasting requires more detailed information. From a disaggregated modeling point of view, there are significant associations between trips and the activity participation of travelers (Kitamura et al. 1997). Furthermore, travelers with varying socio-demographic and socio-economic characteristics in the region have divergent time-use activity patterns. This dissertation presents a new disaggregated travel demand microsimulation model framework that is sensitive to the mix of variables connected to travelers' decisions. A new pattern recognition model is developed to identify population clusters with homogeneous time-use daily activity patterns, and to derive a representative set of activity patterns in each cluster. The representative behavior within each cluster is then used as an information guide for innovative agent-based modeling of the 24-hour activity schedule and the travel linked to it.

To date, numerous travel demand models have been developed, using both aggregated and disaggregated approaches, for modeling short-term and long-term choices of travelers, such as activity participation, timing, transport mode, activity location, route choice,

work/residential location, and vehicle ownership (Oppenheim 1995; Ortuzar and Willumsen 2011).

The first generation of travel demand models, commonly known as four-stage models, were developed to evaluate the short-term and long-term effects of transport-related infrastructure investments in the late 1950s (Weiner 1999). The conventional four-stage models, which were later improved and known as trip-based models, have been employed widely to forecast traffic flows and volumes at the aggregated spatial level, such as traffic analysis zones (Goran 2001; McNally 2007). Trip productions and attractions are predicted at the aggregated level based on the attributes of the zone. Subsequently, trips are distributed among origin-destination pairs of traffic zones (Ortuzar and Willumsen 2011). Further, features of the four-stage models were to forecast the mode choice and trip assignment of predicted trips in the network.

Despite the pervasive application of these models in transportation planning, these models also have some major limitations (Boyce and Williams 2015). For instance, interdependencies between trips during the day or trips belonging to the same trip chain are not captured. Furthermore, joint trips of individuals belonging to the same household, and how household interaction impact trip scheduling, are not taken into account. Another limitation is the higher temporal and spatial aggregation level. Centroids of zones are considered as single points for trip origins and destinations and modeling tends to differentiate only peak traffic times versus off-peak traffic times. Physical and institutional constraints, as well as the traveler's imperfect knowledge of their environment, are not considered in scheduling (Clarke 1986; Ettema, Borgers and Timmermans 1993; Jang, Chiu and Zheng 2013).

To overcome some of the major limitations of trip-based models, and to provide more precise forecasting models, in the late 1970s transport modelers developed the second generation of travel demand models, commonly known as disaggregated trip-based models (Marcotte and Nguyen 2013). The disaggregated trip-based models are more sensitive to the sociodemographic characteristic of the population and land use attributes (Garling, Kwan and Golledge 1994). Disaggregated travel demand models can precisely capture the effects of factors that impact travel behavior, such as socio-demographic attributes and time allocation (Boyce and Williams 2015). Time allocation patterns have noteworthy implications for traffic congestion, demand estimation, and air quality.

Comparison between disaggregated trip-based models and actual daily life decision-making processes of individuals revealed that there are connections between trips and the activity participation of individuals (Kitamura, Chen and Pendyala 1997; Ben-Akiva and Bowman 1998b). Furthermore, due to the rapid growth in auto usage, the increased level of air pollution, noise pollution, and road congestion became more important policy issues for transport planners (Stopher, Hartgen and Li 1996; Bhat and Koppelman 1999). Finding a solution and policy for such complex transportation and environmental issues motivated transport modelers to develop the third generation of travel demand models, commonly known as activity-based travel demand models (Clarke 1986; Recker, McNally and Root 1986a; Recker, McNally and Root 1986b). Activity-based approaches focus on predicting traveler activity patterns. This includes modeling activity time-use patterns, activity episodes scheduling with associated attributes, and corresponding linked travel accomplished by travelers.

1.2 Activity-Based Travel Demand Modeling

Activity-based travel demand modeling is a recently emerged advanced disaggregated travel demand forecasting approach primarily introduced by Hagerstrand in the 1970s. The activity-based travel demand models take into consideration space-time constraints and the associations among activities and travel at the individual or household level (Hagerstrand 1970; Ellegard 1999). These considerations allow the travel demand models to more realistically incorporate the effects of travel circumstances on activity and travel selections (Goulias 1999). In comparison to the conventional four-stage models, activity-based travel demand models generate activities first; then, the destinations for the activities are determined, followed by transportation modes being identified, and, finally, the exact transportation routes used for each trip are forecast (Arentze and Timmermans 2000; Bowman and Ben-Akiva 2001; Fosgerau 2002). Incorporating space-time constraints and the effect of detailed individual and household level attributes into the modeling process substantially increases the model's ability to offer better forecasts of future travel patterns (Scott and Kanaroglou 2002; Arentze and Timmermans 2009; Timmermans and Zhang 2009; Liao, Arentze and Timmermans 2013).

Figure 1.1 demonstrates a typical activity-based model structure used in practice in North America. Activity generation and scheduling, tour and trip destination choice, tour and trip time of day, tour and trip mode decision, and network assignment are universal modules for most activity-based travel demand models (Recker 2001; Li, Lam and Wong 2014). These models are able to predict both long-term decisions (e.g. work/residential location and vehicle ownership) and short-term decisions (e.g. activity purposes, timing, transport mode, and locations) of a given synthetic population (Ben-Akiva and Bowman 1998a;

Hildebrand 2003). Different alternative policy scenarios on transport, land use, environmental impacts, and economic development can be assessed using the outcomes of activity-based travel demand models (Kitamura 1988).

To date, researchers and practitioners have employed different approaches, such as computational process-based and econometric-based methods for the development of activity-based models. A number of important specifications which measure the performance of such models are prediction accuracy, reproducibility, computational time, large scale operation capability, and performance at the household level (Kitamura 1988; Jovicic 2001; Rasouli and Timmermans 2014).

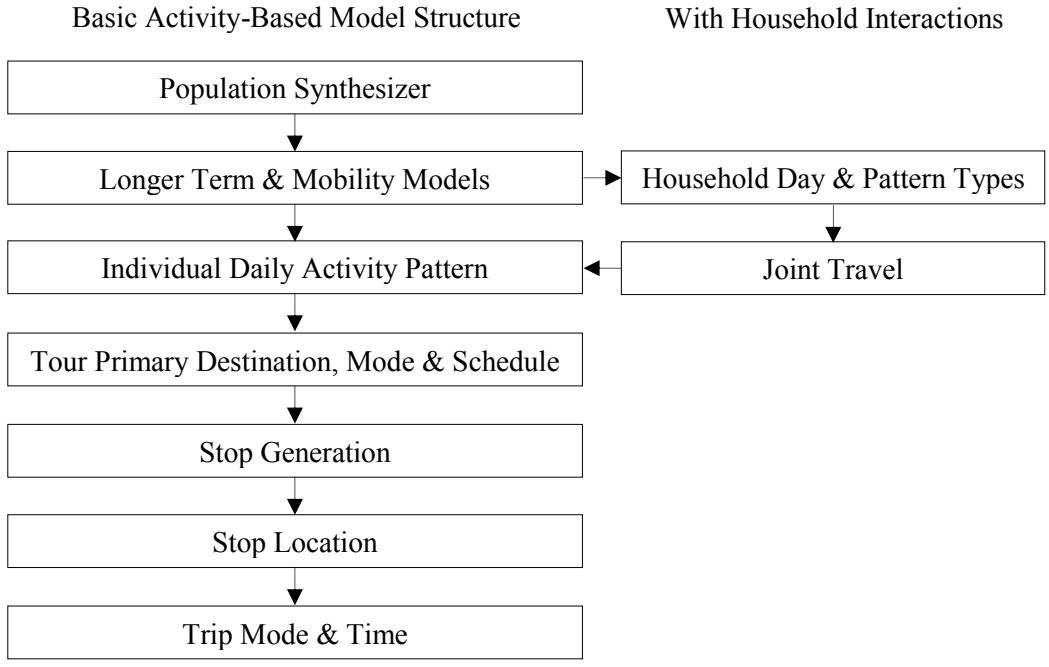


Figure 1.1 Typical activity-based model structures (Castiglione et al. 2015, p.107)

1.3 Motivations and Context of This Study

The primary motivation of this contribution is to model and micro-simulate various aspects of activity-travel decisions, including activity selection and scheduling behavior of population cohorts within an activity-based travel demand modeling system. To this end, this study presents the development of the Scheduler for Activities, Locations, and Travel (SALT) disaggregated travel demand micro-simulation model. The SALT model system, as an ongoing research program in the Department of Civil and Resource Engineering of Dalhousie University, is a new generation of travel demand model system that attempts to micro-simulate the formation of 24-hour activity schedules of individuals with varying characteristics and behavior for each population cohort, and model short-term and long-term travel and mobility decisions. The SALT model is comprised of five main components: population synthesizer, time-use activity pattern recognition, tour mode choice, activity destination choice, and activity/trip scheduling (Daisy 2018a).

The SALT model system is designed based on the multi-layer hybrid machine learning techniques. A series of behaviorally realistic advanced econometric and ensemble learning modules are incorporated in the SALT model system for modeling behavioral mechanisms and time-use activity patterns of populations. Particularly, this dissertation addresses the development of the state-of-art machine learning and pattern recognition models to identify representative time-use activity patterns, infer the scheduling behavior of individuals, and enable the generation of a synthetic population for the region. In addition, this study offers the modeling of the daily time-use activity patterns and estimates emission factors for university commuters, considered as a special trip generator in regional travel demand models. The ensemble learning micro-behavioral modules introduced in this study

are developed using advanced machine learning techniques that are novel in travel behavior analysis.

1.4 Objectives and Scope

The overall objective of this dissertation is to develop the Scheduler for Activities, Locations, and Travel (SALT) disaggregated travel demand micro-simulation model. Numerous advanced machine learning based micro-behavioral modules, including new pattern recognition and inference models within the SALT modeling framework, are developed to model and micro-simulate various aspects of activity selections and scheduling behavior of travelers. It is also hypothesized that people's daily activity patterns are strongly influenced by their socio-demographic attributes, such as age, household size, income, etc. The identification of homogeneous population clusters can also improve the accuracy in the estimation of activity-based travel demand models. To accomplish these goals, the following major objectives were carried out during the development of the SALT model system and the advancement of its machine learning micro-behavioral modules:

- To develop an inclusive pattern recognition modeling framework that can identify individuals with homogeneous daily activity patterns and group them into clusters;
- To establish a representative set of model individuals, who represent homogeneous cohorts in each identified cluster, and to develop a decision tree that can infer cluster membership for the selected traveler;

- To develop an inference model to learn and replicate activity engagement patterns of the population groups for daily agenda formation, including stop number, activity type, and activity sequential arrangement;
- To develop a scheduler model to predict temporal information associated with the traveler's daily activity schedule and build up the 24-hour activity schedule of individuals with varying characteristics and behavior for each cluster;
- To produce a baseline synthetic population for the study area at the dissemination area and regional levels by matching the distribution of household and individual sociodemographic characteristics with census data; and
- To investigate and model time-use activity patterns and estimate emission factors of a synthetic baseline population for university commuters, considered as a special trip generator in regional travel demand models.

1.5 Thesis Structure

To address the objectives discussed above, this dissertation is divided into nine chapters. Chapter two and chapters four to nine are organized as independent journal paper publications towards the development of the SALT model system. A summary of the content of each chapter is outlined below:

Chapter 1 provides an introduction on the research subject, outlines the objectives, and the organization of the dissertation. Chapter 2 provides a comprehensive review relevant to the activity-based travel demand models. Chapter 3 provides an overall description of the data and methods used in this study. Chapter 4 describes a comprehensive pattern recognition modeling technique for identifying individuals with homogeneous daily

activity patterns. Chapter 5 presents details of inference modeling technique for learning and replicating activity engagement patterns of different population groups. Chapter 6 describes a scheduling model to build up the 24-hour activity schedule. Chapter 7 focuses on the development of a population synthesizer model. Chapter 8 presents the modeling and finding of the daily time-use activity patterns and estimates emissions factors for university commuters. Finally, chapter 9 summarizes the findings of this research and key conclusions stemming from this research, along with recommendations for future work.

Chapter 2 Activity-Based Travel Demand Modeling: Progress and Possibilities[1]

2.1 Introduction

The goal for transportation models is to depict reality as precisely as possible. These models can be employed to investigate and find solutions for different transportation problems such as traffic congestion, transport-related GHG emissions, impacts on the economy, and traffic accidents. Transportation models are usually used to make forecasts in uncertain conditions, support management decision making, and inform policies to develop infrastructure and change travel behavior patterns of people (De Palma et al. 2011; Daisy et al. 2018a; Hafezi et al. 2018). The conventional four stage models, trip-based travel demand models, and activity-based travel demand models represent three major generations of travel demand models (Bates 2007). All three of them are able to provide travel demand forecasts with a suitable level of accuracy for the circumstances in which they were developed. Based on the model's characteristics, they are adequate to predict short-term decisions (e.g. activity purposes, timing, transport mode, and locations) and long-term decisions (e.g. work/residential location and vehicle ownership) related to transport problems (Nakamura, Hayashi and Miyamoto 1983; Hunt, Kriger and Miller 2005; Bates 2007).

The conventional four stage models belonged to the first generation of travel demand models. These models are able to provide travel demand forecasts, including traffic flows

[1] A version of this chapter has been published:
Hafezi, M. H., Millward, H., and L. Liu. (2018). "Activity-based travel demand modeling: Progress and possibilities". Peer reviewed ASCE proceedings of the International Conference on Transportation and Development (ICTD). Pittsburgh, Pennsylvania, USA.

and volumes at the scale of the aggregated traffic analysis zone (Goldner 1971). Commonly, the overall four stage modeling framework comprises four components: trip generation, trip distribution, mode choice, and traffic assignment. In this modeling approach, trip productions and trip attractions are estimated based on the attributes of respective zones (Shan, Zhong and Lu 2013). Over the past 50 years, the limitations of these models are becoming more challenging and questionable for transport modelers, due in part to a shifting emphasis on policy planning rather than regional planning (such as assessing long-term investment strategies). Some of the major shortcoming of these models are that they fail to consider interdependencies between trips during the day belonging to the same trip chain, lack modeling for joint trips of individuals belonging to the same household, and operate at coarse levels of temporal and spatial aggregation (Rasouli and Timmermans 2014; Boyce and Williams 2015; Hafezi, Liu and Millward 2018a).

These limitations of conventional four stage models, and the need for more sensitive forecasting models with better implementation capability for policy analysis, motivated transport planners to develop a second generation of travel demand models known as disaggregate trip-based models. Generally, disaggregate trip-based models, unlike four stage models, analyze each individual trip as independent and isolated (Boyce and Williams 2015). Aggregated and disaggregated trip-based models are two major types of the second generation of travel demand models. Although this type of modeling approach could overcome some major limitations of conventional four stage models, however, they were still insufficiently integrated for more sensitive forecasting needs (Dong et al. 2006). Some of the trip-based travel demand models' limitations are as follows. First, the time

component is not taken into account: the focus is on individual trips and/or tours without considering the temporal-spatial constraints between all trips and activities conducted in a given day. The sequential information is missing in the modeling process. Secondly, individuals within the same household are considered as isolated decision-makers. Third, and perhaps most importantly, they neglected the fact that the demand for travel is derived from the demand for activity engagement (Krizek 2003). Further analyses on the outcomes of disaggregate trip-based models and comparison with actual daily life decisions of individuals revealed that there are relations between trips and the activity participation of individuals. Consequently, the third generation of travel demand models, known as activity-based travel demand models, was developed (Dong et al. 2006). This study presents a comprehensive overview of computational based activity scheduling models. In addition, this study adds to the current literature by reviewing and assessing recent and ongoing rule-based machine learning models.

2.2 Activity-Based Modeling

Activity-based models of travel demand represent the third generation of travel demand models and are based on the concept that travel is a derived demand, initiating from the need of individuals to engage in out-of-home activities. The basic concept is that activities take place in both time and space, Hagerstrand formalized the notion of space-time in the 1970's, and developed the preliminary activity-based model (Hagerstrand 1970). In his time-geography theory, individuals are considered as living in a space-time prism in which their participation in activities is impacted by three constraints, as follows. Firstly, there are capability constraints that emphasize biological needs and existing resources that can necessitate or bound an individual's participation in an activity. Secondly, there are

coupling constraints that emphasize both the spatial and temporal requirements for an individual who joins with other individuals to perform a certain activity. Lastly, there are authority constraints that limit the individual's entry to certain activity locations or times. The theory posited that a decision to participate in a certain activity at a certain time and place is a joint outcome of numerous conditions and constraints (Ellegard 1999; Ellegard and Vilhelmson 2004; Ellegard and Svedin 2012; Widen, Molin and Ellegard 2012).

Activity-based travel demand models are able to better replicate travel decisions at the individual or disaggregated level, and may therefore yield better predictions of future travel patterns (Dong et al. 2006). In recent years, activity-based modeling has received much consideration and seen significant progress. A wide variety of modeling methods has been developed to model various components of activity-based models, such as activity type, activity sequence, activity frequency, sequential activity location, activity duration, and transport mode for the next trip (Auld et al. 2016; Bao et al. 2016; Jiang, Ferreira and Gonzalez 2017). Two of the most widely used approaches in activity-based travel demand models (Arentze and Timmermans 2000) are discussed in the following section.

2.2.1 Econometric Activity-Based Modeling

Econometric activity-based models are developed based on the random utility theory (McFadden 1980). This theory argued that individuals constantly desire to maximize the utility of their activity schedule. DAYSIM (2001), CEMDAP (2004), and MORPC (2002) are some well-known econometric activity-based travel demand models. The decision-making processes in econometric models are executed by employing a logit or nested logit

model. Furthermore, daily activity and travel patterns are considered as a set of tours in the modeling process (Bowman and Ben-Akiva 2001). In this respect, the home-based tour is formed by adding a series of travel and activity episodes that started and ended at the home-location. Based on the specific model characteristics, tour type is defined. Broadly, the primary tour contains activity episodes that have highest priority in the individual's daily activity agenda, such as work or school. Secondary tours comprise non-mandatory activity episodes such as shopping or leisure activities (Wen and Koppelman 2000). In most of the econometrics models, the activity priority sequences are specified as follows: work or school, personal maintenance (e.g., food shopping), and discretionary activities (Vovsha, Petersen and Donnelly 2004). Nevertheless, this priority sequence might be amended by activity duration or joint activity with others. The complexity of choice set within the nested-logit formation is controlled by predefining tour sequences such as home-work-home or home-shopping-school-home. These models also estimate the number of tours, number of stops, and vehicle allocation within the household context (Ettema, Borgers and Timmermans 1996).

Despite pervasive use of the econometric type of model in activity-based travel demand models, this modeling type has been criticized for requiring a heavy computationally rigorous scheduling system, when it employs a structurally high sequential logit model (Garling, Kwan and Golledge 1994). In addition, the predefined choice set for selection of daily activity patterns in the modeling process may not represent all possible alternatives for individuals' daily activity patterns. Furthermore, since this modeling approach uses only a restricted number of time periods for analysis (e.g. morning peak, noon, afternoon

peak), the models have also been criticized for how they capture time-of-day properties (Ettema, Borgers and Timmermans 1996).

2.2.2 Computational Based Activity Scheduling Models

More recently, the computational process modeling approach was developed following the context-dependent choice preferences theory (Arentze and Timmermans 2000). The advanced models contain the decision-making process as an internal loop in the modeling framework that is designed using a set of straightforward heuristic rules (e.g. if-then statements). The cognitive model (Hayes-Roth and Hayes-Roth 1979), is one of the earliest activity-based model developed using the computational process modeling framework. Potential choices for forming individual's agenda at various stages of abstraction are produced through the application of a series of heuristic rules for activity planning processes. In the following section several models classified as computational process models (CPMs) are overviewed.

Combinatorial Algorithm for Rescheduling Lists and Activities (CARLA)

The CARLA model attempts to simulate activity-travel patterns at the household level. It incorporates spatio-temporal constraints in the process of activity scheduling. Activities are selected and added to the schedule if they satisfy the constraints and set-up rules (Jones et al. 1983). Four rules categories are defined in the model: logical rules presume one unique activity at a time at one location, environmental rules denote authority constraints (access time restrictions to different places) and travel times between locations, inter-personal rules indicate coupling constraints (joint activities with other household members), and personal rules refer to personal preferences.

Simulation of Travel/Activity Responses to Complex Household Interactive Logistic Decisions (STARCHILD)

The STARCHILD model was one of the first attempts at developing a CPM. The activity generation module in STARCHILD is comprised of three steps. First, all feasible household travel activity patterns that are within defined temporal-spatial constraints are identified. Next, similar activity-travel patterns are identified using a pattern recognition technique and grouped in three to ten different groups. Lastly, a representative pattern engagement utility function in each group is recognized through the logit choice model. The scheduler module in STARCHILD is customized through a series of heuristic rules. These rules insert possible activities into individual daily agenda, with scheduling conforming to all constraints (Recker, McNally and Root 1986a; Recker, McNally and Root 1986b). The model's assumption that individuals conduct a comprehensive exploration for feasible patterns is questionable, and its integration with the population synthesizer module appear to be unclear.

SCHEDULER

The SCHEDULER model attempts to model activity choices, location, and travel in the individual's daily activity agenda. Activities in SCHEDULER are selected with the long-term calendar module and scheduling is carried out through a cognitive map. Activity attributes such as timing and utility are regularized in the long-term calendar. The

cognitive map comprises a set of heuristic rules that control for temporal-spatial constraints. Activities with the highest priority and duration are assumed to have more precedence compared to other activities in the scheduling process (Garling, Kwan and Golledge 1994). This model has been criticized because the assumption of activity priorities in the scheduling process may result in overestimating the occurrence of high priority activities.

Simulation Model of Activity Scheduling Heuristics (SMASH)

The SMASH model incorporates features of both discrete-choice modeling and CPMs in the activity scheduling process. Primarily, the pretip planning stage is explored, including activity selections, activity locations, activity start times, activity sequences, and travel modes to the various activity sites. The subjects' activity schedules are predicted based on their activity agenda and information about their spatio-temporal situations. The model estimates the utility functions for the choice rules, associated to inserting, deleting, or replacing activities. The model is tested under different settings with the activity schedules considered by the subjects themselves (Ettema, Borgers and Timmermans 1996). Further test should be performed in order to examine the computational efficiency of the SMASH model.

GIS-Interfaced Computational process model for Activity Scheduling (GISICAS)

The GISICAS model is an advanced version of the SCHEDULER model with two main improvements. Spatial data of home and work locations of population and potential destinations for non-mandatory activities are added to the model. Moreover, a set of spatial search heuristics are merged in the model in order to advance the neighborhood searching

process (Kwan 1997; Kwan and Golledge 1997). The fundamental goal of development of the GISCAS was to be a decision support system for advanced travel system information. Therefore, prediction of the individuals' daily activity-travel behavior is limited in the model context.

Activity Mobility Simulator (AMOS)

The AMOS model includes a policy sensitive loop that enables the model to modify existing activity-travel patterns according to the new situation. Alternative activities are generated from simulation, neural network, and time allocation models. Activity purposes, time budget, frequencies, location, and priority list are inter-connected in AMOS (Kitamura et al. 1996). The strength of this model is to forecast short-term responses to particular policy changes. On the other hand, the model requires detailed survey data comparable to a travel diary survey, which limits its application to a specific data set and problem.

A Learning-BAsed TRansportation Oriented Simulation System (ALBATROSS)

ALBATROSS may be regarded as the most comprehensive CPM model developed to date. The core of the model comprises of a decision-making heuristic procedure with learning systems. Fixed and flexible activities are defined. Initially, fixed activities are added into the scheduling model, along with their temporal and spatial information, into the agenda. Next, flexible activities are added into the agenda with respect to their order of preference and choice of time-of-day (Arentze and Timmermans 2000). ALBATROSS utilizes a machine learning technique known as a CHAID decision tree to predict temporal attributes of activities including start time and duration. The model relies on cross-sectional diary

data. Therefore, this model has been criticized for how to implement learning and adaptation attributes on the short-term basis.

Travel and Activity Scheduler for Household Agents (TASHA)

The TASHA model was developed with similar concepts to those used in SCHEDULER. Initially, activities with similar socio-demographic and temporal features are identified and grouped into different classes. This process is done using a series of empirical data analyses. Next, probability distribution functions for activity start time and duration in each class are computed. Several heuristic rules are defined in TASHA, and used for inserting, adjusting, and producing a completed schedule. Rules relating to activity priorities are used for conflict resolution in the model (Miller and Roorda 2003). TASHA also employs an econometric modeling technique for tour formation in the model. Adopting various sets of explanatory variables outcomes in various groupings and produces different probability distributions. Finding the best set of explanatory variables that have the best fit for classifying the population in the dataset is a challenging and time-consuming issue within TASHA.

Agent-Based Dynamic Activity Planning and Travel Scheduling Model (ADAPTS)

Both econometric modeling and rule-based techniques are incorporated in the ADAPTS model. A hazard-based formulation is employed for identifying activities with similar characteristics and producing activity temporal information for the scheduling engine. ADAPTS extends the conflict resolution rules in TASHA by considering more potential situations for conflicts between activities (Auld and Mohammadian 2009). Furthermore, ADAPTS is integrated with a dynamic traffic assignment procedure that allows for

rescheduling activities based on the new situations. The various computational based models have brought new insight into activity-based travel demand modeling, with substantial improvements in model structure and computational efficiency. However, these models have been criticized for incorporating the activity pattern generation module as an exogenous component in the modeling process (Ettema, Borgers and Timmermans 1996), which may impact the reproducibility of the model. An additional concern with models of this type is that they do not fully incorporate decision processes and behavioral mechanisms that lead to observed activity-travel decisions in the modeling process (De Palma et al. 2011).

2.2.3 Rule-Based Machine Learning Approaches

Over the last two decades or so, the use of emerging machine learning techniques in activity-based travel demand models has received much consideration and seen significant progress (Joh et al. 2002; Allahviranloo and Recker 2013; Hafezi, Liu and Millward 2017b; Hafezi, Liu and Millward 2017c). Machine learning employs algorithms that, without being explicitly programmed, mimic natural learning behavior to identify and differentiate complex patterns in data, and make a seemingly intelligent resolution (Bishop 2007). Machine learning techniques may be used to competently handle the various challenges in modeling different components of activity-based travel demand models, with the objective of boosting model accuracy by distinguishing complex data patterns. While these techniques are well known in the statistics fields and computer science, there have been only a few applications in activity-based travel demand modeling.

A recent application by Allahviranloo et al. in 2016 proposed a clustering technique called k-mean to group populations with similar daily activity patterns (Allahviranloo, Regue and Recker 2016). Their new technique is an alternative approach to traditional activity generation modules (using explanatory variables for grouping the population) in the overall Household Activity Pattern Problem (HAPP) modeling framework that was initially developed by Recker et al. in 1986 (Recker, McNally and Root 1986a; Recker, McNally and Root 1986b). As another example, Li and Lee (2017) employed a context-free grammars technique to generate a set of activities in individuals' daily agenda. Their approach was an alternative for replacing the predefined choice set for selection of daily activity patterns originally developed by Bowman and Ben-Akiva in 2001.

Other examples of machine learning application techniques in activity-based models include incorporating the AdaBoost algorithm in predicting temporal information of activities, and utilizing the support vector machine (SVM) in a daily activity sequence recognition process (Allahviranloo and Recker 2013). Missing values are automatically handled in the AdaBoost algorithm. Furthermore, variables do not require transformation, very few parameters need to be tweaked, and the algorithm doesn't overfit easily. On the other hand, the algorithm is sensitive to noisy data and outliers. The SVM algorithm is able to approximate complex nonlinear functions and automatically create nonlinear features. In contrast, interpretation is difficult when applying nonlinear kernels, and it takes a longer time to train compared to other algorithms.

2.3 Discussion and Conclusions

Urban development and rapid technological progress continue to increase the complexities in individual travel behavior. The latest generation of travel demand models, known as activity-based travel demand models, aim to more accurately forecast future travel patterns in a transportation system. To do so, they take as their starting point the fact that travel is derived from activity engagement, and then model activity engagement schedules for proxy individuals representing distinct population groups. Travel and location selections are restricted by certain time and space constraints. Activities for a continuous 24-hour period are estimated/predicted from socio-demographic characteristics of the proxy individual, including the household context, and from relevant land use and locational information.

In this study, we have overviewed the model structures of several computational based activity scheduling models. In addition, we have reviewed recent modeling developments employing machine learning techniques. Although various activity-based travel demand models have been developed and some of them are being implemented in practice (Vovsha, Petersen and Donnelly 2002; Outwater and Charlton 2006; VDOT 2009), there is undoubtedly substantial room for model improvement in terms of prediction accuracy, reproducibility, computational time, model structure, large scale operation capability, and performance at the household level. The discussion in this study confirms that during the last two decades significant progress has been made in improving many aspects of activity-based travel demand models by incorporating machine learning techniques in the modeling process. One fruitful avenue for future study may be to produce a more detailed overview

of machine learning based models. Based on a review of the models identified in this study, potential directions for future work are recommended as follows:

- Most of the current computational based activity scheduling models assume a priority order of activities in the scheduling process, that may result in overestimating the occurrence of high priority activities. Extra effort will be required in future work to modify this assumption in order to better replicate travel decisions processes.
- Many current models extract input information for modeling from the entire sample data without considering any segmentation of the population. This could potentially fail to capture important latent parameters operating on activity types and duration. Segmenting the input data into a number of homogeneous population groups is recommended for further investigations in order to gain a better capture these latent parameters.
- In most current models information about work activities is taken as the main or only input to the model. This may impact the dynamic activity planning aspect of activity-based models. Therefore, producing non-work activity in the agenda, and its specification as part of the inference procedure, is recommended for future work.
- The analysis unit in most previous models was the weekday. Modeling a weeklong scheduling period can help transport modelers to better understand variation in activity travel patterns of the population and result in improved scheduling prediction. Therefore, development of an integrated time-use data technique that

can efficiently, and without being intrusive to respondents, collect traveler information over a week is recommended.

- The use of many machine learning techniques, such as the random forest, fuzzy c-means clustering and the CART classifier algorithms, are still not explored for use in travel behavior analysis.
- Previous studies revealed the need for establishing sub-models for considerable sub-populations or special trip generators, such as large hospitals or large universities. Therefore, development of such sub-models in regional travel demand models is recommended for future work.

Chapter 3 Data and Methods

3.1 Scheduler for Activities, Locations, and Travel (SALT)

The overall goal of this study focuses on the development of the Scheduler for Activities, Locations, and Travel (SALT) disaggregated travel demand microsimulation model. The SALT modeling framework adopts the concept of activity-based travel demand modeling approaches and theories. In the initial stage, the SALT model system utilizes a pattern recognition approach that identifies population clusters with homogeneous time-use activity patterns. Later, a series of behaviorally realistic advanced econometric and rule-based models were developed for modeling behavioral mechanisms and time-use activity patterns for each identified cluster. As shown in Figure 3.1, the SALT conceptual framework consists of the following five major modules:

- Population synthesizer: This generates duplicates of sample households concerning the marginal data on individual and household attributes. These synthesized households are geographically located and spatially determined to represent the entire population of the study area.
- Time-use activity pattern recognition: At the core of the SALT modeling framework, this module identifies population groups with homogeneous daily activity patterns and mobility decisions. Each identified population cluster contains crucial information on people's activity patterns, such as activity type, timing, sequential arrangement of activities, and duration probability distribution.
- Tour mode choice: This module estimates primary tour destinations, number of tours per day, number of intermediate stops, and mode choices for shaping

individual daily activity patterns. Socio-demographic attributes, trip attributes, and land use characteristics are incorporated into the modeling process to understand daily tour better, stops generation and mode choice behavior.

- Activity destination choice: This module generates the daily activity agenda. It determines of the type of activities in the agenda, activities frequency, and their priority order in the sequence of activities.
- Activity/trip scheduling: This module estimates the timing and duration for every activity type in the agenda. Predicted activities are inserted into the skeleton schedule, and activities are scheduled according to their priority importance and empirical guide information gained from the representative activity pattern in each cluster.

Specifically, this study focuses on developing state-of-the-art machine learning and pattern recognition models to identify the representative time-use activity patterns, infer the scheduling behavior of individuals, enable the generation of a synthetic population for the region, and model the daily time-use activity patterns along with the estimation of emission factors for university commuters, considered as a special trip generator in regional travel demand models. Figure 3.2 illustrates the advanced ensemble learning modules incorporated in the SALT model system. In the following section, main sub-models developed in this study are overviewed.

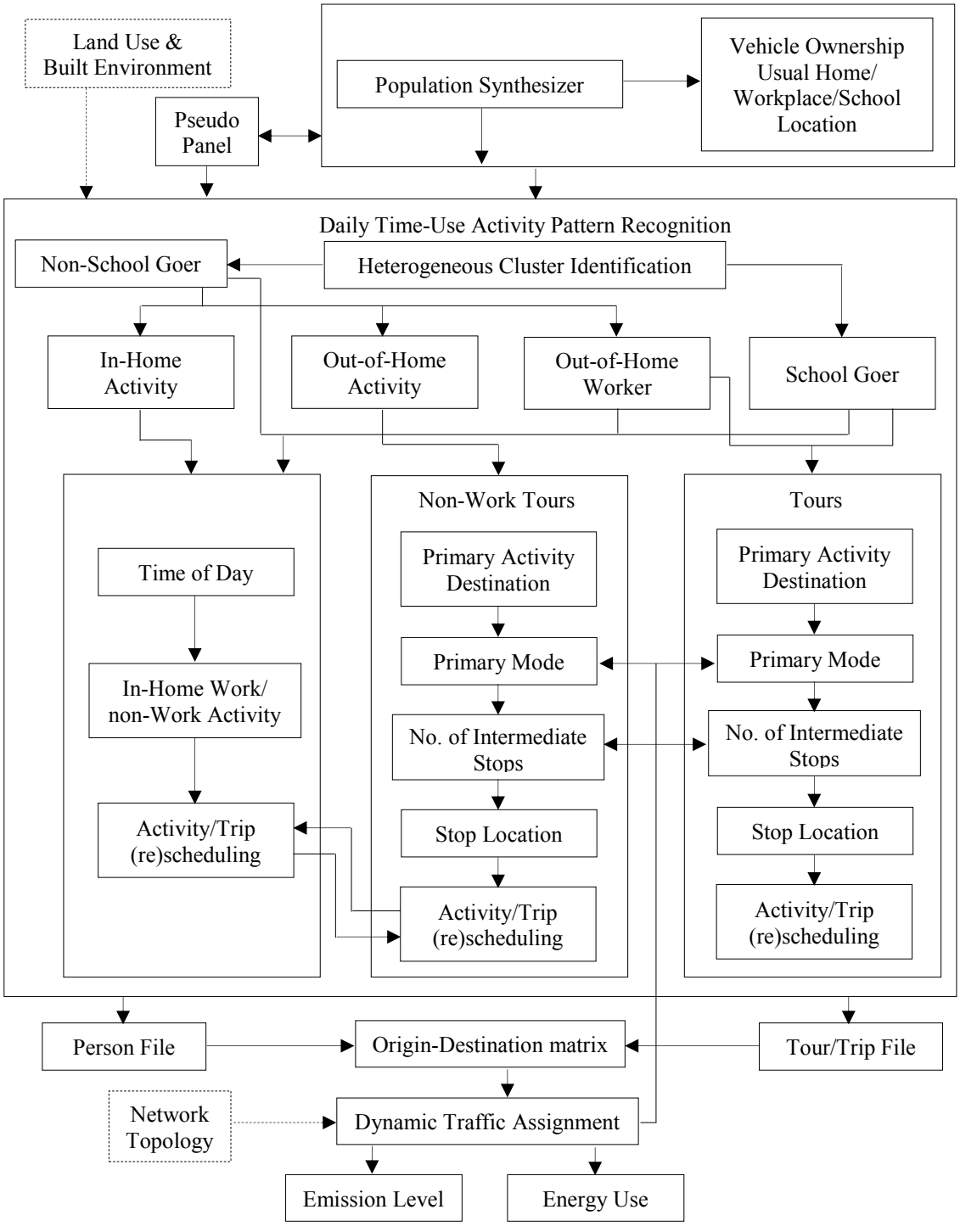


Figure 3.1 A conceptual framework of the Scheduler for Activities, Locations, and Travel (SALT)

3.2 Pattern Recognition

In this stage, a new and efficient pattern recognition modeling framework consisting of four sequential phases is developed, with the goal of producing population clusters with homogeneous activity patterns. Initially, a dynamic subtractive clustering algorithm is employed to initialize both cluster number and cluster centroids. Next, pattern complexity of activity sequences in the dataset is recognized using the Fuzzy C-Means (FCM) clustering technique. The clustering algorithm identifies individuals with homogeneous daily activity patterns and groups them into clusters. Next, a set of representative activity patterns are identified using the Multiple Sequence Alignment (MSA) technique. Finally, the cluster memberships are characterized through their socio-demographic attributes using the classification and regression tree (CART) classifier algorithm. The decision tree is able to properly recognize which cluster travelers belong to, based on the socio-demographic attributes of travelers. To implement the pattern recognition model, the 24-hour activity patterns are allocated into 288 three dimensional five minute intervals. The first dimension contains temporal information on activities, the second dimension contains socio-demographic characteristics associated to activities, and the third dimension comprises spatial information related with activities. Each interval comprises information on activity types, timing, location, and travel mode if relevant. Aggregated statistical assessment and Kolmogorov-Smirnov tests are utilized to assess and analyze clusters in terms of heterogeneous diversity of temporal distribution, and differences in a variety of socio-demographic variables. A time-use activity pattern recognition model produces crucial information such as activity type, activity start time, activity duration probability distribution, and sequential arrangement of activities, all of which are necessary for

modeling succeeding micro-behavioral modules in the SALT model. A detailed explanation of the pattern recognition model along with respective results are presented in Chapter 4.

3.3 Agenda Formation

The choice of daily activity sequences differs between individuals based on their socio-demographic characteristics and their health and/or mobility status. The aim of this stage is to provide an improved methodology for learning and modeling the daily activity engagement patterns of individuals using a state-of-the-art machine learning algorithm. The dependencies between activity type, activity frequency, activity sequence, and socio-demographic characteristics of individuals are taken into account by employing a Random Forest model. In order to capture the heterogeneity and diversity among the predictor variables, two different methods for split selection in the Random Forest algorithm are employed: CART and Curvature Search. These two methods are examined under two different layer settings. In the first setting, the algorithm grows trees using all alternative predictor variables, whereas in the second setting, the predictor variable's importance is estimated and then the algorithm grows trees using only high-score predictor variables. The estimation accuracy of the proposed models is evaluated using confusion matrix, transition matrix, and sequential alignment techniques. Ultimately, the inference model is able to learn and predict various aspects of the daily activity agenda, such as stop number, activity type, and activity sequential arrangement. A detailed explanation of the inference model along with respective results are presented in Chapter 5.

3.4 Scheduling

The aim of this stage is to develop a new modeling framework that is able to model and produce temporal information associated with the traveler's daily activity schedule for use in the SALT model. A scheduling model is developed using a precise and efficient machine learning technique known as Random Forest (RF). The RF model is formulated based on the socio-demographic characteristics of travelers and temporal features of their activities. Start time and activity duration for every activity type are allocated to a set of bins. Eight different bin structures, varying in their time interval, are designed as response variables. In addition, a heuristic decision rule-based algorithm is developed to build up the 24-hour activity schedule of population groups. Using a rule-based algorithm, the predicted activities are inserted into the traveler's skeleton schedule. An algorithm is then employed to schedule travelers' activities based on activity importance level and empirical guide information gained from the population cluster's representative pattern. A detailed explanation of the scheduling model along with respective results are presented in Chapter 6.

3.5 Population Synthesis

In this stage, a population synthesizer model is developed in order to generate a synthetic populations for the different geographical units in the study region. The population synthesizer model is used to replicate a sample of households with respect to data on their individual and household attributes. Individual and household sample data are drawn from the Public Use Microdata File (PUMF) and respective marginal data are drawn from the Canadian Census data. The synthetic algorithm is employed using three sub models: first,

using household level control tables (HL model); second, using individual and household level control tables (HPL model); and third, weighting individual and household level control tables (WHPL model). Error percentages and goodness-of-fit are used for validation of the model. The population synthesis model generates synthetic populations both at the regional and dissemination area (DA) levels. Furthermore, the model produces 100% synthetic populations for all four commuter groups at Dalhousie city campuses. A detailed explanation of the population synthesis model along with respective results are presented in Chapters 7 and 8.

3.6 Daily Activity Patterns and Pollution Estimation

The purpose of this stage is twofold. Firstly, it attempts to examine the activity engagement, and the sequencing and timing of activities, for student, faculty, and staff commuter groups at the largest university in the Maritime Provinces of Canada. Secondly, transport-related Greenhouse Gas (GHG) emissions based on the population characteristics and living zone in relation to campus areas are estimated. The daily activity patterns (DAP) of all university community groups is modeled using the CART classifier algorithm. In general, five zones are designated for emission estimation: on-campus zone (ONC), inner-city (INC), suburban-area (SUB), inner-commuter belt (IUB) and outer-commuter belt (OCB). Two emission scenarios in respect to changes in transit ridership and auto driving are investigated in order to demonstrate how changing the primary travel mode can impact emissions volume. The Motor Vehicle Emission Simulator (MOVES) 2014a is utilized as a simulation platform for estimating the major air pollutants, including carbon dioxide (CO_2), carbon monoxide (CO), nitrogen oxide (NO_x), particulate matter (PM_{10} and $PM_{2.5}$), total hydrocarbon (THC) and volatile organic compounds (VOC) for a

typical weekday. A detailed explanation of the modeling of university daily activity patterns, along with respective results and pollution estimation, are presented in Chapter 9.

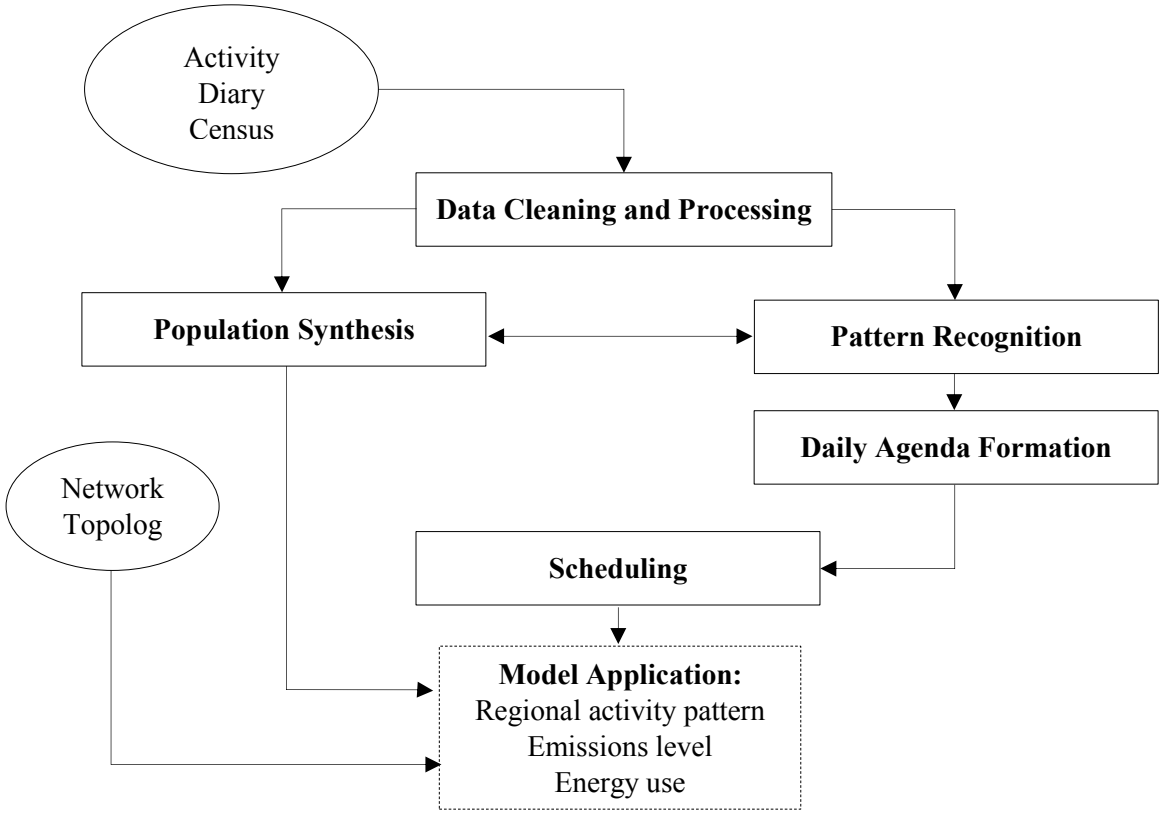


Figure 3.2 Ensemble learning modules incorporated in the SALT model system

3.7 Data

In this section, an overall description of the primary and secondary data sources for developing the micro-behavioral modules of the SALT model are presented. Table 3.1 presents a summary of data sources used for building the SALT modeling system. Numerous Big Data Analytics (BDA) and data mining techniques are used for data

screening and data processing in this study. Detailed data processing and data preparation steps are discussed in each relevant chapter.

3.7.1 Space-Time Activity Research (STAR)

This study utilizes time-diary and GPS geo-coordinate data as the primary data source, from the Space-Time Activity Research (STAR) survey undertaken in Halifax, Canada. The STAR survey represented the world's first large-scale application of global positioning system (GPS) technology for a household activity survey. The unique and rich Halifax STAR project produced a wide variety of data, including: (1) household roster data, (2) main file, (3) vehicle data, (4) time diary (episode and summary data file), (5) activity diary (episode data file), (6) land use database, (7) business hours survey data, (8) places and locations (PAL) directory data, and (9) global positioning systems (GPS) data. Full descriptions of the survey design and the socio-demographic features of respondents can be found in (TURP 2008; Millward and Spinney 2011).

The Halifax STAR project produced survey data from 1,971 randomly designated households in Halifax Regional Municipality (HRM) between April 2007 and May 2008. A primary respondent over age 15 was randomly selected in each household, completed a 2-day time diary, and carried a GPS unit (Hewlett Packard iPAQ hw6955) during all out-of-home activity. The respondent then completed a "day-after" de-briefing through at Computer-Assisted Telephone Interview (CATI), to verify activities, times, and locations, supplemented and verified through GPS tracking. The original STAR dataset comprised 188 activity sub categories defined under ten major activity classes. The activity codes employed were similar to those utilized in Statistics Canada's General Social Survey

(GSS) time-use surveys, and relate to the prime purpose of the activity. Among the available travel diary data in the study region, the Halifax STAR household survey provides the most accurate and GPS-validated time-use data for use in development of any disaggregated travel demand models, as of today.

3.7.2 Environmentally Aware Commuter Travel Diary (EnACT) Survey

In order to better understand and explore the travel behavior characteristics of university commuters in the SALT model, a unique web-based travel survey known as the Environmentally Aware Commuter Travel Diary (EnACT) survey was designed and implemented. The EnACT represents the first university-based travel diary survey in Canada. The questions and activity log were designed to be consistent with the Halifax Space-Time Activity Research survey (STAR) and Canadian General Social Survey (GSS) instruments. The EnACT survey was conducted in Spring 2016 at Dalhousie University, Nova Scotia. Dalhousie University is the largest university in the Maritime Provinces of Canada, with three urban campuses in the city of Halifax and one campus in the town of Truro.

All university populations groups, comprising undergraduate students, graduate students, faculty members, and staff were asked to complete a 24-hour travel log and also to provide detailed individual and household information. The EnACT survey includes six sections: (1) household information, (2) individual information, (3) environmental attitudes and behavior, (4) attitudes toward transportation, (5) information and communications technology (ICT) related information, and (6) a 24-hour travel log.

The survey was dynamically designed (including branching and piping options) in order to reduce response burden. For example, the vehicle type properties question was asked only to people who reported at least one vehicle. The EnACT survey collected detailed information on each respondent's home and work location; make, model and year of motorized vehicles used in their last commute to Dalhousie; travel mode; and average commuting travel time, which are all useful for transportation emissions calculations. Several survey recruitment methods were used, including sending e-mails through university administration to Dalhousie commuters containing login information to access the survey online, promoting the survey in social media (Facebook and Twitter), and distributing survey posters across campuses. From a total of 840 respondents, 570 respondents completed all the sections except the 24-hour travel log, and a total of 364 respondents completed all six sections of the survey. Of these responses, 40.1% are undergraduate students, 34.4% are graduate students, 6.4% are faculty members, and 19.1% are staff. A comprehensive descriptive analysis of all six sections of the EnACT survey can be found in Liu et al. 2016 and Hafezi et al. 2018a.

3.7.3 Other Data Sources

Secondary data sources used in this study were: the 2006 and 2011 Census and Public Use Microdata File (PUMF) derived from Statistic Canada, Halifax Regional Municipality (HRM) database 2012, Environment Canada archive, the 2015 Canadian Vehicle Survey and GeoBase - National Road Network and Environment Canada archive.

Table 3.1 Data sources used in the development of the SALT model system

Data Objects	Data Sources	Data Descriptions	Unit	SALT's Micro-Module(s)
Representative time-use microdata sample at the household level	STAR[1]	Time-diary and GPS geo-coordinate microdata sample data Land use and built environment data	Three dimensions (temporal, socio-demographic and spatial) data with five minutes intervals Parcel level	Pattern recognition Ensemble learning Activity participation Trip Chaining Tour mode choice
Representative time-use microdata sample of university population (undergraduate student, graduate student, staff and faculty)	EnACT[2]	Time-diary microdata sample	Three dimensions (temporal, socio-demographic and spatial)	Travel behavior characteristics of university community Transport-related GHG emissions
Time-use microdata sample at the individual level	GSS[3]	Time-diary microdata sample	Activity duration Episode duration	Synthetic pseudo panel
Microdata sample of the population at the individual and household levels	PUMF[4]	Socio-demographic characteristics of a random microdata sample	Dissemination area (DA) level Regional level	Population synthesis
Marginal population data at the DA and regional level	CCS[5]	Distribution of the socio-demographic characteristics of the marginal population data	Dissemination area (DA) level Regional level	Population synthesis
Road network	NRN[6]	National road network layer in the ArcGIS platform	Street level Highway level	Network building Transport-related GHG emissions
Vehicle characteristic data	CRV[7]	Vehicular age distribution and fuel characteristics	Vehicle type	Transport-related GHG emissions
Meteorological data	ECA[8]	Humidity and temperature data	Regional level	Transport-related GHG emissions
Transport service location and road network	HRM[9]	Transit stop locations, Transit and road networks	Street level	Network building Transport-related GHG emissions

[1]Space-Time Activity Research, [2]Environmentally Aware Commuter Travel Diary Survey, [3]General Social Survey, [4]Public Use Microdata File, [5]Canada's Census, [6]National Road Network and Environment Canada archive, [7]Canadian Vehicle Survey, [8]Environment Canada Archive, [9]Halifax Regional Municipality Geodatabase

Chapter 4 A Time-Use Activity-Pattern Recognition Model for Activity-Based Travel Demand Modeling[2]

4.1 Introduction

In recent years, disaggregate travel demand models have begun to be employed for travel demand forecasting purposes. These models improve upon the traditional four stage modeling method, since they are able to more accurately capture the effects of elements that influence travel behavior and time allocation, such as socio-demographic attributes. Latterly, the activity-based modeling approach, along with other disaggregate travel demand modeling methods such as trip-based modeling, has become more popular and commonly used in both the academic and practitioner sectors. To date, numerous activity-based models have been developed, such as STARCHILD (1986), and SCHEDULER (1989), ALBATROSS (2000), DAYSIM (2001), MORPC (2002), TASHA (2003), and CEMDAP (2004). Activity-based models emphasize that travel is a derived demand, originating from the need of an individual to participate in activities. Activity-based models work on constructing the 24-hour activity schedule and the associated travel linked with activities performed by individuals. Most activity-based models comprise the following universal modules: activity generator and scheduler, tour and trip time of day, tour and trip mode choice, tour and trip destination choice, and network assignment.

For many years, researchers and practitioner have employed different approaches for development of activity-based models. Rasouli and Timmermans reviewed recent work on

[2] A version of this chapter has been published:
Hafezi, M. H., L. Liu., and H. Millward. (2017). "A time-use activity-pattern recognition model for activity-based travel demand modeling". Transportation. 1-26. DOI: 10.1007/s11116-017-9840-9.

activity-based models of travel demand (Rasouli and Timmermans 2014), and argued that these models can be classified into three main concept categories: (1) constraint-based models, (2), discrete choice models, and, (3) computational process models. The constraint-based models consider possible travel patterns with respect to a set of space-time constraints. PESASP (1977), GISICAS (1977), and CARLA (1985) are some examples of activity-based models developed through the constraint-based modeling approach. The second category, discrete choice models (also known as econometric models), consider activity pattern consequences from utility maximizing decisions. DAYSIM (2001), MORPC (2002), and CEMDEP (2004) are some examples of activity-based models developed through the econometric approach. Finally, the computational process models (also known as rule-based models) simulate and model activity patterns through computational processes. SCHEDULER (1989), ALBATROSS (2000), and TASHA (2003) are some examples of activity-based models developed through the rule-based modeling approach.

Some researchers have incorporated both econometric and rule-based modeling approaches in the activity-based modeling framework. The reason was to increase the model computational efficiency and degree of accuracy of outputs. In particular, Auld and Mohammadian (2009) employed a rule-based technique to resolve conflicts in the activity scheduling phase in the ADAPTS econometric activity-based model. Another example is the TASHA model, where Miller and Roorda (2003) for tour formation of their rule-based model by utilizing an econometric modeling approach. In very recent research, borrowing from the computer science field, researchers have used machine learning techniques to develop different components of activity-based models. However, there have been very

limited applications of such techniques in activity-based modeling. For instance, the K-means clustering technique has been used in a pattern-recognition modeling framework (Jiang, Ferreira and Gonzalez 2012; Allahviranloo, Regue and Recker 2016) and support vector machine (SVM) has been used in a daily activity sequence recognition process (Allahviranloo and Recker 2013).

In this study we develop a new solution method for the activity generation module in activity-based travel demand models. We build on progress in activity generation modules by developing a new comprehensive pattern recognition modeling framework which leverages activity data to derive clusters of homogeneous daily activity patterns. Each cluster produces vital information such as activity type, start time, end time, duration probability distribution, and sequential arrangement of activities. Our prime contention is that generating more accurate activity patterns is a significant step in decreasing uncertainty in forecasting the individual's activity engagement decisions and moving current activity-based models closer to replication of reality. Three-dimensional five-minute intervals are used as the basic analysis unit in this study. Several machine learning techniques not previously employed in travel behavior analysis (fuzzy c-means (FCM) clustering algorithm and the CART classifier) are employed in the pattern recognition framework. This study contributes by providing additional insights to the linkage between activity generation and activity scheduling modules in the overall activity-based travel demand modeling framework. Furthermore, the proposed modeling framework in this study may be applied to any applications that contain a group of linked sequences, such as day-to-day variations in transit ridership or station demand at the individual level. Finally, the results of this study are expected to be incorporated within the activity-based travel

demand model, Scheduler for Activities, Locations, and Travel (SALT) for Halifax Regional Municipality (HRM), Nova Scotia, which is currently under development.

The remainder of the study is structured as follows: first, we provide a review of relevant past research concerning activity generation modules in activity-based modeling framework. Secondly we discuss the data used and data transformation necessary for pattern recognition, followed by presentation of the pattern recognition methods and discussion of model results. The study concludes by providing a summary of contributions and future research directions.

4.2 Literature Review

Activity generation modules can play an important role in every activity-based modeling framework. Prediction accuracy of individual travel behavior depends on actual information drawn from activity generation modules. Thus, producing more accurate and homogeneous information from this module will result in increasing prediction accuracy in activity-based travel demand modeling. Since 1970, when Hagerstrand developed the preliminary activity-based model (Hagerstrand 1970), researchers have used several different approaches for the activity generation modules of activity-based models, such as empirical data analysis, decision trees, and hazard functions (Recker, McNally and Root 1986a; Recker, McNally and Root 1986b; Arentze and Timmermans 2000; Miller and Roorda 2003; Auld and Mohammadian 2009).

The activity generation module in the ALBATROSS (A Learning-BAsed TRansportation Oriented Simulation System) model consists of an integrated decision-making heuristic with learning mechanisms (Arentze and Timmermans 2000). Initially, activities are

classed into two sets: fixed and flexible activities. The model then produces the fixed activities and relocates them along with their temporal and spatial information to the scheduler module of the algorithm. The flexible activities are added to schedules based on their order of priority and choice of time-of-day. A hazard-based formulation is employed in the activity generation module in the ADAPTS (Agent-Based Dynamic Activity Planning and Travel Scheduling) model (Auld and Mohammadian 2009). Activities with similar features are recognized and essential information such as activity start times and durations are generated. The activity generation module in the STARCHILD (Simulation of Travel/Activity Responses to Complex Household Interactive Logistic Decisions) model comprises three successive phases (Recker, McNally and Root 1986a; Recker, McNally and Root 1986b). At the beginning, the algorithm generates all the feasible travel activity patterns. Next, all possible activity-travel patterns are found and grouped. Lastly, representative patterns in each group are identified using the logit choice model. A series of empirical data analyses are employed in the activity generation module of the TASHA (Travel and Activity Scheduler for Household Agents) model (Miller and Roorda 2003). Populations are groups with similar distributions of start time, activity type, and frequency, based on different sets of explanatory variables such as age, gender, income, and occupation.

One limitation of using explanatory variables for grouping the population is that using different sets of explanatory variables results in different groupings and generates different probability distributions. Therefore, it is challenging and time consuming to find which set of explanatory variables have the best fit for grouping the population in the dataset. As activity generation modules have a direct effect on prediction accuracy, it is important that

populations are grouped with the most similar characteristics. In recent years, machine learning techniques have brought new insight into the modeling process of different components of activity-based modeling. For instance, Jiang et al. 2012 and Allahviranloo et al. 2016 employed the K-means clustering technique for activity pattern recognition. Another example is the application of the SVM technique in activity scheduling process (Allahviranloo and Recker 2013). Machine learning is a computationally fast and straightforward reproducible technique that without being deliberately programmed is able to naturally learn to recognize complex patterns, and make an intelligent resolution based on the trained data (Bishop 2007; Kubat 2015). While machine learning is well known in the computer science field, nevertheless there have been only limited applications of the technique in activity-based travel demand modeling, and mainly in the activity generation process. For instance, machine learning techniques can be used to solve a sequence alignment problem (Joh et al. 2002). In the following section some of these research efforts are overviewed.

Through aggregation of statistical learning methods and data mining, Jiang et al. (2012) proposed a new modeling framework for clustering daily patterns of individual activities. Numerous machine learning techniques such as the K-means clustering algorithm and the principal component analysis (PCA) were employed to explore the inherent daily activity structure, and populations were clustered based on the similarity of their activities. The modeling framework was implemented for both weekday and weekend data. Their research findings enhance the traditional population divisions into workers, students, and non-workers. In their proposed modeling framework, individuals are clustered based on their activity similarity rather than by explanatory variables such as age or occupation.

Liu et al. (2015) employed profile Hidden Markov Models (HMM) to augment the sequence alignment method (SAM). Their argument is that the SAM alone cannot capture infrequent activities and their related travel episodes. Consequently, they added a supplementary phase to the SAM by converting multiple alignments into a position-specific counting system to capture the probability of all infrequent and frequent activities in the data. A two-stage clustering technique to infer activity time windows was developed by Allahviranloo et al. (2016). Activity pattern recognition is accomplished using aggregation of K-means clustering and affinity propagation methods, in order to capture both frequent and infrequent activities. They extended their work to discover differences between activity patterns by employing the SAM and agenda dissimilarity distance measurement methods. They found that the scheduler executed better when it used the clustered data compare to un-clustered data. Their proposed method clustered populations into eight clusters. In another study, Li and Lee (2017) utilized probabilistic context-free grammars in the modeling and learning of daily activity patterns. Saneinejad and Roorda (2009) measured similarities between routine weekly activity sequences by utilizing the multiple sequence alignment methods.

This study addresses the above-mentioned limitations of activity generation modules by using explanatory variables for grouping the population and by recognizing infrequent activities in the overall activity-based modeling framework. In this study, we tackle the problem from a new standpoint, through development of a new comprehensive pattern-recognition modeling framework that leverages activity data to derive clusters of homogeneous daily activity patterns. Each particular cluster produces essential information such as activity type, start time, end time, duration probability distribution,

and sequential arrangement of activities. Application of this new framework to activity-based modeling not only reveals the strength of machine learning to identify homogeneous clusters, but also yields additional insights into the linkage between two critical activity-based model modules namely activity generation and activity scheduling.

4.3 Data

In this section, an overall description of the Space-Time Activity Research (STAR) survey data and data processing steps is presented. This study uses time-diary and GPS geo-coordinate data, from the STAR survey accomplished in Halifax, Canada. The STAR survey was a combined household activity survey and travel survey, and the world's first large-scale employment of global positioning system (GPS) technology for tracking and verification of out-of-home activities. A brief description follows, and full descriptions of the survey design and the socio-demographic characteristics of respondents can be found in (TURP 2008; Millward and Spinney 2011).

The Halifax STAR project collected survey data from 1,971 randomly selected households in Halifax Regional Municipality (HRM) between April 2007 and May 2008. The survey collected fully geo-referenced 2-day (i.e. 48-h) time diary data from a randomly selected primary respondent aged 15 years or older within each household. Primary respondents carried a GPS data logger (Hewlett Packard iPAQ hw6955) for a 48-hours reporting period, maintained a daily ''activity log'' during that period, and completed a computer-assisted telephone interview (CATI) time-diary survey the day after the two-day reporting period had ended. The respondents' descriptions of their out-of-home activities were prompted and validated by the GPS data.

The original STAR dataset included 188 activity sub categories defined under ten main activity categories. The activity codes utilized were those employed in Statistics Canada's time-use surveys, and relate to the prime purpose of the activity. In addition, entertainment activities were defined by passive attendance, whereas both sports and hobbies were identified by active participation in the activity.

4.3.1 Data Processing

The data processing consisted of three steps. The first step was to identify and eliminate data for non-working days from the STAR survey data. The second step was to clean the database of any missing values to ensure validity, uniformity, and consistency. The resulting data set comprised 2,778 working person-days (1,389 individuals, two days each). Note that the current research did not consider the possible temporal correlation of activity sequences between two continuous days, and treated all person-days as independent samples. As mentioned earlier, a five minutes interval was used as the basic time unit in this study. Therefore, we rounded all time values up/down in the way so that they were evenly divisible by five. Lastly, the final step was to re-categorize the original 188 activity categories. To align with the transportation planning literature and urban studies (Ben-Akiva and Bowman 1998b; Bhat et al. 2004), and based on similarities between some of the primary activities, the original activities were aggregated into 9 activity categories, as shown in Table 4.1. Travel episodes are categorized as a separate time-use category in this study. This feature allows the model to be updated with new data on congested travel times. The new travel times may be measured by operating the activity-based travel demand model in aggregation with a congestion index. In addition, for the purpose of this study, we categorized in-home activities into three major categories.

Table 4.1 Proposed cluster-based codification for activity episodes

Activity code	In-home/ Out-of-home	Aggregated activity categories	Code	Descriptions
1	In-home activities	Home chores	H	Working at home, eating/meal preparation, indoor or outdoor cleaning, interior or exterior home maintenance, child care, other in home activities.
2		Home leisure	L	Watching TV/listening to radio, reading books/newspapers, etc.
3		Night sleep	N	Night sleep
4	Out-of-home activities	Workplace	W	Work/job, all other activities at work, work related (conferences, meetings, etc.).
5		Shopping & services	P	Shopping for goods and services, routine shopping.
6		School/college	S	Class participation, all other activities at school.
7		Organizational/hobbies	G	Organizational, voluntary, religious activities. Hobbies done mainly for pleasure, cards, board games, all other hobbies activities.
8		Entertainment	E	Eat meal outside of home, all other entertainment activities.
9		Sports	T	Walking, jogging, bicycling, all sports related activities.

4.3.2 Data Transformation

Prior to implementing pattern recognition techniques, it was essential to transform the activity survey data. The data transformation process is illustrated in Figure 4.1. The twenty-four hours were split into 288 five minute intervals, and each interval has 3 dimensions. The first dimension comprises temporal information on activities: each of the 288 cells was coded with one of the 9 major categories as defined in Table 4.1. The second dimension comprises socio-demographic characteristics related to activities, and the third dimension contains spatial information associated with activities. In the current study, we have utilized only the temporal dimension in the modeling framework. However, the other dimensions can also be used for cluster analysis. For example, the spatial information

dimension can be used to identify if there are differences in clusters in terms of daily spatial dynamics (e.g. home-work distance).

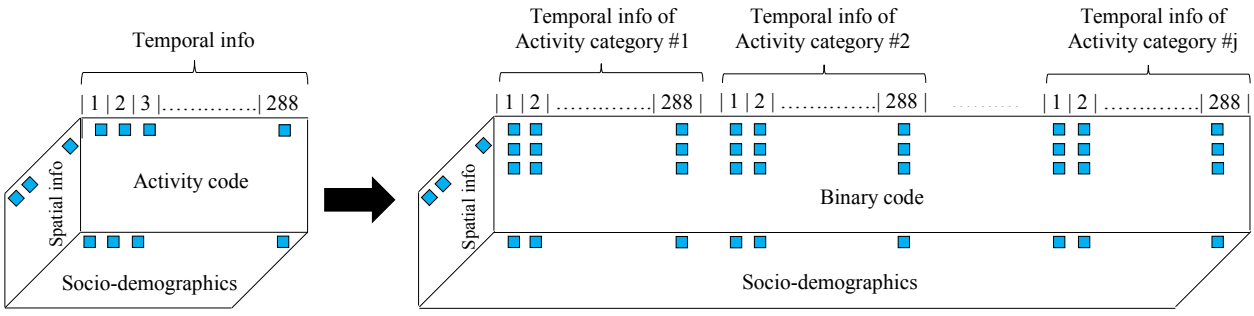


Figure 4.1 Database schema transformation

Consequently, the preliminary transformed matrix dimensions would be 288 x 2,778 x 2,778. The ultimate step in data transformation process was to transform activity survey data to the binary format. Each of the 9 major activity categories were transformed to a "1" or "0" binary code such that if the individual participated in the activity in the particular time interval a code of "1" was recorded, and otherwise "0".

4.3.3 Individual and Aggregated Daily Activity Dissimilarities

Figure 4.2 shows the aggregated daily temporal pattern of 2,778 person-days (1,389 individuals, two days each) with their activities categorized in 9 groups at the aggregated level. The 288 five-minute intervals started at 4:00 am and finished at 3:55 am on the next day. The temporal distribution of in-home and out-of-home activities in Figure 4.2 show very interesting and informative information about household daily activity pattern. In particular, we see that household chores have morning and early evening peaks (breakfast and supper), and household leisure is high in the later evening. Workplace and household chores dominate during the daytime, and are inversely related.

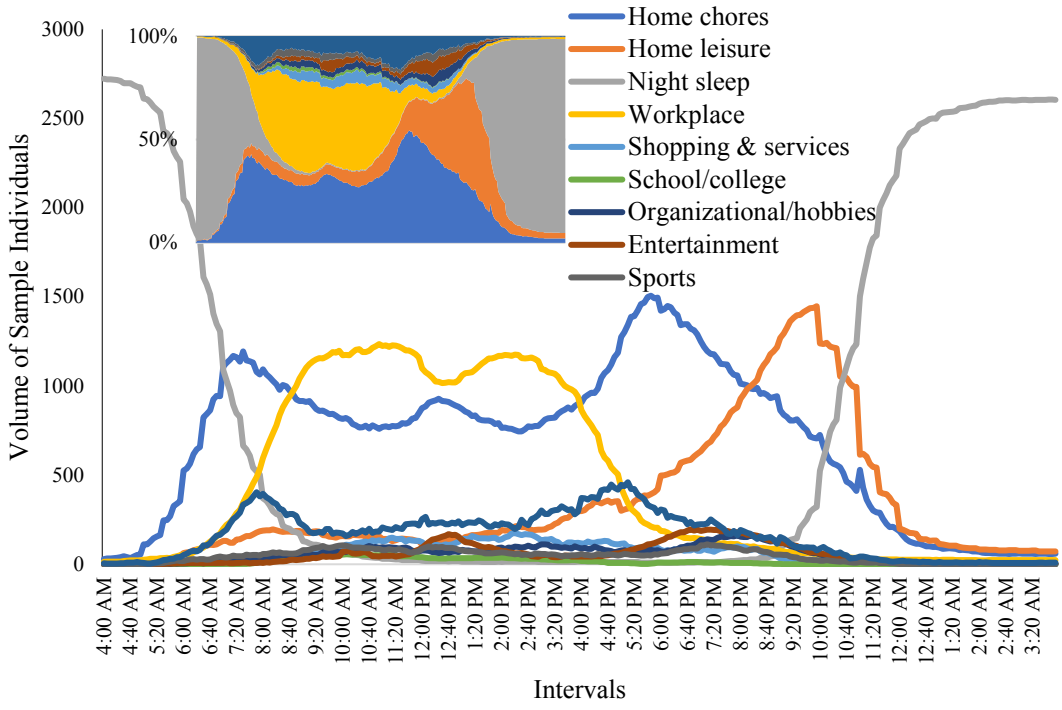


Figure 4.2 Aggregated temporal pattern of person-day activities

Figure 4.3 visualizes sparsity patterns of person-day activities. The spots in Figure 4.3 show the transformed data using 5 minutes intervals. For each activity type, the darker area indicates that the individual participated in the activity (code "1"), and conversely the brighter area indicates that the individual did not participate in the activity (code "0). The horizontal axis denotes time of day starting from 4:00 a.m. and ending at 3:59 a.m.

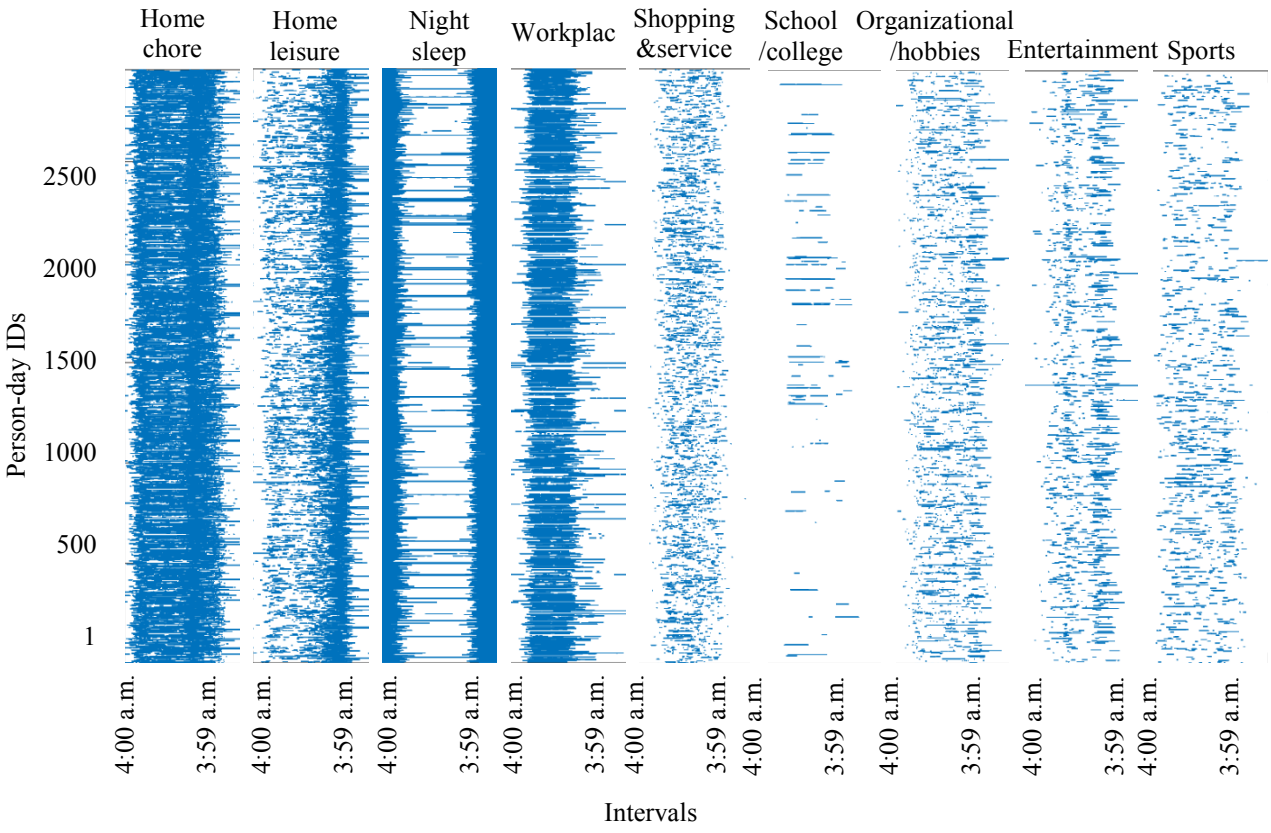


Figure 4.3 Sparsity pattern visualization of person-day activities

4.4 Methods

The pattern recognition modeling framework in this study comprises four modules, as follows. First, we employed a subtractive clustering algorithm for initializing the total cluster number and cluster centroids. Previous studies (Pena, Lozano and Larranaga 1999; Erisoglu, Calis and Sakallioglu 2011) suggested initializing these two values before implementing any clustering algorithm such as K-means or C-means, in order to increase the performance of the main clustering algorithm. We used Dunn's index to measure cluster validity. Next, individuals with similar activity patterns were identified and clustered using the FCM clustering algorithm. Using the multiple sequence alignment method (M-SAM), the sets of representative patterns were achieved. We incorporated the progressive method to calculate the number of steps needed to align multiple sequences.

Finally, the CART classifier algorithm was used to explore inter-dependencies among the attributes in each identified cluster, and to relate the membership of cluster individuals to their socio-demographic characteristics. Broadly, we group the activity sequences into different clusters using FCM and analyze the relationship between demographic features of persons and identified clusters using CART, with an assumption that activity sequence clusters are dependent on demographic features. Although the SAM is commonly used in activity sequence analysis, to the best of our knowledge the FCM clustering has not been explored in activity pattern or travel behavior studies (Hafezi, Liu and Millward 2017b). Further study will include extending the current modeling framework to produce detailed information on activities, such as start time, duration, activity type, travel distance, and location, that are crucial for the scheduling stage of activity-based travel demand modeling.

4.4.1 Initialization of Cluster Number and Cluster Centroids

The first step in the proposed pattern recognition modeling framework is to initialize both cluster number and cluster centroids. For this purpose, a dynamic subtractive clustering algorithm is implemented. The algorithm searches for cluster centers based on the density of neighboring data points. Overall, the subtractive clustering algorithm consists of five phases. In this study, we present only a concise overview of the algorithm, and interested readers are referred to (John Lu 2010; Ngo and Pham 2012; Shieh 2014) for more explanation. The transformed temporal information on activities achieved in section 3.2 is used as input of the subtractive clustering algorithm. The parameter z is defined as the sample size of person-days in the dataset, 2778. For each individual in the population $i \in \{1,2,3,\dots,z\}$ there are 2,592 (9 activity categories x 288 time-intervals) data points $P =$

$\{p_1, p_2, p_3, \dots, p_z\} \in \{0,1\}^z$ such that each p_i has 2,592 dimensions. Each data point represents a transformed person-day activity pattern. The subtractive clustering algorithm begins by initializing the accept ratio $(\bar{\partial})$, reject ratio $(\underline{\partial})$, cluster radius (u_r) and squash factor (ϑ) parameters. These parameters have important effects on finding cluster centers and total cluster number in the database. The u_r is defined as a positive value demonstrating a neighborhood radius. A larger value of u_r results in finding fewer cluster numbers whereas a smaller value of u_r can results in model overfitting. The suggested value for ϑ is $1.25 \leq \vartheta \leq 1.0$ and for u_r is $0.15 \leq u_r \leq 0.30$ (Ngo and Pham 2012). The next step in the subtractive clustering algorithm is to calculate density for all data points.

$$T_i = \sum_{j=1}^{m} e^{-\frac{4}{u_r^2}\|p_i - p_j\|^2} \quad (1)$$

Using the Euclidean distance method, the distance between two data points is computed. In other words, the distance indicates the extent of differences between two person-day activity sequences. The algorithm continues by searching among computed densities for all data points, and the data point (p^*) with highest density (T^*) is designated as the initial cluster center. Next, the algorithm recalculates the density of all data points using the difference between the highest selected density in the last step and the new computed density.

$$T_i = T_i - T_h^* e^{-\frac{4}{\vartheta * u_r}\left\|p_i - p_j^*\right\|^2}; i = 1, \dots, z \quad (2)$$

If $T > \bar{\partial} T^{ref}$ then p^* is nominated as a new cluster center. Otherwise, E_{min} is computed as the shortest distance between p^* and all previously found cluster centers. The process

of finding a new cluster center is continued if $\frac{E_{min}}{u_r} + \frac{T^*}{T^{ref}} \geq 1$. If not, then $T(p^*) = 0$ and w^* is designated with the following maximum density. The algorithm is terminated when $T^* < \underline{\partial} T^{ref}$. Considering the set of cluster centers, the membership degree of data points in each cluster is computed as follows:

$$\beta_{ih} = e^{-\frac{4}{u_r^2}\|p_i - p_h\|^2} \tag{3}$$

The cluster number and cluster centroids identified through the subtractive clustering algorithm are used as inputs for the Fuzzy C-means (FCM) algorithm in the next step of the pattern recognition modeling framework. The FCM algorithm determines the final memberships in each cluster through a fuzzy process.

4.4.2 Identification of Individuals with Homogeneous Activity Patterns

The second step in the proposed pattern recognition modeling framework is to identify individuals with homogeneous activity patterns and group them into clusters. For this purpose, the Fuzzy C-Means (FCM) unsupervised machine learning algorithm is employed. In the FCM each data point that represent a person-day activity has the likelihood to belong to several clusters. This aspect of the algorithm boosts the cluster quality by selecting the best fitted data points. The FCM algorithm uses an iterative process in which the degree of membership for each data point in the cluster is computed at each iteration, and subsequently this information is utilized in updating the cluster membership and cluster centroids in the following iterations. The FCM algorithm is terminated when a termination condition is met. The FCM algorithm employs the following steps:

Initialize membership $R^{(0)} = [k_{ih}]$ for data point p_i (person-day activity) of cluster g^* by randomly choosing membership of all clusters. At the t-th phase, calculate the fuzzy centroid $F^{(t)} = [f_l]$ for $l = 1, \dots, g$, where g is the number of clusters obtained from the previous step.

$$f_l = \frac{\sum_{i=1}^{z} (h_{il})^{\tau} p_i}{\sum_{i=1}^{z} (h_{il})^{\tau}} \tag{4}$$

where h_{il} is the degree of membership of data point p_i in the l^* cluster, τ is the fuzzy parameter and z is the number of data points (2778 person-days). The fuzzy membership h_l is updated as follows:

$$h_l = \frac{1}{\sum_{k=1}^{g} \left(\frac{\|p_i - h_l\|}{\|p_i - h_k\|} \right)^{\frac{2}{(\tau-1)}}} \tag{5}$$

Minimize the following objective function:

$$N_\tau = \sum_{i=1}^{z} \sum_{h=1}^{g} h_{il}^{\tau} \|p_i - h_l\|^2 \tag{6}$$

The updating algorithm is terminated when $\left\| N^{(\tau)} - N^{(\tau-1)} \right\| < \varphi$. The parameter φ is specified as the minimum threshold in the algorithm. The final membership of cluster f_l is obtained as follow:

$$p_i \in f_l \leftrightarrow {h_l}^{N^{(\tau)}} > \beta \tag{7}$$

For each data point p_i, assign p_i to cluster f_l if fuzzy membership h_l of $N^{(\tau)}$ is greater than threshold value β. Activity sequences belonging to members in each identified cluster are used as input for the sequence alignment method (SAM) in the next step of the pattern

recognition modeling framework in this study. The SAM algorithm measures distance between activity sequences based on the number of stages needed to align two sequences of activities.

4.4.3 Identification of Sets of Representative Activity Patterns

The third step in the proposed pattern recognition modeling framework is to identify the sets of representative activity patterns. For this purpose, the Multiple Sequence Alignment (MSA) method is employed. The sequence alignment method is commonly used in the biological sciences to compare strings of chromosomes. One of the main challenges of the sequence alignment problem is to compute the required number of stages in order to align two strings (Chenna et al. 2003). This problem can be solved using various methods such as heuristic methods, approximation algorithms, probabilistic methods, and global optimization. In this study, a new heuristic method is used for solving the sequence alignment problem. This method, named the progressive alignment technique, is composed of three phases. At the beginning, for all existing pairs of sequences in each cluster, pairwise distance scores are computed. Next, a guide tree based on the calculated similarity sequences is produced and similar sequences are assigned close together in the guide tree. Finally, the sequences are aligned according to training collated by the guide tree. Supposing that cluster g has l membership: $g_l \in \{g_1, g_2, g_3, \dots, g_l\}$, the objective is to calculate the optimal alignment for every member of cluster g. The optimal alignment is accomplished through the distance score matrix. The distance score between two members of cluster g is computed as follows:

$$SC = 1 - \frac{r}{s} \tag{8}$$

Where r is the number of corresponding strings g_i and g_j in a bit string, and s is the number of corresponding lengths of d_{g_i} and d_{g_j} in a bit string.

The score for two matched strings is equivalent to $+1$, the penalty for two mismatched strings is $-\rho$ and the gap penalty is $-\sigma$. The guide tree is constructed according to the distance score matrix. Finally, the representative pattern will be achieved through execution of several alignments, including insertion, deletion, and substitution of the entire cluster membership. The edit-distance between two strings g_i and g_j with lengths of d_{e_i} and d_{e_j} is computed as follows:

$$T_{i,j} = max \begin{cases} T_{i-1,j-1} + 1 & if\ g_i = g_j \\ T_{i-1,j-1} - \rho & if\ g_i \neq g_j \\ T_{i-1,j} - \sigma & \\ T_{i,j-1} - \sigma & \end{cases} \tag{9}$$

The $T_{i,j}$ is an alignment with maximum score. In order to understand the relationship between demographic features of persons and identified clusters, we then analyzed each cluster using a CART classifier is used as input in the CART classifier.

4.4.4 Investigation of Inter-Dependencies among the Attributes

The last step in the proposed pattern recognition modeling framework is to discover the inter-dependencies among the socio-demographic attributes of persons in each identified cluster, with an assumption that cluster membership is dependent on demographic features. In doing so, our approach is comparable to recent work by (Jiang, Ferreira and Gonzalez 2012). For this purpose, the CART classifier is performed, to construct the best-fitting decision tree that contains the highest amount of information. Consistent with other

decision tree algorithms such as C4.5, CHAID, and ID3, in the CART algorithm the impurity measure is a decision maker for seeking leaf nodes (Tan, Steinbach and Kumar 2005). The CART algorithm utilizes the Gini index to measure the impurity. The Gini index is calculated for every predictor variable at each node, and the variable that has the minimum value is chosen. In addition, the CART algorithm employs cross-validation as a complementary measurement to choose the optimal decision tree. The Gini index is calculated as follows:

$$I(G) = 1 - \sum_{j=1}^{n} f_j^2 \tag{10}$$

Where:

n is the number of activity categories (9 activity categories as defined in Table 4.1),

f_i is the relative frequency of activity j in the cluster g.

4.5 Discussion of Results

We applied the proposed pattern recognition modeling framework to data associated with 2,778 person-day (1,389 individuals, two days each) drawn from the 2008 Space-Time Activity Research (STAR) travel survey (TURP 2008) in Halifax, Nova Scotia, Canada. The FCM clustering method bundled individual activity patterns into twelve discrete clusters. The Dunn's index showed 12 to be the best number of clusters. The temporal pattern of individual activities for the twelve identified clusters is shown in Figure 4.4, and Table 4.2 presents an analysis of clustered data. In the following section, a discussion for each of the twelve clusters and their socio-demographic attributes is presented.

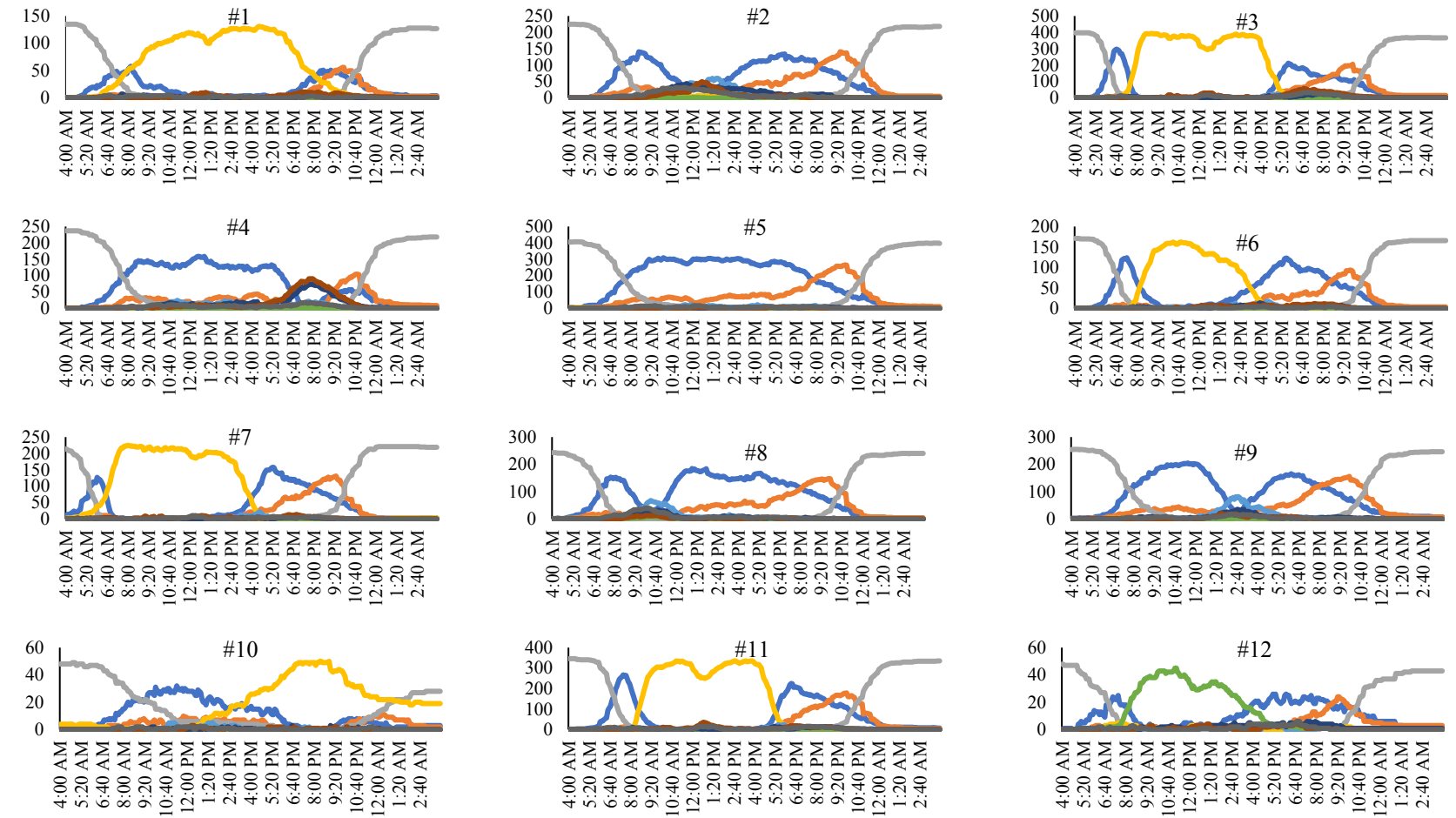


Figure 4.4 Temporal pattern of person-day activities for twelve identified clusters

Table 4.2 Analysis of clustered data: Share of different socio-demographic variables, membership analysis and representative patterns

Social demographic variables	Sample mean (%)		Mean of cluster (%)											
			# 1	# 2	# 3	# 4	# 5	# 6	# 7	# 8	#9	#10	#11	#12
Gender	Female	0.53	0.53	0.53	0.44	0.59	0.56	0.52	0.47	0.54	0.59	0.51	0.53	0.60
Age	Young adults (ages 15-35 years)	0.10	0.12	0.05	0.10	0.10	0.11	0.10	0.05	0.09	0.07	0.15	0.09	0.60
	Middle-aged adults (ages 36-55 years)	0.49	0.67	0.29	0.66	0.38	0.32	0.71	0.72	0.29	0.32	0.66	0.70	0.31
	Older adults (aged older than 55 years)	0.41	0.20	0.66	0.24	0.53	0.57	0.19	0.23	0.63	0.61	0.19	0.22	0.08
Education	Diploma or university certificate	0.67	0.76	0.58	0.76	0.62	0.66	0.85	0.53	0.57	0.65	0.64	0.80	0.38
Occupation	Regular shift	0.53	0.73	0.22	0.93	0.26	0.24	0.87	0.93	0.19	0.24	0.43	0.89	0.13
	Irregular schedule	0.10	0.22	0.10	0.03	0.10	0.11	0.09	0.07	0.07	0.07	0.47	0.08	0.02
	Student	0.03	0.01	0.00	0.00	0.04	0.01	0.01	0.00	0.03	0.02	0.04	0.01	0.67
	Retired	0.23	0.02	0.52	0.02	0.39	0.41	0.01	0.00	0.53	0.41	0.00	0.00	0.08
	Work at home	0.15	0.23	0.10	0.13	0.15	0.16	0.30	0.06	0.11	0.09	0.09	0.26	0.02
Flexible schedule	Have no flexibility in a work schedule	0.50	0.55	0.48	0.54	0.46	0.43	0.44	0.63	0.48	0.51	0.75	0.40	0.43
Job number	Have more than one job	0.07	0.09	0.04	0.04	0.16	0.08	0.05	0.07	0.11	0.05	0.08	0.08	0.00
Income	Low-income (<= $ 40,000)	0.39	0.28	0.44	0.22	0.48	0.49	0.32	0.29	0.53	0.48	0.47	0.26	0.78
	Middle-income ($ 40,000 - $ 100,000)	0.53	0.60	0.46	0.68	0.45	0.45	0.55	0.64	0.42	0.46	0.49	0.59	0.19
	High-income (> $ 100,000)	0.09	0.12	0.10	0.10	0.07	0.07	0.13	0.08	0.05	0.06	0.04	0.15	0.03
Total cluster membership			137	225	401	238	419	171	229	247	262	53	348	48
Percentage in total (number of person-days)			4.9	8.1	14	8.5	15	6.1	8.2	8.8	9.4	1.9	12.	1.7
Home chores (%)			25.0	34.7	27.8	41.3	43.1	33.7	30.1	40.5	40.1	35.0	29.4	35.5
Home leisure (%)			12.0	20.2	16.5	14.4	19.1	16.2	17.7	19.9	19.1	13.6	15.3	13.8
Night sleep (%)			62.8	45.0	55.6	44.1	37.7	50.0	52.1	39.5	40.7	51.3	55.2	50.6
Total in-home (%)			*100*	*100*	*100*	*100*	*100*	*100*	*100*	*100*	*100*	*100*	*100*	*100*
Workplace (%)			**89.2**	4.76	**82.6**	6.41	8.16	**79.5**	**89.3**	10.5	13.1	**90.9**	**86.5**	5.77
Shopping & services (%)			1.61	27.2	3.16	15.3	30.1	5.56	2.41	**32.2**	**38.2**	2.90	2.98	3.38
School/college (%)			0.00	1.00	0.14	1.20	0.42	0.40	0.29	3.16	1.28	0.66	0.15	**69.2**
Organizational/hobbies (%)			2.34	**29.5**	3.24	**31.2**	19.9	5.11	1.76	21.1	21.9	1.57	2.62	11.1
Entertainment (%)			4.67	17.4	6.92	**33.7**	10.3	6.48	3.73	8.75	11.6	1.93	4.30	4.53
Sports (%)			2.12	19.9	3.93	12.0	**30.9**	2.93	2.50	24.2	13.7	2.04	3.40	5.89
Total out-of-home (%)			*100*	*100*	*100*	*100*	*100*	*100*	*100*	*100*	*100*	*100*	*100*	*100*

Cluster #1: extended work-day workers, comprised a group of workers who engaged in work activity for a longer duration, starting from 8:00 a.m. to around 8:00 p.m. A large portion of workers in this cluster were middle-aged females aged between 36 and 55 years old (67.0%), while 76.0% of them had education levels higher than high school. Furthermore, 73.0% of people in this cluster were full-time workers, and they commonly had middle income (60.0%). The major percentage of the workers in this cluster (55.0%) indicated they had no flexibility in their work schedule.

Cluster #2: non-worker, midday activities, consisted of a group of people who participated in organizational/hobbies or entertainment activities mostly in the midday, starting from 10:00 a.m. to around 5:00 p.m. A large proportion of people in this cluster were female (53.0%) and also aged older than 55 years (66.0%). The majority of people in this cluster were educated and belonged to the middle or low income level. A minor proportion of the people in this cluster had work at home (10.0%) while 52.0% of them indicated that they had some flexibility in a work schedule.

Cluster #3: 8-4 workers, was a group of workers who engaged in work activity in a consistent manner, starting from 8:00 a.m. to around 4:00 p.m. The major proportion of workers in this cluster consisted of middle-aged males with education level higher than high school. A major proportion of workers in this cluster were full time workers (93.0%), and they typically had middle income level. Additionally, the workers in this cluster engaged in entertainment activities typically in the evening, starting around 6:00 p.m. for a duration around two hours.

Cluster #4: non-worker, evening activity, involved a group of people who participated in organizational/hobbies or entertainment activities mostly in the evening, starting from 6:00 p.m. to around 10:00 p.m. Similar to cluster 2, most people in this cluster were older females with education level higher than high school. A large proportion of people in this cluster (48.0%) belonged to the low income partition. Furthermore, a minor group of them had work at home (15.0%), while 54.0% of them indicated that they had some flexibility in a work schedule.

Cluster #5: stay-at-homes, comprised a group of people who mostly spent their time at home. The greater number of people in this cluster belonged to the low-income partition. Similar to clusters 2, 4, 8 and 9, a large proportion of people in this cluster consisted of old-aged females. Furthermore, a minor proportion of people in this cluster (4.64%) went out of home in the day for recreational activities. Compared to other activities, sports and shopping activities after in-home activity was most typical. In addition, cluster 5 had the largest membership (15.08%) in comparison with other identified clusters in this study.

Cluster #6: shorter work-day workers, involved a group of workers having work duration typically less than 5 hours in a day, and who finished their work in the early afternoon before 2:00 p.m. A large proportion of workers in this cluster were middle-aged females between 36 and 55 years old (71.0%). Furthermore, a large proportion of workers in this cluster (85.0%) had education level higher than high school. In total, 56.0% of them indicated that they had some flexibility in their work schedule. The workers in this cluster participated in more recreation activities than those in other identified worker clusters.

Cluster #7: 7-3 workers, comprised a group of workers who started work in the early morning around 7:00 a.m. and finished their work in the early afternoon around 3:00 p.m. A large proportion of workers in this cluster were middle-aged males between 36 and 55 years old (47.0%), and the majority had middle-income (64.0%). Furthermore, the majority of workers in this cluster were full time workers (93.0%), while 63.0% of them indicated that they had no flexibility in a work schedule. A minor portion of workers in this cluster (7.0%) have more than one job. It is interesting to note that workers in this cluster typically start and end at the workplace ahead of peak traffic periods both in the morning and afternoon.

Cluster #8: non-worker, morning shopping, involved a group of people who did shopping activities mostly in the morning, starting from 9:30 a.m. to around 12:00 p.m. Similar to clusters 2 and 4, a major proportion of people in this cluster consisted of females and also were aged older than 55 years. Moreover, similar to cluster 4, a large proportion of people in this cluster were educated and belonged to the low income level (53.0%). Furthermore, a minor group of them had work at home (11.0%). In total, 52.0% of them specified that they had some flexibility in a work schedule.

Cluster #9: non-worker, afternoon shopping, consisted of people who did shopping activities mostly in the afternoon. Consistent with other non-worker clusters, a large proportion of people in this cluster were female (59.0%) and also aged older than 55 years. Furthermore, a large proportion of people in this cluster (48.0%) belonged to the low income partition, with education level higher than high school. Compared to other identified non-worker clusters, only a small portion of them had work at home (9.0%), while 51.0% of them indicated that they had no flexibility in a work schedule

Cluster #10: evening workers, was a group of workers who mostly started to work in the evening around 4:00 p.m. and finished their work around midnight. In contrast to cluster 7, a large proportion of workers in this cluster were females (51.0%). Similar to cluster 3, 6 and 7, a large proportion of people in this cluster were middle-aged with education level higher than high school. Moreover, 47.0% of people in this cluster had an irregular working schedule, and they typically had middle or low income level.

Cluster #11: 9-5 workers, involved a group of workers who unlike workers in cluster 7, typically tend to travel to and from work during peak traffic periods both in the morning and afternoon. The majority of workers in this cluster consisted of middle-aged females between 36 and 55 years old (53.0%) with middle-income (59.0%) level. Similar to cluster 6, a large proportion of workers in this cluster had education level higher than high school. Interestingly, around 60.0% of them indicated that they had some flexibility in a work schedule.

Cluster #12: students, involved a group of students who engaged in school activity in a consistent manner. A large proportion of students in this cluster were young adults aged between 15 and 35 years old (60.0%). The majority of students in this cluster (78.0%) belonged to the low income partition. Furthermore, students typically engaged in recreation activities after school time, starting from 4:00 p.m. to around 11:30 p.m. We performed an examination of start time and duration probability distributions of different activities. The purpose of this analysis is to explore the cluster aspects from a temporal point of view. The probability distributions of start time and duration for work activity are shown in Figure 4.5 and Figure 4.6, respectively. It should be noted that we do not show the details of probability distributions of all activities for the sake of brevity.

Figure 4.5 Probability distribution of being at the workplace activity in clusters

Figure 4.6 Probability distribution of workplace activity duration in clusters

Results of the cluster distribution analysis reveal that start time and duration distributions of work activity in most of the clusters are clearly dissimilar. Also, these results reveal that the number of clusters significantly influences cluster features.

We employed the Kolmogorov-Smirnov (KS) test on the activity start time distributions between pairs of clusters, to test for significant differences. Result of KS tests for twelve different clusters and all activity categories are shown in Table 4.3. The KS test is built as a statistical hypothesis test. The null hypothesis (H_0) is that the two samples were drawn from the same population. Values of 1 in Table 4.3 indicate rejection of H_0 at the $p = 0.05$ level. As can be seen, in most of the tests the null hypothesis is rejected and start time distributions may be regarded as significantly different between the two clusters.

Figure 4.7 depicts a set of representative activity patterns that correspond to the centroids of each cluster. It should be noted that members within specific clusters can have activities that due to their lower share or short duration compared to other activities are absent from the representative patterns. Our results show that, by using the FCM clustering algorithm, each activity type is embodied in at least one of the representative patterns, which make it comparable with other clustering algorithm such as k-means (Jiang, Ferreira and Gonzalez 2012; Allahviranloo, Regue and Recker 2016).

Table 4.3 Kolmogorov-Smirnov test on activity start time distribution

Workplace

Cluster number	1	2	3	4	5	6	7	8	9	10	11	12
1		0	1	1	1	1	0	0	1	1	1	1
2			0	1	1	0	1	0	1	1	1	1
3				1	1	1	1	1	1	1	1	1
4					1	0	1	0	1	1	0	1
5						0	0	1	1	1	0	1
6							1	1	1	1	1	1
7								1	1	1	1	1
8									1	1	1	1
9										1	1	1
10											1	1
11												0
12												

Organizational/hobbies

Cluster number	1	2	3	4	5	6	7	8	9	10	11	12
1		1	1	1	1	1	1	1	1	1	1	1
2			1	1	1	1	1	0	1	1	1	1
3				0	0	0	0	1	1	0	1	1
4					1	0	1	0	1	1	0	1
5						1	1	0	1	1	1	1
6							0	1	0	1	1	1
7								1	1	1	1	1
8									1	1	1	1
9										1	1	1
10											1	1
11												0
12												

Shopping & services

Cluster number	1	2	3	4	5	6	7	8	9	10	11	12
1		1	1	1	1	1	1	0	1	1	1	1
2			1	1	1	1	1	1	1	1	1	1
3				0	0	0	1	1	1	1	0	1
4					1	1	1	1	1	1	1	1
5						1	1	1	1	1	1	1
6							1	1	1	1	1	1
7								0	0	0	1	1
8									1	1	1	1
9										1	1	1
10											1	1
11												0
12												

Entertainment

Cluster number	1	2	3	4	5	6	7	8	9	10	11	12
1		0	0	1	0	0	0	0	1	1	0	0
2			1	1	1	1	1	1	1	1	1	1
3				1	0	1	1	1	1	1	0	0
4					0	0	1	1	0	1	0	0
5						1	1	1	1	1	1	0
6							0	0	0	1	0	0
7								1	1	1	1	1
8									1	1	1	1
9										1	1	1
10											0	0
11												0
12												

School/college

Cluster number	1	2	3	4	5	6	7	8	9	10	11	12
1		1	1	1	1	1	1	1	1	1	1	1
2			1	1	1	1	0	1	1	0	1	0
3				0	0	0	1	1	0	0	0	0
4					1	1	1	1	1	1	1	0
5						1	1	0	0	0	1	0
6							0	0	0	0	0	0
7								1	1	1	1	0
8									1	1	1	1
9										0	1	0
10											1	0
11												0
12												

Sports

Cluster number	1	2	3	4	5	6	7	8	9	10	11	12
1		1	1	1	1	1	1	1	1	1	1	1
2			1	1	1	1	1	0	1	1	1	1
3				1	0	1	1	1	0	1	0	0
4					0	0	1	0	0	1	0	1
5						1	1	0	1	1	1	1
6							1	1	1	1	1	1
7								0	0	1	1	1
8									1	1	1	1
9										1	1	1
10											1	1
11												0
12												

**Values of 1 indicate significant differences at the p=0.05 significance level*

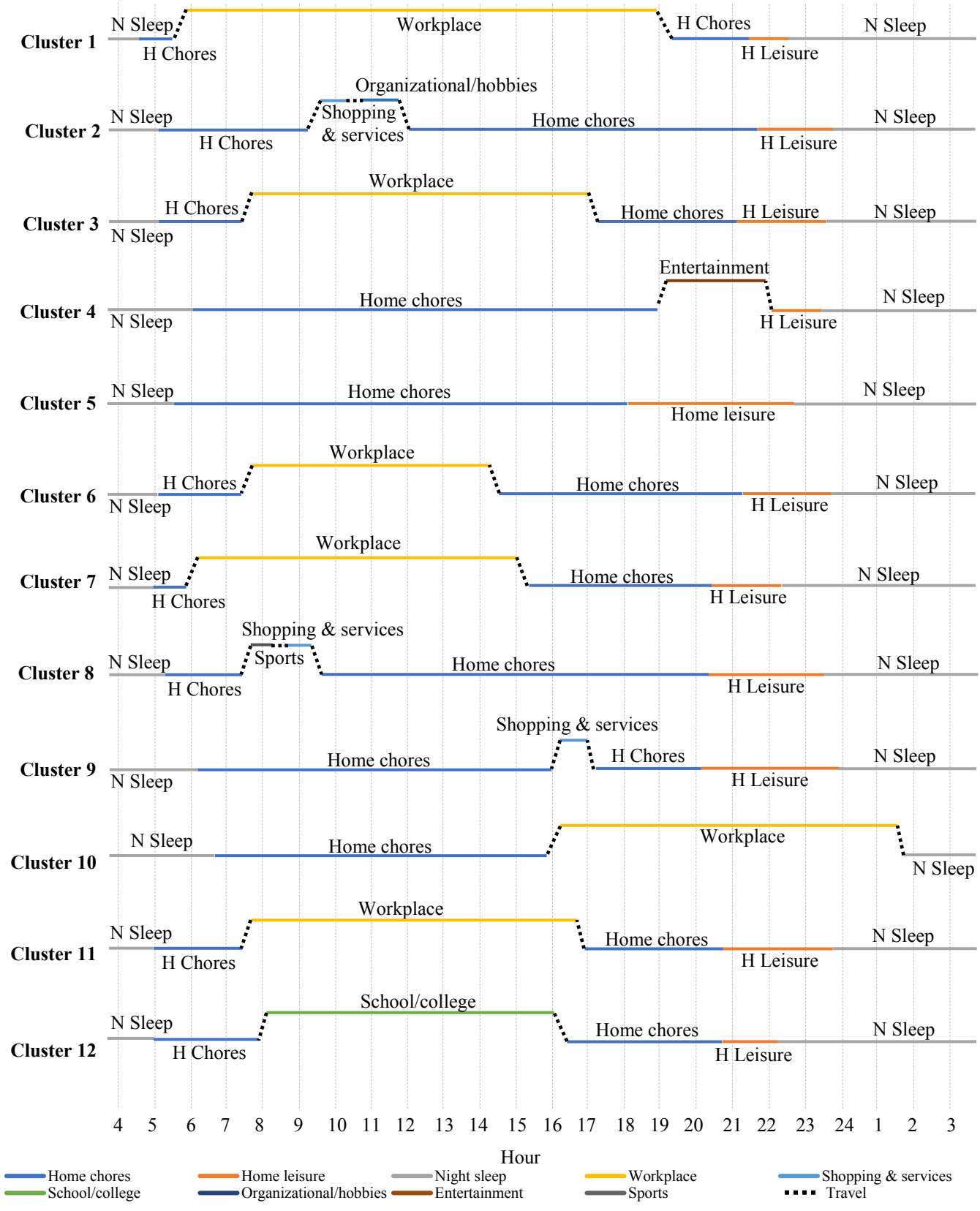


Figure 4.7 Twelve identified representative activity patterns

Figure 4.8 depicts the inter-dependencies among the attributes in each twelve identified clusters. Circles in Figure 4.8 designate leaves and triangles designates branches. The CART algorithm utilized the Gini index for leaf splitting and it fitted a tree with 16 leaves and 15 branches in total. Individuals are classified in the first root of the fitted tree based on their attendance at school or not. From branch 3, branches grow based on non-worker or works in-home versus out-of-home worker. Individuals are then classified as members of clusters 2, 5, 6, 8 or 9 based on income, age, and education level. Note that these clusters have a major non-mandatory activity (i.e., entertainment, sport, shopping) in their daily activities. In contrast, clusters 1, 3, 7, 10 and 11 have a main work activity in the pattern, and are based on income, age, education, and gender criteria.

The CART algorithm found specific clusters for particular leaf nodes based on the high probability that an individual belongs to it. However, it should be noted that, in each particular leaf node, there is a probability that an individual might belong to any of the other clusters. For instance, cluster 4 does not appear in any of the leaf nodes, as can be seen in Figure 4.8. Accordingly, we calculated the probability distributions of each cluster at each leaf node, and these are shown in Table 4.4. This allows us to infer cluster membership of individuals based on random number generation and cumulative probability functions. This CART classifier feature is important for use in future forecasting phases in activity-based travel demand modeling.

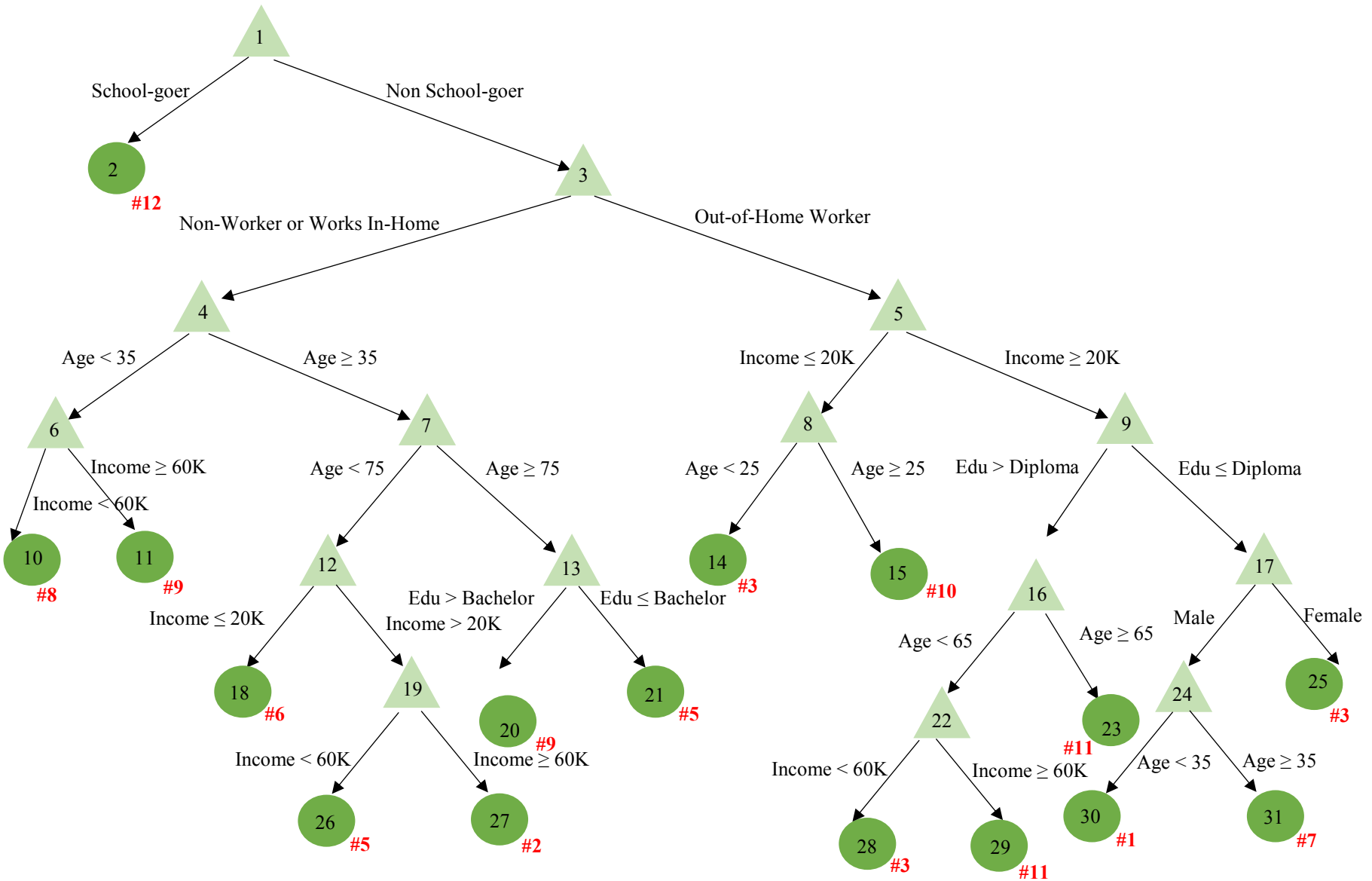


Figure 4.8 Decision tree results: Exploring attribute interdependencies in members of cluster

Table 4.4 Probability of different clusters in the decision tree

#	Leaf 2 – Cluster 12			Leaf 10 – Cluster 8			Leaf 11 – Cluster 9			Leaf 14 – Cluster 3			Leaf 15 – Cluster 10			Leaf 18 – Cluster 6		
	%	PDF	CDF	%	PDF	CDF	%	PDF	CDF	%	PDF	CDF	%	PDF	CDF	%	PDF	CDF
#1	0.08	0.04	0.04	0.04	0.04	0.04	0.03	0.02	0.02	0.14	0.11	0.11	0.04	0.04	0.04	0.00	0.00	0.00
#2	0.00	0.00	0.04	0.00	0.00	0.04	0.02	0.01	0.03	0.00	0.00	0.11	0.07	0.07	0.11	0.00	0.00	0.00
#3	0.00	0.00	0.04	0.00	0.00	0.04	0.00	0.00	0.04	0.32	**0.25**	0.36	0.03	0.03	0.14	0.15	0.15	0.15
#4	0.25	0.13	0.17	0.12	0.13	0.17	0.09	0.07	0.10	0.00	0.00	0.36	0.14	0.14	0.28	0.00	0.00	0.15
#5	0.25	0.13	0.30	0.12	0.13	0.29	0.00	0.00	0.10	0.14	0.11	0.47	0.04	0.04	0.32	0.00	0.00	0.15
#6	0.00	0.00	0.30	0.00	0.00	0.29	0.02	0.02	0.12	0.00	0.00	0.47	0.07	0.07	0.39	0.50	**0.50**	0.65
#7	0.00	0.00	0.30	0.17	0.17	0.46	0.01	0.01	0.13	0.00	0.00	0.47	0.03	0.03	0.42	0.00	0.00	0.65
#8	0.25	0.13	0.43	0.46	**0.46**	0.92	0.00	0.00	0.13	0.00	0.00	0.47	0.13	0.13	0.55	0.00	0.00	0.65
#9	0.00	0.00	0.43	0.00	0.00	0.92	0.73	**0.52**	0.65	0.00	0.00	0.47	0.16	0.16	0.71	0.00	0.00	0.65
#10	0.17	0.09	0.52	0.08	0.08	1.01	0.00	0.00	0.65	0.27	0.21	0.69	0.24	**0.24**	0.95	0.00	0.00	0.65
#11	0.00	0.00	0.52	0.00	0.00	1.01	0.09	0.07	0.72	0.14	0.11	0.80	0.05	0.05	1.00	0.35	0.35	1.00
#12	0.92	**0.48**	1.00	0.00	0.00	1.01	0.39	0.28	1.00	0.25	0.20	1.00	0.00	0.00	1.00	0.00	0.00	1.00

#	Leaf 20 – Cluster 9			Leaf 21 – Cluster 5			Leaf 23 – Cluster 11			Leaf 25 – Cluster 3			Leaf 26 – Cluster 5			Leaf 27 – Cluster 2		
	%	PDF	CDF	%	PDF	CDF	%	PDF	CDF	%	PDF	CDF	%	PDF	CDF	%	PDF	CDF
#1	0.00	0.00	0.00	0.00	0.00	0.00	0.04	0.04	0.04	0.08	0.08	0.08	0.01	0.01	0.01	0.00	0.00	0.00
#2	0.10	0.10	0.10	0.16	0.16	0.16	0.13	0.13	0.17	0.07	0.07	0.15	0.18	0.18	0.18	0.26	**0.26**	0.26
#3	0.00	0.00	0.10	0.00	0.00	0.16	0.04	0.04	0.22	0.19	**0.19**	0.34	0.02	0.02	0.20	0.00	0.00	0.26
#4	0.10	0.10	0.20	0.09	0.09	0.24	0.10	0.10	0.32	0.06	0.06	0.40	0.16	0.16	0.36	0.17	0.17	0.43
#5	0.30	0.30	0.50	0.37	**0.37**	0.61	0.13	0.13	0.45	0.13	0.13	0.52	0.26	**0.26**	0.61	0.19	0.19	0.61
#6	0.00	0.00	0.50	0.00	0.00	0.61	0.04	0.04	0.49	0.06	0.06	0.59	0.00	0.00	0.62	0.01	0.01	0.63
#7	0.00	0.00	0.50	0.00	0.00	0.61	0.04	0.04	0.53	0.12	0.12	0.71	0.00	0.00	0.62	0.00	0.00	0.63
#8	0.10	0.10	0.60	0.26	0.26	0.87	0.11	0.11	0.64	0.04	0.04	0.75	0.20	0.20	0.81	0.21	0.21	0.84
#9	0.40	**0.40**	1.00	0.13	0.13	1.00	0.09	0.09	0.73	0.07	0.07	0.82	0.18	0.18	0.99	0.13	0.13	0.97
#10	0.00	0.00	1.00	0.00	0.00	1.00	0.00	0.00	0.73	0.02	0.02	0.84	0.00	0.00	0.99	0.00	0.00	0.97
#11	0.00	0.00	1.00	0.00	0.00	1.00	0.27	**0.27**	1.00	0.15	0.15	0.99	0.00	0.00	1.00	0.01	0.01	0.99
#12	0.00	0.00	1.00	0.00	0.00	1.00	0.00	0.00	1.00	0.01	0.01	1.00	0.00	0.00	1.00	0.01	0.01	1.00

#	Leaf 28 – Cluster 3			Leaf 29 – Cluster 11			Leaf 30 – Cluster 1			Leaf 31 – Cluster 7		
	%	PDF	CDF	%	PDF	CDF	%	PDF	CDF	%	PDF	CDF
#1	0.08	0.08	0.08	0.10	0.10	0.10	0.31	**0.31**	0.31	0.05	0.05	0.05
#2	0.07	0.06	0.15	0.03	0.03	0.13	0.00	0.00	0.31	0.05	0.05	0.10
#3	0.25	**0.25**	0.40	0.16	0.16	0.29	0.25	0.25	0.56	0.15	0.15	0.25
#4	0.05	0.05	0.44	0.07	0.07	0.37	0.00	0.00	0.56	0.06	0.06	0.31
#5	0.06	0.06	0.50	0.06	0.06	0.43	0.08	0.08	0.63	0.08	0.08	0.39
#6	0.12	0.12	0.62	0.14	0.14	0.57	0.15	0.15	0.79	0.06	0.06	0.45
#7	0.04	0.04	0.67	0.09	0.09	0.66	0.15	0.15	0.94	0.31	**0.31**	0.76
#8	0.00	0.00	0.67	0.02	0.02	0.68	0.00	0.00	0.94	0.05	0.05	0.81
#9	0.06	0.06	0.73	0.06	0.06	0.74	0.05	0.05	0.99	0.05	0.05	0.86
#10	0.01	0.01	0.75	0.02	0.02	0.76	0.01	0.01	1.00	0.04	0.04	0.90
#11	0.25	0.25	0.99	0.24	**0.24**	1.00	0.00	0.00	1.00	0.10	0.10	1.00
#12	0.01	0.01	1.00	0.00	0.00	1.00	0.00	0.00	1.00	0.00	0.00	1.00

#: represents cluster number; %: represents share of each cluster in the identified cluster (leaf node in the decision tree); PDF: Probability distribution function; and, CDF: Cumulative distribution function

As an example, assume a non-worker or works in-home individual in the test set with the following socio-demographic characteristics: 32 years old and average income of 65K. According to Figure 4.8, the individual is allocated to leaf 11 in the decision tree. Subsequently, with respect to the probability distribution calculated in Table 4.4, the individual has a 52% likelihood to belong to cluster 9. However, as mentioned before, there is a 48% chance that the individual might belong to the other clusters. Suppose that the random generated number is 0.8. According to Table 4.4 and the cumulative probability functions, this value falls within the range of [0.72, 1.00] and indicates that the individual will be assigned to cluster 11.

4.6 Conclusions

Due to lack of full data on all individuals of the population, transport modelers are not able to predict or model the travel behavior of all individuals in the territory. Consequently, the best policy is to predict or model travel behavior for a representative set of model individuals, who represent homogeneous cohorts. Accordingly, aggregation is both inescapable and essential in travel demand modeling. The significant original contribution of this study is to develop a new inclusive pattern recognition modeling framework that leverages activity data to derive clusters of homogeneous daily activity patterns for use in activity-based travel demand modeling. We modeled the 2-day in-home and out-of-home time-use activity patterns of individuals drawn from the large Halifax STAR travel diary survey, the world's largest deployment of global positioning system (GPS) technology for a household activity survey. Each individual's daily pattern of activity was segmented into 288 three dimensional five-minute intervals, and information on activities, socio-demographics of the individual, and spatial information related to activities were coded to

the three interval dimensions. The clustering algorithm rapidly converged and resulted in twelve clusters of person-days, each with homogeneous activity patterns.

The proposed pattern recognition modeling framework in this study comprises four modules. The first module, subtractive clustering algorithm, initialized the cluster centroids and total number of clusters. The initial number of clusters found in this phase was validated through utilizing Dunn's index. The subtractive clustering algorithm is an alternative method to deal with problems including high resolution. The algorithm considers data points as potential sources which resulted in finding cluster centroids with higher accuracy. In the next module, person-days in the dataset were bundled into dissimilar clusters based on comparable routine activity sequences using the novel and efficient Fuzzy C-Means (FCM) clustering algorithm. When compared to other potential clustering algorithms such as K-means, FCM yields better convergence of the local minima of the squared error principle. This is directly associated to the choice of cluster centroids and to cluster membership. In the 3rd module twelve representative activity patterns were recognized, corresponding to cluster centroids. The progressive alignment method yields more accurate results by improving SAM through iterative profile-alignment of tree portions to maximize sum of pairs score. Finally, in the last module inter-dependencies among the attributes of each identified cluster were investigated using the CART algorithm. Compared to other potential classifiers such as C4.5, CART classifier improves decision tree performance by adding additional cross-validation step. The heterogeneous diversity among clusters in terms of their distributions of activity type, start time, activity duration, and end time, were confirmed by use of the non-parametric KS test.

In this study, we demonstrated how cluster analysis can be used to detect differences in the socio-demographic characteristics of population groups with different daily activity patterns. Our results show that individuals belonging to the non-worker or student clusters have different income, age, gender, and education level. Individuals with a stay-at-home pattern seem to be identified primarily by age, gender, and income level, while workers are found to have statistically dissimilar education, income, and flexible schedules. Lastly, and not surprisingly, students are found to have remarkably dissimilar marital status, age, and education level.

Numerous detailed information on activities, such as start time, activity duration, activity type, location, and travel distance can be extracted from each identified cluster. Such precise information is crucial for the scheduling step of activity-based travel demand modeling. The proposed method enriches the traditional methods such as using socio-demographic variables for classifying the population, and provides clusters based on the powerful computerized pattern recognition technique. For instance, discovering activity patterns over longer time periods, such as weekly, monthly, and seasonally, can be accomplished in a short period of time using the proposed algorithm in this study. Another advantage of the developed new method is that unlike previous approaches, the algorithm has the ability to recognize people who typically tend to avoid travel in peak traffic periods. Our model particularly recognized cluster #7 and cluster #11 of workers who commonly travel to and from work before and after peak traffic periods both in morning and afternoon peaks, respectively.

Compared to previous studies that used complex methods to capture frequent and infrequent activities in a dataset (Liu et al. 2015), the proposed FCM clustering algorithm

in this study is more straightforward and easy to implement in practical activity-based travel demand models. Furthermore, the cluster memberships selection in the FCM is comparable to the k-means clustering algorithm proposed by Jiang et al. (2012) and Allahviranloo et al. (2016). In the FCM each data point has the likelihood of belonging to several clusters, and this results in producing more homogeneous activity patterns in each cluster. For instance, we identified two different workers in cluster#7 and cluster#11 that, regardless of their similarity in activity sequences, are distinguished by start time and end time at the workplace. The application of this study is not restricted only to the transportation area: the presented new modeling framework can be harmonized for any applications that contain a set of connected sequences, such as recognition of functionally significant regions, or day-to-day variations in transit ridership and station demand at the individual level.

To build on this study and further demonstrate the potential of our proposed method, we are proposing several avenues of research. Firstly, it is possible to explore seasonal activity patterns by taking advantage of the wealth of data in the large-scale Halifax household travel diary survey (STAR). Secondly, and in line with growing worldwide interest in employing GPS locational data, we aim to investigate additional linkages between the STAR GPS data and travel diary data, and incorporating them in the proposed modeling framework in this study. Thirdly, we intend to extend our work to study the interaction between individual and household activity patterns, using data from the STAR survey. The latest generation of activity-based travel demand models includes interactions between household members, as these have a significant impact on others' travel. In addition, a conceivable further step of this work is to establish a hybrid framework employing discrete

choice models in combination with the output from pattern recognition to recognize likelihoods of activity participation and predict activity patterns of individuals with greater accuracy.

In summary, the modeling framework presented in this study provides a straightforward and easy-to-implement tool for urban and transport modelers to understand time-use activity patterns for different kinds of individuals. The results of this study are expected to be implemented within the activity-based travel demand model, Scheduler for Activities, Locations, and Travel (SALT) for Halifax, Nova Scotia.

Chapter 5 Learning Daily Activity Sequences of Population Groups Using Random Forest Theory[3]

5.1 Introduction

In recent years the activity-based modeling approach has received much attention from transport modelers and policy makers. Complexities in individual travel behavior such as flexible working hours, self-employment, and online shopping have increased considerably due to rapid technological progress. Moreover, policies related to road congestion and air pollution are having greater impacts on travel behavior. Clearly, we need more complex and disaggregated models that address the above-mentioned changes. Integrity, interdependencies, higher temporal and spatial aggregation, and behavioral basis are the four main reasons argued by Rasouli and Timmermans (2014) to change from the traditional four stage travel demand models to activity-based travel demand models. Numerous modeling approaches have been developed for activity-based travel demand models. The constraints-based modeling approach was one of the earliest techniques developed, based on the theory that travel is a derived demand, originating from the need of an individual to engage in activities (Hagerstrand 1970). With the concept that individuals desire to maximize the utility of their activity schedule, the random utility theory was then used in the development of activity-based models (McFadden 1980). More recently, the computational process modeling approach was developed using the theory of context-dependent choice preferences and emphasizing their scheduling aspect (Arentze

[3] A version of this chapter has been published:
Hafezi, M. H., L. Liu., and H. Millward. (2018). "Learning daily activity sequences of population groups using random forest theory". Transportation Research Record: Journal of the Transportation Research Board.

and Timmermans 2000). PCATS (2000), ALBATROSS (2000) and ADAPTS (2009) are some examples of activity-based travel demand models developed through the above-mentioned approaches. The broad modules for most activity-based travel demand models are: activity generator and scheduler, activity engagement, tour and trip destination, and mode choice and network assignment, among others.

The activity engagement patterns component provides explicit details on activity type, and the frequency and sequence of engaged activities. Daily activity engagement patterns of individuals are crucial components in any activity-based travel demand model, as individual's travel demand originates from their need to engage in particular activities. Many empirical and theoretical approaches have been employed for the activity engagement module, such as multinomial logit model, probabilistic grammars, decision tree, etc. (Bowman and Ben-Akiva 2001; Arentze and Timmermans 2007; Auld and Mohammadian 2009; Li and Lee 2017; Daisy et al. 2018a).

Despite all of the progress made in activity engagement modules through rule-based or econometric techniques, there have been few attempts to utilize the capability of machine learning to derive daily activity engagement patterns. In this study, we developed a new model that is able to learn and replicate individual's daily activity engagement patterns. Inspired by Random Forest (RF) theory, decision trees were grown using both CART and Curvature Search techniques. The resulting models were trained from observed activity sequences. A previous study reveals that ensemble methods performed better when they were applied to clustered data (Allahviranloo and Recker 2013). Therefore, in this study, we applied the models to twelve unique population clusters derived using a novel pattern recognition model, and compared the results to each other. We used confusion matrix,

transition matrix, and sequential alignment methods to compare the estimation accuracy for replication of activity type, activity position, and activity sequences. The remainder of this study is organized as follows: first, we review relevant past research concerning daily activity engagement patterns of travelers. Secondly, we discuss the data used for the activity pattern recognition. The modeling approach and the planned layer settings are explained in the next section, followed by a discussion of model results. The study concludes with a summary of contributions and a brief discussion of future research directions.

5.2 Literature Review

Daily activity engagement patterns differ between individuals based on their socio-demographic characteristics and their health and/or mobility status. In a seminal early work on time geography, Hagerstrand (1970) proposed three types of constraints that shape individual activity sequences. The first type are capability constraints, which focus on biological needs and available resources that can require or limit an individual's participation in an activity (e.g., eating meals, drinking, and sleeping). The second type are coupling constraints, which refer to the spatial and temporal necessities for an individual who joins with other individuals to conduct a certain activity. The third type are authority constraints, which limit the individual's access to certain activity locations or times. Thus, a decision to participate in a particular activity is a combined result of several decisions and constraints (e.g. household interaction, choices on activity type, location, timing, duration, destination, etc.).

Following Hagerstrand's time-space constraints, a wide variety of modeling approaches has been employed to model various aspects of daily activity sequences, such as activity type, activity sequence, activity frequency, sequential activity location, duration, and transport mode for the next trip. A three-level structural model was proposed by Bowman and Ben-Akiva (2001) for modeling daily activity pattern. In the first phase, the daily activity sequences choice set is established. In the next two phases, tour and trip decisions with regard to the choice of daily activity pattern are modeled. Following this concept, Bhat et al. (2004) modeled daily activity sequences in a more disaggregate manner and considered separate models for non-workers and workers. In other studies, Bhat et al. (2013) and Leszczyc and Timmermans (2002) estimated activity durations using the multiple discrete continuous extreme value (MDCEMV) and hazard models, respectively. Daily activity sequences can be modeled within tour formulation. Initially, the primary activity and its associated tour type (i.e. activity destination, sequence, and number of stops) is modeled, and in the next phase the number and purpose of secondary tours are estimated (Ben-Akiva and Bowman 1998b; Ho and Mulley 2013).

In addition to inclusive developed econometric and rule-based models, in recent years there has been growing interest in applying machine learning practices to activity-based modeling (Allahviranloo and Recker 2013; Li and Lee 2017; Hafezi, Liu and Millward 2018b). A wide range of applications have been developed using machine learning techniques, mostly in the computer science and statistics fields, but to date there have been very limited applications of the technique in activity-based modeling, and especially in modeling of activity engagement patterns. Li and Lee 2017 modeled individual's activity sequences by employing probabilistic context-free grammars. They adopted language

concepts to the daily activity sequences and estimated activity sequences. In another recent study by Allahviranloo and Recker 2013, daily traveler activity engagement patterns were modeled using Support Vector Machines (SVM). They employed several machine learning techniques in their proposed model. Using conditional random fields, the dependencies between activity sequence, activity type, and socio-demographic data were identified, and then the sequential choice of activities was captured by employing the markov chain model.

In this study we propose an algorithm inspired by Random Forest theory to model the activity engagement patterns of individuals. The existing applications of Random Forest in transportation fields are limited to transport mode recognition and traffic incident detection (Shafique and Hato 2015; You, Wang and Guo 2017). The proposed conceptual model in this study is able to learn and replicate individual's daily activity engagement patterns with regard to heterogeneity characteristics. The application of this framework to activity-based modeling not only discloses the efficiency of machine learning to model daily traveler activity engagement patterns, but also contributes additional insights to the linkage between activity agenda formation and activity scheduling modules.

5.3 Data

The data used in this study are drawn from the Space-Time Activity Research (STAR) survey undertaken in Halifax, Canada. The STAR survey was a joint household activity survey and travel survey, and the world's first large-scale employment of global positioning system (GPS) technology for tracking and confirmation of out-of-home activities. A brief description follows, and full descriptions of the survey design and the

socio-demographic features of respondents can be found in (TURP 2008; Millward and Spinney 2011; Spinney and Millward 2011). The survey period was between April 2007 and May 2008, equally spaced through days of the week and months of the year. It yielded fully geo-referenced two-day (i.e. 48-h) time diary data from 1,971 randomly designated primary respondents, aged 15 years or older. Respondents carried a GPS data logger, maintained a daily ‘‘activity log’’ during that period, and completed a day-after computer-assisted telephone interview (CATI) time-diary survey. The respondents' descriptions of their out-of-home activities were prompted and confirmed by the GPS data.

In this study, we use data from twelve identified unique clusters of respondents as a result of applying a novel pattern recognition modeling framework to the STAR data. Interested readers are referred to previous research by the authors (Hafezi, Liu and Millward 2017b; Hafezi, Liu and Millward 2017c) for more details on clustering methods and execution of the pattern recognition framework. In general, the pattern recognition model identified six clusters for workers (cluster#1: extended worker, cluster#3: 8-4 worker, cluster#6: shorter worker, cluster#7: 7-3 worker, cluster#10: evening worker and cluster#11: 9-5 worker), one cluster for students (cluster#12: students), four clusters for non-worker and non-student (cluster#2: non-worker midday activities, cluster#4: non-worker evening activities, cluster#8: non-worker morning shopping and cluster#9: non-worker afternoon shopping), and one cluster for individuals who mostly spend their time at home (cluster#5: stay-at-home). Individuals in each cluster differ notably from individuals in other clusters in terms of their socio-demographic characteristics. Moreover, individuals within each cluster have homogenous activity sequences whereas there is a heterogeneous diversity

between clusters in terms of their distributions of start time, activity duration, activity type, and socio-demographic characteristics.

5.4 Methods

Random Forest (RF) theory is based on the use of many decision trees that have been grown using the bagging method (Breiman 2001; Suthaharan 2015). Each tree plays as a weak learner in the algorithm and the combination of these weak inputs builds up a robust ensemble learning model. Results obtained from the RF model are based on the majority votes in the ensemble models. One of the main challenges in most decision tree models is to select the best split predictor method (Breiman 2001; Biau and Scornet 2016). In this study, we introduced two new techniques: CART and Curvature Search, each with two-layer settings for split predictor selection. Although the RF method is used in other transportation fields such as transport mode recognition and traffic incident detection, to the best of our knowledge establishing RF models using the CART and Curvature Search have not been explored in activity engagement pattern modeling or travel behavior studies.

5.4.1 The Random Forest (RF) Model

The Random Forest (RF) structure for predicting a set of activity type in individual's agenda is shown in Figure 5.1. Practically, the RF model take the following steps. Initially, cross validation partition for train data (D_r) and test data (T_e) are performed. Prior to the growing of each tree, r_n observations are randomly drawn from the sampled data points $r_n \in \{1, \dots, n\}$. Next, at each node in the decision tree a number of predictor variables are randomly selected from all the available predictor variables (Q). The default number of selected predictor variables is the square of the total number of existing predictor variables

($\sqrt{Q}$). Best split at each node (B_{lit}) is accomplished using the split predictor selection search over b_{lit} directions. As the tree grows, this process is repeated and continues until the information is saturated. In RF model, trees are not required for cost complexity pruning. The new arrival data at the testing stage $t_n \in \{1, \ldots, e\}$ is propagated down to all of the trees grown (m trees) in the RF model, and a set of prediction results from all trees is obtained $\{\hat{P}_1, \hat{P}_2, \hat{P}_3, \cdots, \hat{P}_M\}$. The final prediction result gained is based on majority votes $\hat{P}$.

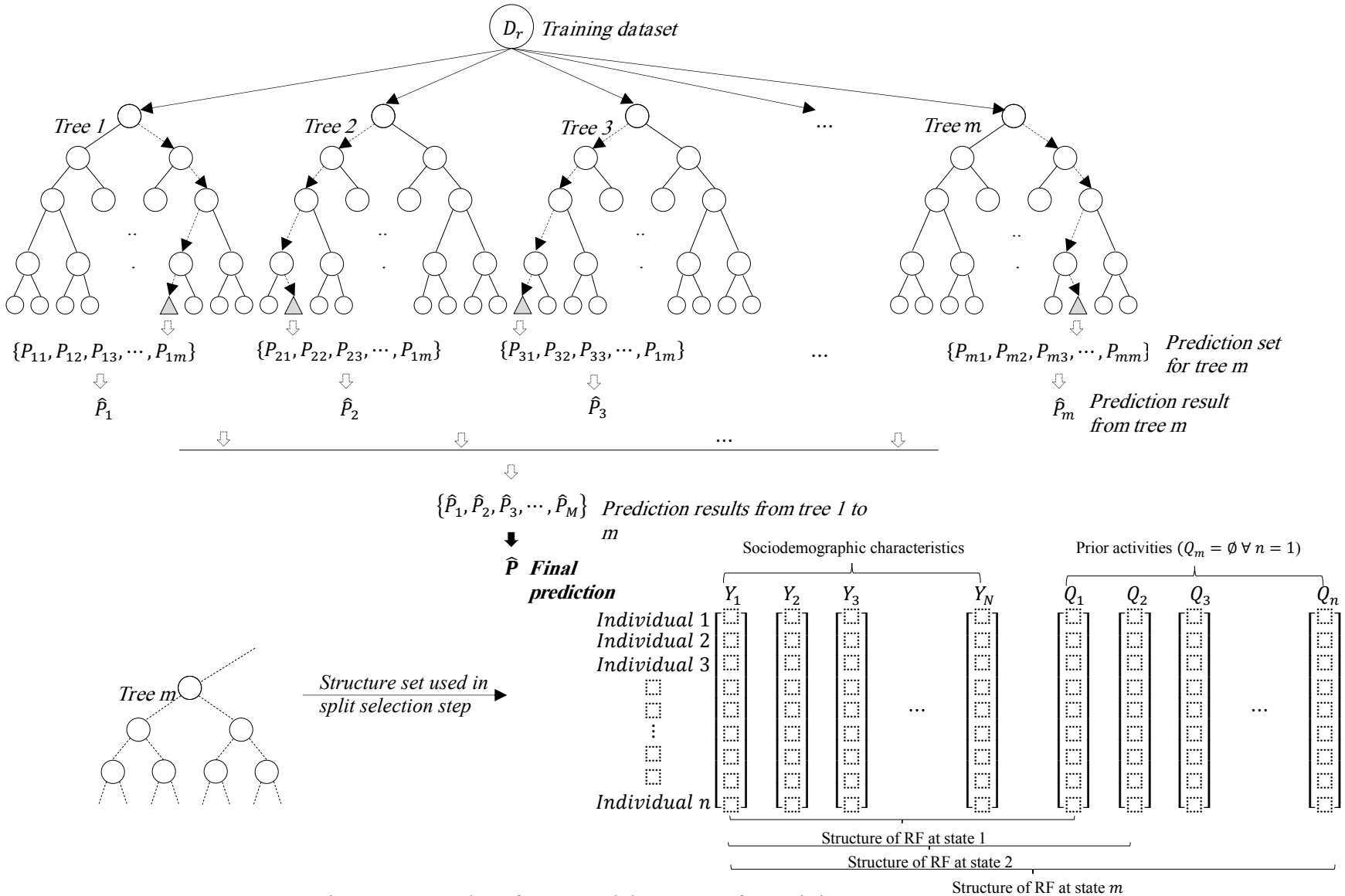


Figure 5.1 Random forest model structure for activity pattern sequence

5.4.2 Decision Splits: CART and Curvature Search

In this section, we describe how the CART and Curvature Search methods are operated in the RF models. We provide here only a brief overview of the algorithm, and interested readers are referred to (Breiman 2001; Biau and Scornet 2016) for more details. We define a training sample $D_r = ((X_i, Y_1), \dots, (X_i, Y_n))$ where X_i represents the response variable X = [home chores, home leisure, night sleep, workplace, shopping & services, school/college, organizational/hobbies, entertainment, sports] and Y_n represents the set of predictor variables (selected from Table 5.1). In this respect, we assume $E_n(C)$ as a subset of training data D_r where C is a pair of (u, w). u is the randomly selected predictor variable, and w is the place of the split along the u-th correspondent, within the limits of C. We postulate the set of all such feasible cuts in C as T_C. The split criterion $S_{clas,n}(u, w)$ is calculated as follows:

$$S_{clas,n}(u,w) = \frac{1}{E_n(C)}\sum_{i=1}^{n}(Y_i - \bar{Y}_C)^2 \exists_{X_i \in C} - \frac{1}{E_n(C)}\sum_{i=1}^{n}(Y_i - \bar{Y}_{C_L} \exists_{X_i^{(u)} < w} - \bar{Y}_{C_R} \exists_{X_i^{(u)} \geq w})^2 \exists_{X_i \in C} \quad (1)$$

$$X_i = \left(X_i^{(1)}, \dots, X_i^{(p)}\right) \forall (u,w) \in T_C \quad (2)$$

$$C_L = \{x \in C : x^{(u)} < w\} \quad (3)$$

$$C_R = \{x \in C : x^{(u)} \geq w\} \quad (4)$$

$\bar{Y}_C$ is the average of the Y_i such that X_i belongs to C.

In the CART, the best cut (u_n^*, w_n^*) is computed as follows:

Table 5.1 Proposed predictor variables for learning daily activity engagement patterns

Predictor	Subcategories
Gender	Male, female
Age	15 to 19, 20 to 24, 25 to 29, 30 to 34, 35 to 39, 40 to 44, 45 to 49, 50 to 54, 55 to 59, 60 to 64, 65 to 69, 70 to 74, 75 to 79, 80 to 84, 85+
Marital status	Married, living common-law, widowed, separated, divorced, single-never married,
Household size	1,2,3,4,5,6
Highest education level	Masters or earned doctorate, bachelor or undergraduate degree, diploma or certificate, some university, some community college, trade, technical or business college, high school/secondary, other
Full/part-time student	Full-time student, part-time student
Paid/self employed	A paid worker, self-employed, an unpaid family worker
Flexible work schedule	No, yes
Work at home	No, yes
Total personal income	Under $20,000, $20,000–$39,999, $40,000–$59,999, $60,000–$79,999, $80,000–$99,999, $100,000 or more
Total household income	Under $20,000, $20,000 - $39,999, $40,000 - $59,999, $60,000 - $79,999, $80,000 - $99,999, $100,000 or more
Dwelling type	Single unit residential, duplex or semi-detached, townhouse, multi-unit residential-less than 6 stories, multi-unit residential-6 or more stories, mobile dwelling, other-specify
Dwelling owned/rented	Owned, rented
Valid driver's license	No, yes
Buss pass	No, yes
Number HH vehicles	0,1,2,3,4,5,6
Number HH motorcycles	0,1,2,3,4,5,6,7,8,9,10
Number HH bicycles	0,1,2,3,4,5,6,7,8,9,10
Usually mode to work	Car, truck or van - as driver, car, truck or van - as passenger, public transit, walk to work, bicycle, motorcycle, taxicab, other method
Usually mode to school	Car, truck or van - as driver, car, truck or van - as passenger, public transit, walk to work, bicycle, motorcycle, taxicab, other method
State of health	Excellent, very good, good, fair, poor
Prior activities	Home chores *(working at home, eating/meal preparation, indoor or outdoor cleaning, interior or exterior home maintenance, child care, other in home activities)*, Home leisure *(watching tv/listening to radio, reading books/newspapers, etc.)*, Night sleep, Workplace *(work/job, all other activities at work, work related, etc.)*, Shopping & services *(shopping for goods and services, routine shopping)*, School/college *(class participation, all other activities at school)*, Organizational/hobbies *(organizational, voluntary, religious activities. hobbies done mainly for pleasure, cards, board games, all other hobbies activities)*, Entertainment *(eat meal outside of home, all other entertainment activities)*, Sports *(walking, jogging, bicycling, all sports related activities)*.

$$(u_n^*, w_n^*) \in \underset{\substack{u \in B_{lit} \\ (u,w) \in T_C}}{argmax} S_{clas,n}(u, w) \quad (5)$$

$argmax$ represents the maximization $S_{clas,n}(u, w)$ over B_{lit} and T_C. In the Curvature Search, the ideal split is computed where the density curves intersect, extracted from the roots (R_1,R_2).

5.4.3 Variable Importance Measures: Mean Decrease Accuracy (MDA)

RF models have the capability to measure the importance of variables and rank them in order to use in the algorithm. Mean Decrease Accuracy (MDA) and Mean Decrease Impurity (MDI) are the two well known measures of variables significance (Breiman 2001; Biau and Scornet 2016). In this study, we used the MDA technique that relies on the out-of-bag (OOB) error estimate. The MDA of the variable $X^{(u)}$ is calculated by balancing the difference in OOB error (O_n) prior and subsequent to the permutation over all trees. The prior OOB error of each tree is computed by testing the RF model using OOB data. The later OOB error of each tree is calculated by adding noise to the sample data of the feature randomly and retesting the OOB error. The MDA for randomly selected variable $X^{(u)}$ is calculated as follows:

$$\text{MDA}(X^{(u)}) = \frac{1}{m}\sum_{l=1}^{m}\left[O_n\left[d_n(.;\Theta_m), I_{m,n}^u\right] - O_n\left[d_n(.;\Theta_m), I_{m,n}\right]\right] \quad (4)$$

$$O_n[m_n(.;\Theta_l), I] = \frac{1}{|I|}\sum_{i:(X_i,Y_i)\in I}(Y_i - m_n(X_i;\Theta_l))^2 \quad (5)$$

Where $I_{m,n}$ is the out-of-bag data set of the m-tree, $I = I_{m,n}^u$ is the permuted data for variable u and $O_n(.;\Theta_m)$ is the estimation for the m-th tree.

5.4.4 Model Calibration and Validation

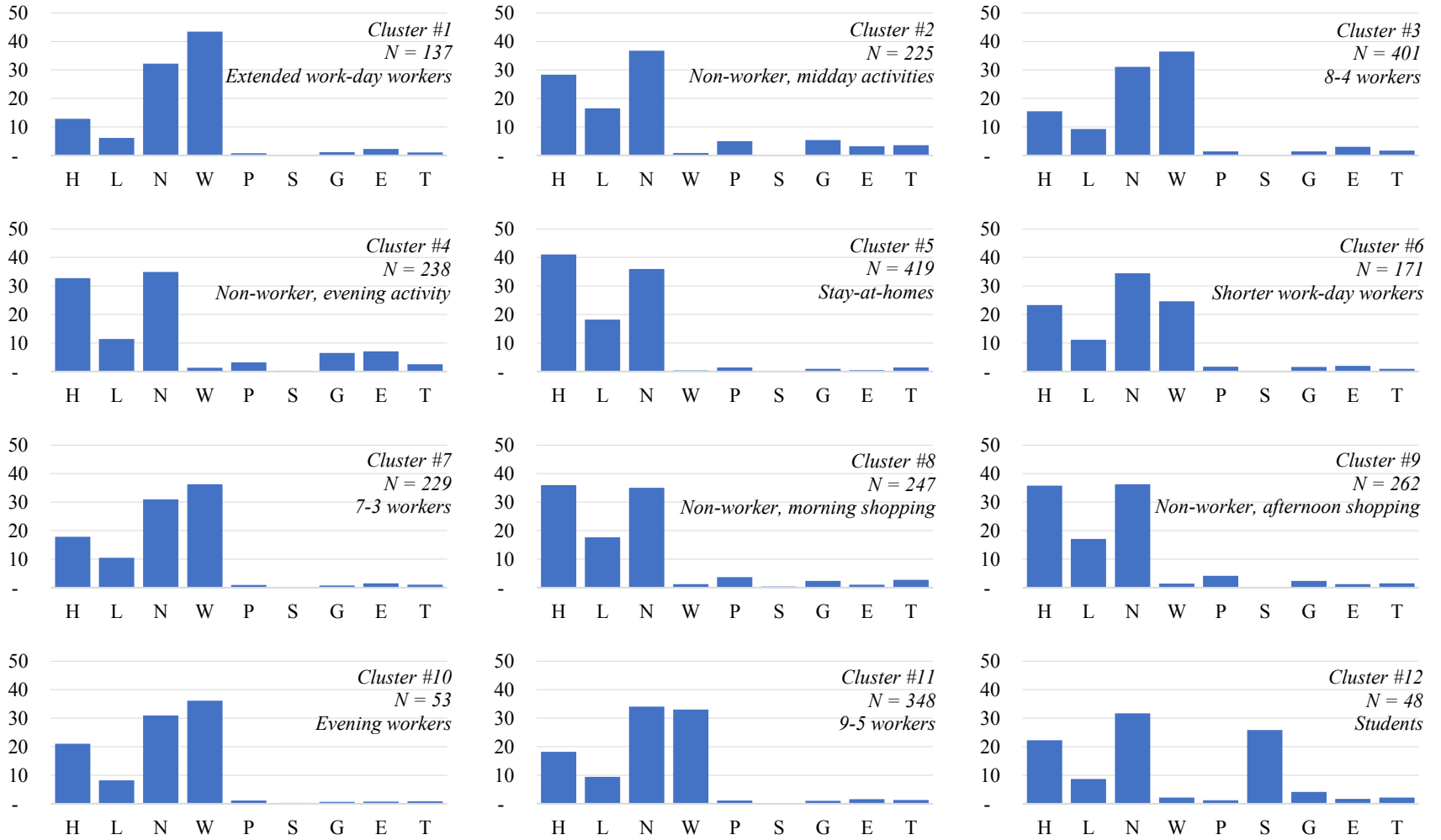
Compared to other machine learning algorithms such as support vector machine (SVM) and back propagation neural network (BPNN), in the RF algorithm few parameters need to be initialized and calibrated. The initial value of m (number of trees) is set to 1000. However, it needs to be validated and adjusted if the model cannot be converged within this value. B_{lit} (best split at each node) needs to be calibrated. In this study we used CART and Curvature Search techniques to determine the best split at each node. T_C (cutoff) is a vector of length equal to the number of classes. T_C comprises three sub-parameters ($c1$, $c2$, $c3$) that are initially each randomly selected in the range between [0,1] with the condition of total sum equal to 1. For all above parameters, we used the OOB error rate to cross-validate and obtain the best parameter values for the RF models.

5.5 Discussion of Results

In order to analyze the impacts of decision splits on the RF model, we trained RF with two different layer settings. In the first layer, we grew trees where predictor variables were randomly selected from all nominated variables. For simplicity's sake in this study, we refer to this setting's layer as RF_CART_I and RF_CURV_I for CART and Curvature Search techniques, respectively. In the second layer, we estimated the variable importance and included only those variables in the RF model that had more than 70% importance compared to others. For simplicity's sake in this study, we refer to this setting's layer as RF_CART_II and RF_CURV_II for CART and Curvature Search techniques, respectively. The response variables are defined as nine main activities, as described in Table 5.1. Previous studies reveal that choice of activities in the schedule not only depends

on an individual's socio-demographic characteristics but also is impacted by prior activities (Kitamura, Chen and Pendyala 1997). Recent empirical work by Allahviranloo and Recker (2013) also shows that SVM as a robust unsupervised machine learning technique has better performance when response variables are associated to both socio-demographic characteristics and previous activities. In this respect, we set up predictor variables in the RF models to choose from a set of socio-demographic variables and prior activities conducted by individuals in a given day. Hence, depending on an activity's position in the agenda, the number of predictor variables varies. Furthermore, additional predictor variables such as flexibility level of work schedule and health state of individuals enter into the analysis. The list of nominated response variables is shown in Table 5.1.

The OOB rate error cross-validation indicates that after $m > 850$ (m is the number of trees), the OBB error rate tends to be stable. Therefore, it is realistic to accept m as 1000 at first. On that basis, we obtain the best optimal parameters for RF models as follows: $c1$ = 0.25, $c2$ = 0.35 and $c3$ = 0.40. Figure 5.2 illustrates the distribution of daily activity engagement for all twelve clusters. As can be seen, most individuals in the worker clusters (#1, #3, #6, #7, #10, #11), in addition to in-home activities, only participate in the out-of-home work activity. A similar result is achieved for the student cluster (#12), where most students only participate in out-of-home school activity in addition to in-home activities. Individuals in other clusters (non-worker and non-school goer) were found to have different out-of-home activity participation patterns (mostly engaging in entertainment or/and recreation activities).



Vertical axis: Probability in empirical data; Horizontal axis: Activity type (H = home chores, L = home leisure, N = night sleep, W = workplace, P = shopping & services, S = school/college, G = organizational/hobbies, E = entertainment, T = sports).

N = cluster size

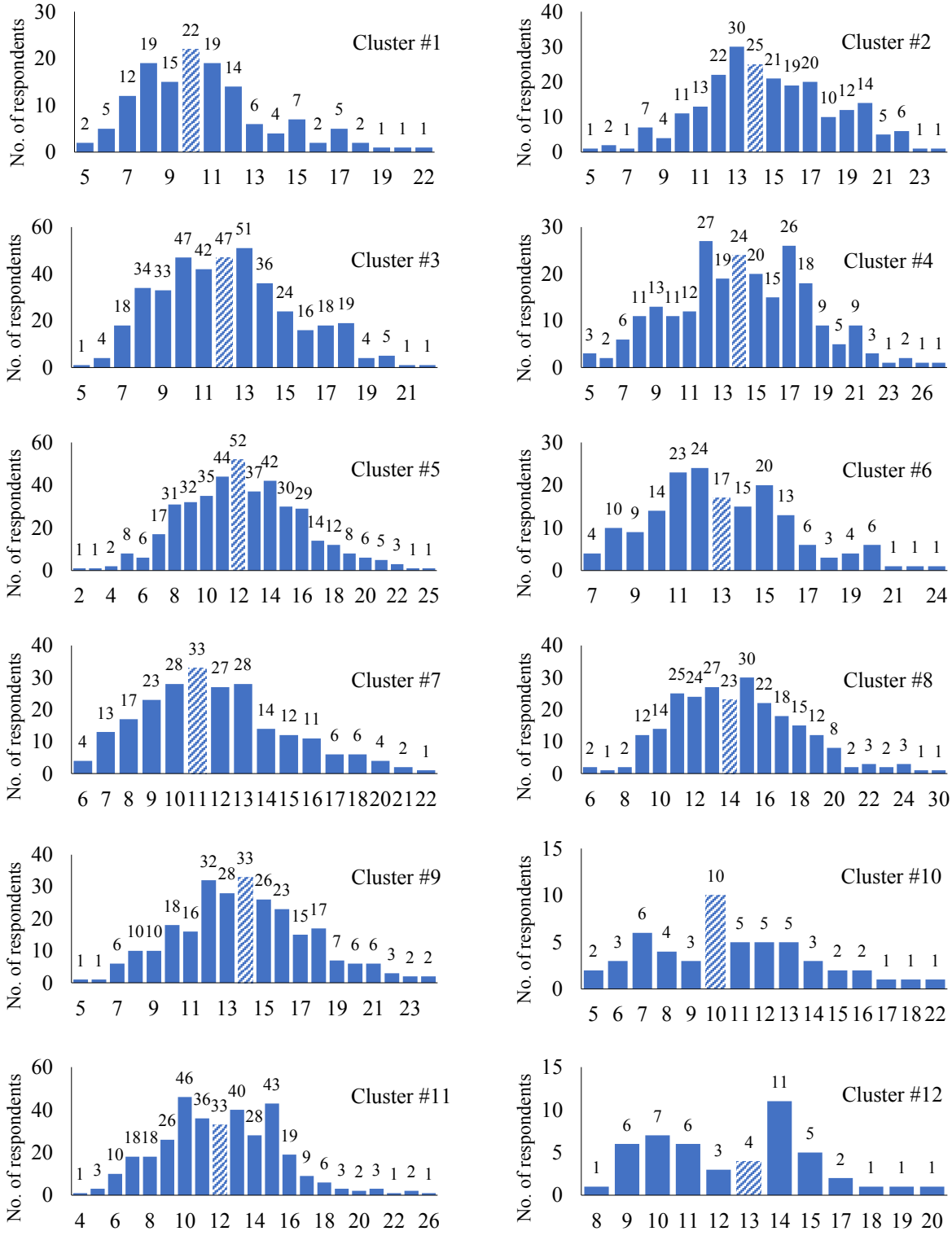
Figure 5.2 Distribution of daily activity engagement for all twelve clusters

Distribution patterns of number of daily activity episodes for individuals in all twelve clusters are shown in Figure 5.3. Our intent is to relate modeling to the most typical member in each cluster. In this respect, we assume the median number of episodes in each cluster as the alternative state number for daily activity participation in the modeling. Moreover, we assume the ultimate activity state for the day as the night sleep. Among twelve clusters, cluster #2 (non-worker midday activity) has the most complex activity pattern and cluster #12 (students) has the least complex activity pattern.

We performed a cross validation partition to our dataset which resulted in 70% for training and 30% for testing model performance. Comparison of the estimation accuracy between four proposed RF models for test dataset and training dataset is shown in Figure 5.4. Based on the estimated accuracy, we can conclude that RF_CART_I provides the best results for most states, followed by RF_CURV_I.

Explicitly, estimation accuracy of RF_CART_I for the test and training sets exceeds 69% for all clusters, and is particularly high for clusters 5 (stay-at-homes) and 12 (students) (74% each). The lowest accuracy is for cluster #2 (non-worker midday activities, at 69%). The average for reproducing activity types using the RF_CART_I model is 71.4% for the entire population. Allahviranloo and Recker (2013) modeled daily activity engagement using the SVM technique with 84% accuracy in the replicating agenda. In their study, five activity types for modeling nine states were used, and the modeling period was between 5:00 a.m. and 23:00 p.m. Given that we are modeling so many more activity types (three in-home and six out-of-home activities) for a period of 24 hours, we can conclude that our model is capable of replicating agenda within a reasonable accuracy range.

A comparison of estimation accuracy between the twelve clusters in Figure 5.4 demonstrates that prediction percentages decline from first state to successive states for all clusters. The decline is most noticeable in more complex activity sequences, suggesting that for these patterns the algorithm needs more sophisticated structure. Further work will include improving the estimation accuracy of the RF models by adding an optimization tool such as Bayesian optimization to tune the hyperparameters in the split selection phase.



Horizontal axis: No. of separate activity episodes (of t minutes or more); Dashed column represents cluster median

Figure 5.3 Number of respondents against number of activity episodes, for all twelve clusters

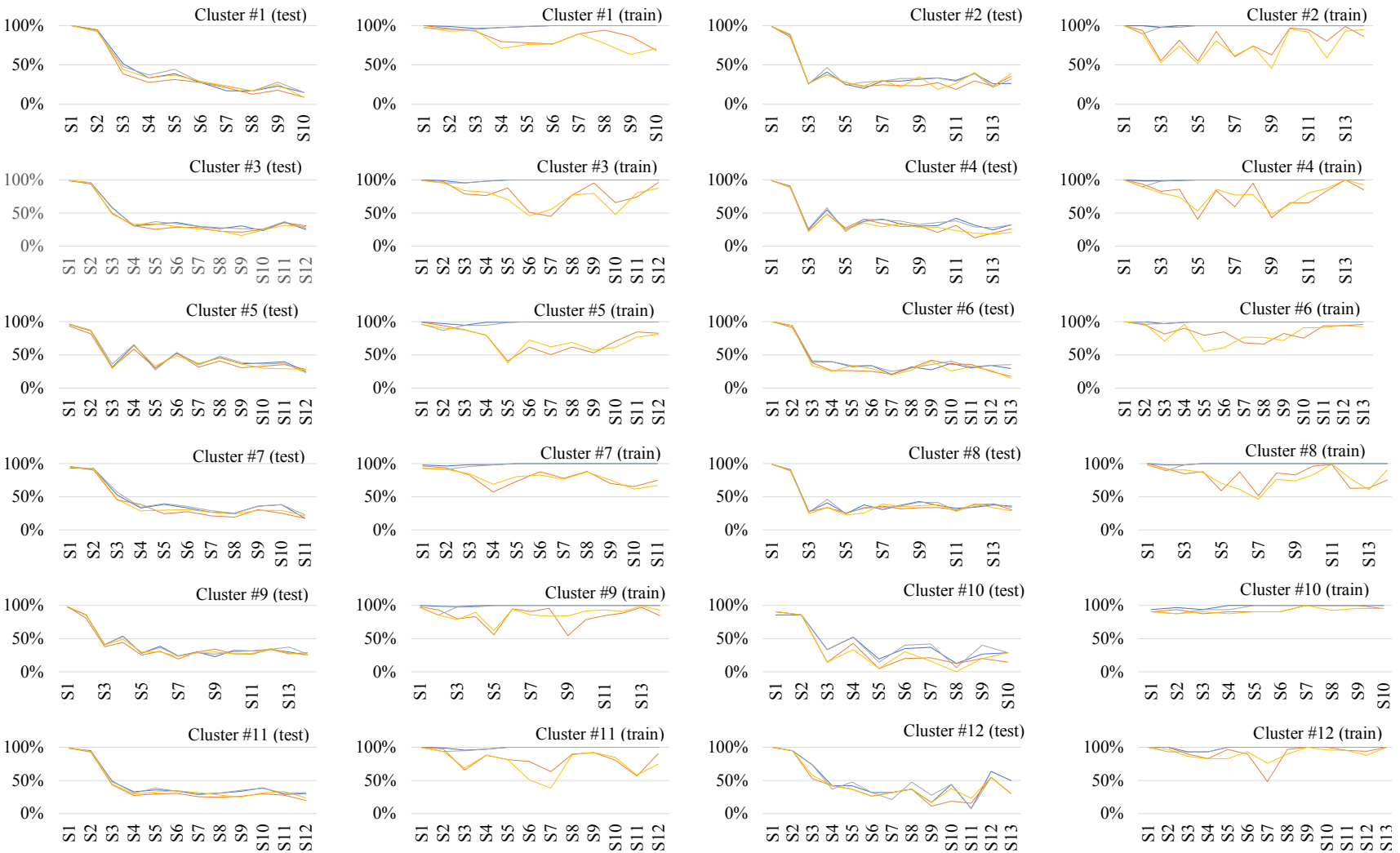


Figure 5.4 Comparing the estimation precision between four RF models for test and training dataset

The confusion matrix for the best-performing RF model, RF_CART_I, is shown in Table 5.2. The confusion matrix determines the ability of the RF model to replicate the existence of each activity type in the dataset. We computed the confusion matrix for all twelve clusters. At each state, the observed activity of the individual is compared with the individual's activity in the estimation, regardless of whether the individual's prior or following activity was replicated properly. The estimation accuracy in each cluster is calculated by the sum of crosswise cells of the confusion matrixes. According to the results presented in Table 5.2, the highest estimation accuracy is obtained for clusters 12 and 5 (80%) and the lowest estimation accuracy obtained is for cluster 2 (76%).Overall, the RF_CART_I model estimation accuracy is 77.7% for the entire population.

For each cluster, the agenda episodes and frequency of the observed data and the results of the best-performing RF models (RF_CART_I), including both in-home and out-of-home activity episodes, was examined by constructing a transition matrix. A transition matrix between successive activity episodes illustrates the likelihood that a consecutive episode of a certain class will happen, given an episode of a current class (Lockwood, Srinivasan and Bhat 2005). Furthermore, it presents detailed information on trip chaining patterns of different market sections in a compact way. Table 5.3 provides the comparison results between observed and replicated patterns through activity episode transitions matrixes for all twelve clusters. The columns in Table 5.3 signify the class of the subsequent activity episode (difference between observed and replicated patterns), while the rows in transition matrix signify the class of the current activity episode (difference between observed and simulated patterns). According to the results presented in Table 5.3,

the mean absolute error (MAE) for all twelve clusters is 7.26%, signifying that RF models could successfully replicate episodes and position in agenda.

In order to test for similarity of activity sequence between observed and replicated patterns, the order of activities was compared using sequential alignment methods (Needleman and Wunsch 1970). In this comparison, the activities are assumed to be independent of start times or activity duration. The distance between activity sequences is computed as the number of phases needed to align two orders of activities. The smaller distance between strings designates higher similarity. Figure 5.5 illustrates the distribution patterns of edit distance between observed and replicated activity sequences (outcomes of the RF_CART_I) for all twelve clusters, where the dashed line indicates the mean distance. In general, RF_CART_I successfully replicated activity sequences of more than 70% of the population in each cluster, with the mean distance equal to 0.47. In Table 5.2 and Table 5.3: H = Home chores, L = Home leisure, N = Night sleep, W = Workplace, P = Shopping & services, S = School/college, G = Organizational/hobbies, E = Entertainment, T = Sports.

Table 5.2 Confusion matrixes for random forest model (RF_CART_I)

#1	H*	L	N	W	P	S	G	E	T
H	.267	.009	.008	.022	.002	.000	.002	.002	.000
L	.016	.065	.002	.014	.000	.000	.002	.002	.000
N	.010	.002	.145	.006	.001	.000	.000	.000	.001
W	.025	.005	.001	.166	.003	.000	.007	.000	.000
P	.011	.001	.002	.007	.040	.000	.000	.002	.000
S	.000	.000	.000	.000	.000	.000	.000	.000	.000
G	.010	.000	.002	.012	.002	.000	.048	.000	.000
E	.012	.000	.000	.006	.002	.000	.002	.030	.000
T	.006	.001	.001	.006	.001	.000	.000	.001	.011

Accuracy #1 = 77.3

#2	H	L	N	W	P	S	G	E	T
H	.309	.018	.004	.000	.016	.000	.004	.000	.002
L	.037	.113	.001	.000	.007	.000	.003	.000	.001
N	.010	.004	.103	.000	.001	.000	.000	.000	.000
W	.003	.001	.000	.009	.001	.000	.000	.000	.000
P	.033	.011	.000	.000	.112	.000	.004	.002	.001
S	.001	.000	.000	.000	.001	.000	.000	.000	.000
G	.019	.007	.000	.000	.008	.000	.059	.000	.000
E	.014	.002	.000	.000	.005	.000	.001	.028	.000
T	.010	.004	.000	.000	.002	.000	.001	.000	.025

Accuracy #2 = 76.0

#3	H	L	N	W	P	S	G	E	T
H	.278	.014	.004	.018	.003	.000	.001	.001	.000
L	.028	.087	.001	.008	.001	.000	.001	.001	.001
N	.013	.004	.122	.001	.000	.000	.000	.000	.000
W	.022	.003	.000	.150	.002	.000	.002	.003	.001
P	.013	.004	.000	.009	.045	.000	.001	.000	.000
S	.000	.000	.000	.001	.000	.000	.000	.000	.000
G	.010	.003	.000	.008	.001	.000	.040	.001	.000
E	.012	.003	.001	.005	.001	.000	.000	.034	.001
T	.007	.003	.000	.004	.000	.000	.001	.000	.022

Accuracy #3 = 77.8

#4	H	L	N	W	P	S	G	E	T
H	.342	.011	.002	.001	.011	.000	.003	.002	.001
L	.032	.112	.001	.000	.006	.000	.003	.002	.001
N	.008	.002	.108	.000	.001	.000	.001	.002	.000
W	.005	.001	.000	.013	.001	.000	.000	.000	.000
P	.031	.007	.001	.000	.082	.000	.002	.001	.000
S	.002	.001	.000	.000	.000	.001	.000	.000	.000
G	.024	.006	.000	.000	.002	.000	.059	.000	.001
E	.021	.002	.000	.000	.004	.000	.000	.039	.000
T	.009	.004	.000	.000	.002	.000	.000	.001	.021

Accuracy #4 = 77.6

#5	H	L	N	W	P	S	G	E	T
H	.390	.027	.003	.000	.003	.000	.001	.000	.001
L	.059	.167	.002	.000	.005	.000	.001	.000	.001
N	.015	.006	.123	.000	.000	.000	.000	.000	.000
W	.002	.001	.001	.005	.000	.000	.000	.000	.000
P	.024	.008	.001	.000	.055	.000	.001	.000	.000
S	.001	.000	.000	.000	.000	.000	.000	.000	.000
G	.010	.005	.001	.000	.000	.000	.024	.000	.000
E	.004	.002	.000	.000	.000	.000	.000	.006	.000
T	.009	.007	.000	.000	.001	.000	.000	.000	.024

Accuracy #5 = 79.5

#6	H	L	N	W	P	S	G	E	T
H	.311	.019	.003	.013	.004	.000	.001	.000	.001
L	.035	.100	.001	.007	.001	.000	.001	.000	.000
N	.013	.003	.118	.000	.000	.000	.000	.000	.000
W	.021	.004	.000	.111	.001	.000	.002	.002	.000
P	.021	.003	.000	.007	.057	.000	.000	.000	.000
S	.000	.000	.000	.000	.000	.000	.000	.000	.000
G	.012	.001	.000	.010	.001	.000	.041	.000	.000
E	.011	.002	.000	.003	.001	.000	.001	.021	.000
T	.008	.001	.000	.000	.000	.000	.000	.000	.013

Accuracy #6 = 77.4

#7	H	L	N	W	P	S	G	E	T
H	.289	.023	.006	.017	.002	.000	.001	.000	.000
L	.032	.106	.002	.010	.002	.000	.000	.000	.001
N	.014	.003	.124	.001	.001	.000	.000	.000	.000
W	.023	.005	.001	.143	.001	.000	.003	.003	.000
P	.009	.006	.000	.003	.040	.000	.000	.000	.000
S	.000	.000	.000	.001	.000	.000	.000	.000	.000
G	.007	.002	.000	.009	.001	.000	.033	.000	.000
E	.009	.002	.000	.006	.000	.000	.000	.028	.000
T	.007	.001	.000	.003	.000	.000	.000	.000	.013

Accuracy #7 = 77.9

#8	H	L	N	W	P	S	G	E	T
H	.339	.027	.001	.001	.011	.000	.003	.000	.002
L	.049	.143	.002	.000	.007	.000	.001	.000	.001
N	.010	.004	.105	.000	.001	.000	.000	.000	.000
W	.005	.001	.000	.011	.001	.000	.000	.000	.000
P	.028	.010	.000	.000	.092	.000	.002	.001	.001
S	.001	.000	.000	.000	.000	.003	.000	.000	.000
G	.015	.007	.000	.000	.003	.000	.047	.000	.000
E	.007	.002	.000	.000	.001	.000	.000	.010	.000
T	.009	.005	.000	.000	.002	.000	.001	.000	.026

Accuracy #8 = 77.6

#9	H	L	N	W	P	S	G	E	T
H	.320	.024	.003	.001	.019	.000	.001	.000	.001
L	.044	.144	.001	.000	.010	.000	.001	.000	.000
N	.014	.005	.109	.000	.001	.000	.000	.000	.000
W	.003	.002	.000	.013	.002	.000	.000	.000	.000
P	.034	.012	.000	.000	.111	.000	.001	.000	.000
S	.001	.000	.000	.000	.000	.001	.000	.000	.000
G	.012	.007	.001	.000	.006	.000	.040	.000	.000
E	.005	.002	.000	.000	.004	.000	.000	.014	.000
T	.008	.003	.000	.000	.003	.000	.000	.000	.017

Accuracy #9 = 76.8

#10	H	L	N	W	P	S	G	E	T
H	.307	.017	.002	.021	.002	.000	.000	.000	.000
L	.023	.117	.002	.002	.006	.000	.000	.000	.000
N	.004	.010	.134	.006	.000	.000	.000	.000	.002
W	.025	.004	.006	.096	.004	.000	.000	.002	.000
P	.013	.006	.002	.006	.056	.000	.000	.002	.000
S	.000	.002	.000	.002	.000	.000	.000	.000	.000
G	.008	.006	.000	.010	.000	.000	.029	.000	.000
E	.006	.002	.000	.008	.002	.000	.000	.015	.000
T	.008	.002	.000	.004	.002	.000	.000	.000	.013

Accuracy #10 = 76.6

#11	H	L	N	W	P	S	G	E	T
H	.290	.012	.002	.019	.003	.000	.001	.000	.001
L	.029	.086	.000	.008	.002	.000	.001	.000	.001
N	.013	.004	.125	.002	.000	.000	.000	.000	.000
W	.023	.006	.001	.146	.002	.000	.002	.001	.000
P	.016	.003	.001	.006	.046	.000	.001	.001	.000
S	.001	.000	.000	.000	.000	.000	.000	.000	.000
G	.012	.002	.001	.008	.000	.000	.039	.000	.001
E	.009	.002	.000	.006	.001	.000	.001	.027	.000
T	.006	.002	.000	.004	.001	.000	.001	.000	.022

Accuracy #11 =78.1

#12	H	L	N	W	P	S	G	E	T
H	.318	.014	.002	.000	.004	.016	.002	.005	.004
L	.016	.095	.002	.000	.000	.004	.002	.002	.000
N	.009	.009	.113	.000	.000	.000	.000	.000	.000
W	.004	.002	.000	.011	.000	.004	.004	.000	.000
P	.011	.005	.000	.000	.034	.004	.002	.000	.000
S	.018	.000	.000	.000	.000	.122	.002	.000	.005
G	.016	.004	.000	.000	.000	.007	.050	.000	.004
E	.002	.002	.000	.000	.000	.005	.004	.023	.000
T	.005	.000	.000	.000	.000	.005	.002	.002	.029

Accuracy #12 = 79.4

Table 5.3 Activity episode transitions matrix: Comparison between observed and replicated patterns (in %)

#1	H*	L	N	W	P	S	G	E	T	
H	*0.0*	4.0	7.8	35.2	4.6	0.0	4.2	2.1	4.6	
L	26.6	*0.0*	7.8	11.7	9.0	0.0	7.0	5.3	1.3	
N	11.5	0.5	*0.0*	5.8	1.6	0.0	2.1	1.6	0.1	
W	22.1	4.6	13.7	*0.0*	2.7	0.0	7.2	23.2	7.4	
P	15.7	12.3	16.2	35.5	*0.0*	0.0	1.3	7.4	0.0	
S	0.0	0.0	0.0	0.0	0.0	*0.0*	0.0	0.0	0.0	
G	11.2	0.6	18.2	22.3	7.0	0.0	*0.0*	0.4	0.2	
E	24.4	11.1	17.4	52.7	1.4	0.0	1.0	*0.0*	2.6	
T	7.4	7.4	13.3	18.5	0.0	0.0	9.6	13.3	*0.0*	

MAE cluster 1 = 7.8

#2	H	L	N	W	P	S	G	E	T	
H	*0.0*	5.3	2.8	0.2	23.4	0.0	5.1	3.8	0.0	
L	25.3	*0.0*	4.1	1.0	16.0	0.3	8.4	3.6	0.1	
N	0.1	3.1	*0.0*	0.0	2.4	0.0	0.4	0.0	0.4	
W	33.3	20.8	20.8	*0.0*	16.7	0.0	13.9	11.1	0.0	
P	8.0	5.3	8.3	16.7	*0.0*	2.9	11.1	21.3	0.4	
S	0.0	0.0	0.0	0.0	0.0	*0.0*	0.0	0.0	0.0	
G	8.7	2.4	3.2	0.4	14.0	0.0	*0.0*	1.6	8.0	
E	30.1	11.3	3.8	4.9	9.5	0.0	0.2	*0.0*	30.8	
T	8.4	10.8	0.0	0.0	18.3	0.0	6.7	7.4	*0.0*	

MAE cluster 2 = 6.6

#3	H	L	N	W	P	S	G	E	T	
H	*0.0*	23.4	4.8	1.8	21.2	0.1	5.1	4.5	0.9	
L	18.4	*0.0*	9.8	1.0	10.6	0.0	9.9	5.2	1.4	
N	1.2	2.3	*0.0*	0.0	1.5	0.0	1.4	0.4	0.3	
W	8.1	7.6	17.5	*0.0*	1.6	0.0	14.6	10.0	2.1	
P	18.1	3.1	11.8	16.8	*0.0*	0.0	10.6	3.7	1.9	
S	5.0	0.0	0.0	0.0	0.0	*0.0*	0.0	5.0	0.0	
G	20.8	5.6	20.2	1.1	5.0	1.4	*0.0*	3.6	4.0	
E	20.9	12.4	17.7	7.0	4.7	0.0	8.9	*0.0*	30.3	
T	18.1	1.8	7.6	1.5	16.5	0.0	4.8	10.6	*0.0*	

MAE cluster 3 = 6.8

#4	H	L	N	W	P	S	G	E	T	
H	*0.0*	8.9	7.6	32.6	6.5	0.0	5.7	3.5	1.9	
L	25.4	*0.0*	1.1	8.9	7.6	0.0	3.5	4.5	0.2	
N	7.6	3.9	*0.0*	0.6	1.8	0.0	0.0	0.8	0.5	
W	29.3	10.9	5.4	*0.0*	2.6	0.0	7.8	25.1	10.0	
P	30.1	8.3	16.7	28.0	*0.0*	4.7	8.4	12.1	1.9	
S	0.0	0.0	0.0	0.0	0.0	*0.0*	0.0	0.0	0.0	
G	9.4	2.7	9.9	34.3	5.2	0.0	*0.0*	4.4	2.7	
E	29.9	14.7	6.1	56.1	5.3	0.6	4.6	*0.0*	5.2	
T	21.5	6.1	13.0	11.7	4.0	0.0	8.0	2.1	*0.0*	

MAE cluster 4 = 8.2

#5	H	L	N	W	P	S	G	E	T	
H	*0.0*	18.4	4.6	0.6	11.4	0.0	3.3	1.3	2.0	
L	14.9	*0.0*	4.2	0.3	7.9	0.0	1.7	0.6	0.2	
N	3.0	1.7	*0.0*	0.9	1.3	0.0	0.2	0.0	0.7	
W	17.6	7.3	19.2	*0.0*	2.0	0.0	5.4	2.7	3.8	
P	9.3	20.0	12.1	17.2	*0.0*	0.0	3.7	14.9	5.6	
S	0.0	0.0	0.0	0.0	0.0	*0.0*	0.0	0.0	0.0	
G	5.8	1.5	3.6	2.8	5.2	0.0	*0.0*	0.8	0.5	
E	25.9	20.0	10.0	15.4	31.8	0.0	6.7	*0.0*	15.4	
T	22.9	3.7	12.8	0.8	5.4	0.0	0.7	0.5	*0.0*	

MAE cluster 5 = 5.7

#6	H	L	N	W	P	S	G	E	T	
H	*0.0*	5.2	1.2	24.4	7.4	0.0	5.6	2.3	3.3	
L	21.9	*0.0*	1.0	6.4	6.3	0.0	5.8	3.4	1.0	
N	7.5	3.3	*0.0*	1.0	1.5	0.0	2.0	0.0	0.3	
W	11.3	7.1	2.4	*0.0*	1.3	0.0	2.3	18.4	3.4	
P	24.4	5.3	12.4	19.3	*0.0*	0.0	16.6	0.6	6.8	
S	0.0	0.0	0.0	0.0	0.0	*0.0*	0.0	0.0	0.0	
G	14.1	5.1	11.3	4.8	12.4	2.9	*0.0*	6.3	0.7	
E	26.2	12.6	0.0	53.9	5.1	0.0	8.0	*0.0*	2.0	
T	9.9	0.3	22.2	5.7	4.8	0.0	3.7	3.7	*0.0*	

MAE cluster 6 = 6.4

#7	H	L	N	W	P	S	G	E	T	
H	*0.0*	4.9	1.6	26.7	5.3	0.0	4.2	3.1	0.6	
L	21.8	*0.0*	0.2	10.0	5.4	0.0	4.0	2.7	0.0	
N	9.2	5.0	*0.0*	1.9	0.7	0.0	0.9	0.4	0.3	
W	21.0	7.0	6.6	*0.0*	2.8	0.7	2.0	26.0	8.7	
P	6.4	27.2	9.7	21.0	*0.0*	1.0	2.8	0.9	6.7	
S	0.0	0.0	0.0	100.0	0.0	*0.0*	0.0	0.0	0.0	
G	8.3	7.5	7.9	24.6	1.6	0.0	*0.0*	2.5	3.1	
E	26.4	22.0	12.5	9.9	6.9	0.0	0.9	*0.0*	1.2	
T	6.7	4.2	9.1	13.5	4.6	0.0	0.8	9.9	*0.0*	

MAE cluster 7 = 7.6

#8	H	L	N	W	P	S	G	E	T	
H	*0.0*	32.1	3.4	1.6	20.8	0.4	4.6	2.0	0.6	
L	21.3	*0.0*	2.7	0.1	12.4	0.2	5.6	1.8	1.3	
N	3.1	0.0	*0.0*	0.7	0.7	0.0	0.7	0.4	0.6	
W	8.5	0.9	2.8	*0.0*	14.8	0.0	7.5	5.3	2.8	
P	29.6	19.7	6.3	20.4	*0.0*	0.0	11.7	20.1	3.4	
S	20.8	12.5	0.0	0.0	41.7	*0.0*	0.0	33.3	0.0	
G	17.6	3.9	6.3	7.4	11.8	1.2	*0.0*	1.2	5.3	
E	23.9	7.6	0.0	18.5	10.9	0.0	5.5	*0.0*	18.5	
T	5.0	7.1	9.7	3.2	17.3	1.1	10.8	0.0	*0.0*	

MAE cluster 8 = 8.0

#9	H	L	N	W	P	S	G	E	T	
H	*0.0*	31.4	1.4	0.2	29.4	0.2	4.8	1.6	3.0	
L	20.3	*0.0*	0.8	1.0	15.1	0.0	3.2	1.3	0.4	
N	0.3	3.0	*0.0*	0.4	2.7	0.0	0.4	0.7	0.0	
W	9.7	4.2	4.4	*0.0*	6.5	0.0	3.8	0.8	0.8	
P	34.9	15.3	10.6	17.2	*0.0*	2.1	21.6	11.3	8.7	
S	0.0	0.0	0.0	0.0	0.0	*0.0*	0.0	0.0	0.0	
G	19.8	3.6	13.8	0.7	9.2	0.0	*0.0*	2.3	3.8	
E	34.4	10.0	13.3	3.3	3.3	0.0	28.9	*0.0*	22.2	
T	29.0	7.9	7.5	2.2	29.4	0.0	0.2	2.4	*0.0*	

MAE cluster 9 = 7.2

#10	H	L	N	W	P	S	G	E	T	
H	*0.0*	3.9	8.9	32.3	5.5	0.0	3.3	3.1	3.6	
L	22.3	*0.0*	5.3	19.7	8.1	0.0	4.3	1.3	3.1	
N	8.0	3.0	*0.0*	7.6	1.3	0.0	0.6	2.5	1.1	
W	31.3	2.3	7.2	*0.0*	5.5	0.0	5.0	19.7	0.5	
P	4.8	0.9	21.7	20.1	*0.0*	0.0	6.7	19.2	9.6	
S	0.0	0.0	0.0	0.0	0.0	*0.0*	0.0	0.0	0.0	
G	17.9	13.0	23.9	42.9	6.0	12.5	*0.0*	6.5	0.0	
E	19.1	0.2	19.4	67.3	22.6	0.0	6.5	*0.0*	0.0	
T	26.3	20.0	12.5	15.0	0.0	0.0	18.8	0.0	*0.0*	

MAE cluster 10 = 9.1

#11	H	L	N	W	P	S	G	E	T	
H	*0.0*	6.4	4.8	31.2	5.5	0.1	4.9	3.2	3.5	
L	20.1	*0.0*	1.3	11.4	5.9	0.0	2.2	2.5	0.5	
N	10.0	3.8	*0.0*	1.6	2.1	0.0	1.1	0.7	0.7	
W	24.8	7.1	7.1	*0.0*	0.2	0.0	3.1	26.9	9.1	
P	6.5	13.3	17.6	31.2	*0.0*	0.0	5.0	3.5	2.3	
S	0.0	0.0	0.0	0.0	0.0	*0.0*	0.0	0.0	0.0	
G	1.6	10.2	11.2	27.7	2.8	0.0	*0.0*	3.9	1.9	
E	24.6	14.2	5.6	57.7	3.5	0.0	8.2	*0.0*	1.5	
T	4.2	5.7	10.1	8.2	9.5	0.0	1.8	1.1	*0.0*	

MAE cluster 11 = 6.9

#12	H	L	N	W	P	S	G	E	T	
H	*0.0*	3.5	8.7	2.1	3.1	26.7	6.9	5.5	2.0	
L	18.6	*0.0*	5.7	1.9	3.8	0.8	11.3	1.9	4.6	
N	4.9	1.3	*0.0*	0.0	0.0	0.0	3.6	0.0	0.0	
W	16.7	0.0	0.0	*0.0*	0.0	16.7	0.0	0.0	0.0	
P	5.0	10.0	10.0	25.0	*0.0*	10.0	0.0	5.0	5.0	
S	1.5	6.6	1.6	4.3	9.7	*0.0*	1.8	25.1	5.0	
G	5.6	7.5	1.4	4.7	9.1	23.3	*0.0*	1.6	6.1	
E	3.3	11.1	11.1	5.6	5.6	2.2	20.0	*0.0*	14.4	
T	7.3	13.6	18.2	0.0	9.1	13.9	9.1	6.7	*0.0*	

MAE cluster 12 = 6.8

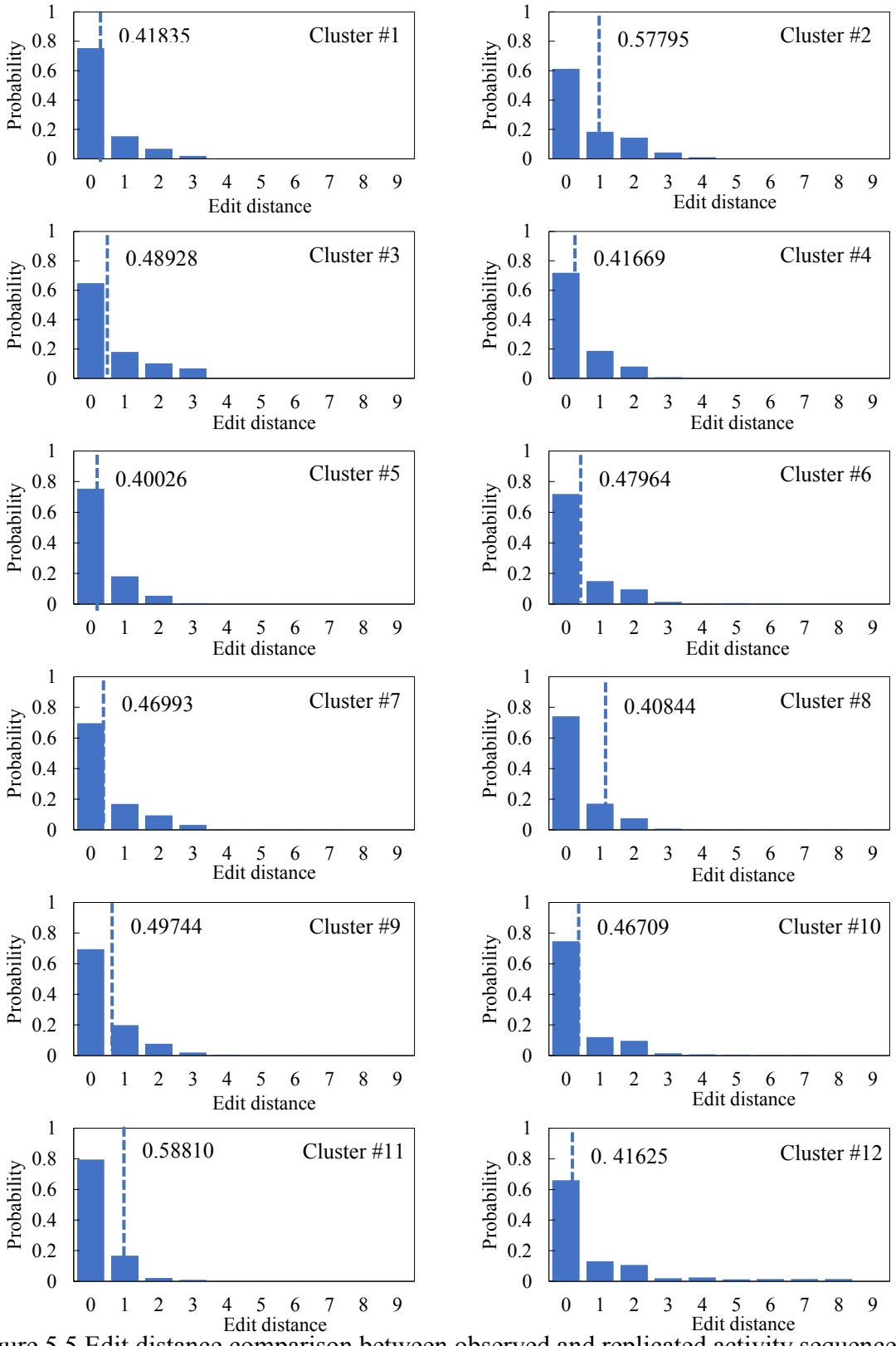


Figure 5.5 Edit distance comparison between observed and replicated activity sequences in 12 clusters

Our empirical evidence from modeling twelve unique clusters shows that performance of the RF models is meaningfully correlated with the sample size used in training model. This is most evident in cases comprising complex patterns. Further studies to overcome this limitation associated with the cluster data with small sample size are planned, in particular employing the population synthesis technique.

5.6 Conclusions

Daily activity pattern models play an essential role in many activity-based travel demand models. Individuals' travel demands originate from the need to participate in certain activities, and accordingly they plan their daily activity sequences based on these needs. The choice of daily activity sequences differs between population groups with varying socio-demographic characteristics. In this study, we developed a new model using Random Forest (RF) theory to learn and replicate activity engagement patterns of population groups. The RF models were trained with two different decision split techniques, CART and Curvature Search. Each of the proposed method was tested under two different layer settings. In the first setting, RF_CART_I and RF_CURV_I, the algorithm randomly selects predictor variables from all nominated variables and grows trees, whereas in the second setting, RF_CART_II and RF_CURV_II, the algorithm selects only predictor variables with high importance and grows trees. Various predictor variables, such as socio-demographic characteristics (age, gender, education level, income level, number of vehicles, etc.), health and/or mobility status of the individual, and preceding activities are employed in the training of RF models.

We applied the proposed RF models to twelve unique population clusters. Individuals in each cluster differ significantly from individuals in other clusters in terms of socio-demographic characteristics and daily activity sequences. Our results demonstrate that the RF model with the CART split selection method (RF_CART_I) can replicate activity sequences and agenda of the entire population with 70.0% and 77.7% accuracy, respectively. We considered nine activity categories (three in-home and six out-of-home) and modeled for a 24-hour period. Comparison of estimated activity sequences for twelve unique clusters demonstrated that RF model performance is associated with level of complexity patterns. In situations involving complex patterns, the algorithm requires more sophisticated structure in terms of predictor and split selections. Relative to other alternative machine learning algorithms such as SVM and BPNN, the RF model is less likely to overfit. This is due to the generalization error that converges to a particular amount when the forest size gradually increases. It can be used to assess the importance of attributes, which is suitable to analyze the features of the studied system. Furthermore, fewer parameters require calibration in the model, and the algorithm may be applied to datasets containing a large number of hidden attributes (Breiman 2001).

This study illustrates the utility of the RF machine learning technique to modeling of sequential activity selection, an application not previously employed in travel behavior analysis. The developed RF models can yield the type and frequency of activities in the schedule, and their sequential order, for use in activity-based travel demand modeling. Compared to multinomial logit regression modeling, which yields interpretable coefficients, the RF model is designed as a black box. It is trained to evaluate the labels of the data and its result is a set of support data-points and their respective weights. In

addition, compared to other artificial intelligence methods such as the support vector machine, the proposed method in this study is able to automatically handle missing values in the algorithm. Furthermore, variables do not need to be transformed, very few parameters need adjustment, and the algorithm doesn't overfit easily. The proposed RF model is also very efficient in computational time.

The empirical results from modeling twelve population clusters show that although RF_CART_I provides superior results than other proposed models, its performance is correlated with the sample size used in the training model, and also to heterogeneity and diversity among the predictor variables. Moreover, model with random selection of predictor variables performed better than including only high importance predictor variables. Further behavioral investigation can be made for interpretation of such a result. Further improvement involves adding an optimization tool to tune the hyperparameters in the split selection step of RF models. In this study we only addressed the activity type and activity sequencing. Various other aspects of daily activity sequences, such as sequential activity location, duration, and mode can be synthesized using the RF models. Moreover, incorporating a probabilistic model such as a hidden Markov model into the RF model structure can be explored in future extensions of this study. Several studies have illustrated the importance of intra-household interaction, in which decisions of activity participation by one household member are often constrained or correlated to the decisions of other household members. Extensions to our modeling procedure are required to account for such correlations. Finally, the results of this study are expected to be implemented within the activity-based travel demand model, Scheduler for Activities, Locations, and Travel (SALT) for Halifax, Nova Scotia.

Chapter 6 Modeling Activity Scheduling Behavior of Travelers for Activity-Based Travel Demand Models[4]

6.1 Introduction

In the late 1950s, investments in new road infrastructure increased considerably due to the substantial increase in vehicle ownership and usage for daily trips. To evaluate the short term and long term impacts of investments, transport planners developed the first generation of travel demand models known as four stage models (Goran 2001). The four stage models comprise four components (i.e. trip generation, trip distribution, mode choice, and traffic assignment) to forecast traffic flows and volumes among traffic zones (McNally 2007). However, with continued urban development there was increased need for more sensitive forecasting models with better implementation capability for policy analysis and decisions. Transport planners established a new generation of travel demand models that can capture the complex behavior of travelers at the disaggregate level (Goran 2001; Bhat et al. 2004). From 1977, second generation models known as disaggregate trip based models and third generation models known as activity-based travel demand models were developed. Broadly, disaggregate trip based models, unlike four stage models, analyze each individual trip as independent. Disaggregate trip based models were utilized widely in the assessment of many large-scale projects world-wide during the 1980's and 1990's (Goran 2001). Further analyses on the results of disaggregate trip based models and comparison with actual daily life decisions of travelers revealed that there are

[4] An earlier version of this chapter has been published:
Hafezi, M. H., L. Liu., and H. Millward. (2018). "Modeling activity scheduling behavior of travelers for activity-based travel demand models". Peer reviewed proceedings of the 97th Annual Meeting of Transportation Research Board (TRB), Washington, D.C., USA.

associations between trips and the activity participation of individuals (Ettema, Borgers and Timmermans 1993; Garling, Kwan and Golledge 1994; Kitamura, Chen and Pendyala 1997; Ben-Akiva and Bowman 1998b). Concurrently, due to a fast growth of car usage, policy issues associated with complexities in individual travel behavior (i.e. flexible working hours, self-employment, e-shopping, etc.), road congestion, and air pollution became more important to analyze for transport planners and policy makers (Stopher, Hartgen and Li 1996; Ben-Akiva and Bowman 1998b; Fosgerau 2002; De Palma et al. 2011; Daisy, Liu and Millward 2017a).

During past 30 years, interest in employing activity-based travel demand models has increased considerably due to the growing importance of testing complex policy measures. Activity-based models focus on the 24-hour activity schedule, and travel episodes are linked with activities performed by individuals (Kitamura et al. 2000; Bhat et al. 2004; Auld, Mohammadian and Doherty 2009; Hafezi, Millward and Liu 2018b). Generating more accurate activity patterns decreases uncertainty in generating temporal information for the scheduling engine in activity-based models (Rasouli and Timmermans 2012; Hafezi, Liu and Millward 2017a; Daisy, Liu and Millward 2017b). As of today, researchers employ various techniques for producing temporal information associated with individual's daily activity schedules, such as probability distribution function, hazard function, and decision tree.

Over the past fifteen years, machine learning techniques have become more popular, and computers are more efficient at storing and processing large amounts of data. However, there have been only limited efforts to incorporate such techniques into activity-based travel demand models. The current study presents a new modeling framework utilizing a

well-known machine learning technique, the Random Forest algorithm. The goal is to learn and predict temporal information associated with travelers' activities in their daily agenda. The results of this study are expected to be implemented within the activity-based travel demand model, Scheduler for Activities, Locations, and Travel (SALT). The SALT model is comprised of five main components: population synthesizer, time-use activity pattern recognition, tour mode choice, activity destination choice, and activity/trip scheduling. The model adopts a pattern recognition approach which identifies population clusters with homogeneous time-use activity patterns. A series of behaviorally realistic econometric models and rule-based models are then developed for modeling time-use activity patterns in each identified cluster. Finally, this study contributes by providing additional insights to the scheduling modules in the overall activity-based modeling framework. The remainder of the study is structured as follows: following the literature review, a discussion of the data used in the modeling is presented. The methods for modeling activity-travel scheduling behaviors are described in the next section, followed by a discussion of model results. The study concludes by providing a summary of contributions and future research directions.

6.2 Literature Review

Over the last half century, many theoretical and practical activity-based travel demand models have been developed. Broadly, these models can be classified into three major groups based on their structure and purpose (De Palma et al. 2011): constraints-based models, random utility models, and computational process models. In constraints-based models, individual's activity patterns are shaped based on Hagerstrand's time-geography theory (Hagerstrand 1970). In random utility models the time component is modelled as a

discrete component (Bhat et al. 2004), while in computational process models it is modeled as a continuous component (Arentze and Timmermans 2000).

These models include a series of universal components such as activity generator and scheduler, tour mode choice, tour and trip time of day, tour and trip destination, and network assignment. One way to capture the uncertainty of departure time and start time in the modeling of individual's scheduling of travel behavior is to generate more accurate and homogeneous activity patterns (Oberkampf et al. 2002; Rasouli and Timmermans 2012). A wide array of theory and methods have been developed to produce information for the scheduling module in activity-based travel demand models. For instance, In CARLA (Combinatorial Algorithm for Rescheduling Lists and Activities), activities are generated and added to the individual's schedule using four rules: logical rules that refer to the presumption of one unique activity at a time at one location, environmental rules that refer to authority constraints (access time restrictions to different places), and travel times between locations, inter-personal rules that refer to coupling constraints (joint activities with other household members), and personal rules that refer to personal preferences (Jones et al. 1983). In STARCHILD (Simulation of Travel/Activity Responses to Complex Household Interactive Logistic Decisions), activities are generated in three steps. First, all possible alternatives to participating in different activities with respect to all constraints are explored. Next, through a series of statistical tests, similar alternatives are clustered in three to ten groups. Finally, a representative activity travel pattern is chosen for each group. The ultimate activity choices are estimated by the multinomial logit model and activities are scheduled through employing a series of rules (Recker, McNally and Root 1986a; Recker, McNally and Root 1986b).

In the cognitive model, alternative decisions for shaping individual's agenda at various levels of abstraction are generated through the application of a series of rules for activity planning processes. The cognitive model of planning can be accounted as the first rule-based simulation activity-based model (Hayes-Roth and Hayes-Roth 1979). In AMOS (Activity Mobility Simulator), the choice of different potential activities for the individual is generated from alternative activity travel patterns. Activities are ordered in agenda through a rule-based scheduling engine, and an activity adjuster is used for conflict resolution. Activity purposes, frequencies, time budget, duration, location, and priority list are inter-connected in AMOS (Kitamura et al. 1996). In SMASH (Simulation Model of Activity Scheduling Heuristics), a set of alternative activity travel patterns along with type, timing, travel mode, travel time, and location for each activity are generated at the first step. Next, the searching process considers ties in activity time and adds the activities that have been prioritized as high in the schedule. Decisions for adding or rescheduling activities in SMASH are made based on the choice of activities, sequencing, travel mode, travel time, location, and choice of joint activity (Ettema, Borgers and Timmermans 1993).

Some researchers have used explanatory data analysis and statistical methods to generate activities. For instance, in TASHA (Travel and Activity Scheduler for Household Agents), activities with similar socio-demographic and temporal characteristics (e.g. start time) are grouped into classes through a series of empirical data analysis. Start time and activity duration are generated through a probability distribution function. A series of heuristic rules (e.g. add, delete, shift or truncating activities) are used to schedule individual's agenda (Miller and Roorda 2003). In ADAPTS (Agent-Based Dynamic Activity Planning and Travel Scheduling Model), activities with similar characteristics are identified through

a hazard function, and essential information for the scheduling engine, such as activity start times and durations, are generated. Activities are added to individual's agenda using a set of heuristic rules based on the TASHA model (Auld and Mohammadian 2009). In ALBATROSS (A Learning-Based Transportation Oriented Simulation System), activities are generated and added to individual's agenda based on their flexibility level and spatial-temporal constraints. These constraints include location, travel mode, and time budget availability. Start time and activity duration are predicted using the CHAID decision tree. The scheduling engine in ALBATROSS uses these constrains to order activities. Fixed activities are added to individual's agenda at first and flexible activities are then scheduled with respect to prior fixed activities (Arentze and Timmermans 2004).

In recent years there has been growing interest in incorporating machine learning techniques in the modeling of activity generation and activity scheduling steps in activity-based travel demand models (Liao et al. 2007; Allahviranloo 2016; Hafezi et al. 2017; Li and Lee 2017). For instance, Allahviranloo (2016) used a k-mean clustering algorithm to identify unique clusters and utilized the AdaBoost algorithm to predict start time and activity duration based on the socio-demographic characteristics of individuals. In another study, a probabilistic context-free grammars technique is adopted to produce a set of activities in individual's daily agenda (Li and Lee 2017). Also, hierarchical markov models have been used for predicting individual's activity choices (Liao et al. 2007).

Despite all of the progress made in generating temporal information for the scheduling engine in activity-based travel demand models, there is undeniably considerable room for improving models' performance in terms of estimation accuracy, computational efficiency, and their practical application. In this study, we introduce a new modeling

framework for predicting temporal information associated with the traveler's daily activity. The framework may also be adapted for modeling other activity-based model components, such as transport mode, and work and residential location choice models.

6.3 Data

This study uses time-diary and GPS geo-coordinate data, from the Space-Time Activity Research (STAR) survey undertaken in Halifax, Canada. The STAR survey represents the world's first large-scale employment of global positioning system (GPS) technology for a household activity survey. The unique and rich Halifax STAR project produced a wide variety of data, including the household roster data, main file, vehicle data, time diary (episode and summary data file), activity diary (episode data file), land use database, business hours survey data, places and locations (PAL) directory data, and global positioning systems (GPS) data. Full descriptions of the survey design and the socio-demographic features of respondents can be found in (TURP 2008; Millward and Spinney 2011). The Halifax STAR project produced survey data from 1,971 randomly designated households in Halifax Regional Municipality (HRM) between April 2007 and May 2008. A primary respondent over age 15 was randomly selected in each household, and completed a 2-day time diary, supplemented and verified through GPS tracking. In this study, nine aggregated classes of activities were included in the proposed model: home chores, home leisure, night sleep, workplace, shopping & services, school/college, organizational/hobbies, entertainment, and sports activities. The final data set after data cleaning comprised 2,778 person-days (1,389 individuals, two days each).

In this study, we use data from twelve clusters of respondents previously-established through application of a novel pattern recognition modeling framework to the STAR data. Table 6.1 and Table 6.2 presents an analysis of the cluster data, showing socio-demographic characteristics and cluster membership, while Figure 6.1 and Figure 6.2 illustrates the observed temporal pattern of individual activities for identified worker and non-worker clusters, respectively. Interested readers are referred to previous research by the authors (Hafezi, Liu and Millward 2017b; Hafezi, Liu and Millward 2017c) for more details on clustering methods and execution of the pattern recognition framework.

Overall, the pattern recognition model recognized six clusters for out-of-home workers (cluster#1: extended worker, cluster#3: 8-4 worker, cluster#6: shorter worker, cluster#7: 7-3 worker, cluster#10: evening worker and cluster#11: 9-5 worker), four clusters for non-worker and non-student (cluster#2: non-worker midday activities, cluster#4: non-worker evening activities, cluster#8: non-worker morning shopping and cluster#9: non-worker afternoon shopping), one cluster for students (cluster#12: students), and one cluster for individuals who mostly spend their time at home (cluster#5: stay-at-home). Each cluster carries various information associated to the travelers' daily activity patterns such as activity type, activity sequencing, probability of start time, and activity duration. Individuals in each cluster differ considerably from individuals in other clusters with regards to their socio-demographic characteristics. Moreover, individuals within each cluster have homogeneous activity sequences, whereas there is a heterogeneous diversity between clusters in terms of their distributions of activity type, activity start time, activity duration, and socio-demographic features.

Table 6.1 Share of various socio-demographic variables for twelve respondent clusters

Social demographic variables		Sample mean (%)	Mean of cluster (%)											
			#1	*#2*	*#3*	*#4*	*#5*	*#6*	*#7*	*#8*	*#9*	*#10*	*#11*	*#12*
Gender	Female	0.53	0.53	0.53	0.44	0.59	0.56	0.52	0.47	0.54	0.59	0.51	0.53	0.60
Age	Young adults (ages 15-35 years)	0.10	0.12	0.05	0.10	0.10	0.11	0.10	0.05	0.09	0.07	0.15	0.09	0.60
	Middle-aged adults (ages 36-55 years)	0.49	0.67	0.29	0.66	0.38	0.32	0.71	0.72	0.29	0.32	0.66	0.70	0.31
	Older adults (aged older than 55 years)	0.41	0.20	0.66	0.24	0.53	0.57	0.19	0.23	0.63	0.61	0.19	0.22	0.08
Education	Diploma or university certificate	0.67	0.76	0.58	0.76	0.62	0.66	0.85	0.53	0.57	0.65	0.64	0.80	0.38
Occupation	Regular shift	0.53	0.73	0.22	0.93	0.26	0.24	0.87	0.93	0.19	0.24	0.43	0.89	0.13
	Irregular schedule	0.10	0.22	0.10	0.03	0.10	0.11	0.09	0.07	0.07	0.07	0.47	0.08	0.02
	Student	0.03	0.01	0.00	0.00	0.04	0.01	0.01	0.00	0.03	0.02	0.04	0.01	0.67
	Retired	0.23	0.02	0.52	0.02	0.39	0.41	0.01	0.00	0.53	0.41	0.00	0.00	0.08
	Work at home	0.15	0.23	0.10	0.13	0.15	0.16	0.30	0.06	0.11	0.09	0.09	0.26	0.02
Flexible schedule	Have no flexibility in a work schedule	0.50	0.55	0.48	0.54	0.46	0.43	0.44	0.63	0.48	0.51	0.75	0.40	0.43
Job number	Have more than one job	0.07	0.09	0.04	0.04	0.16	0.08	0.05	0.07	0.11	0.05	0.08	0.08	0.00
Income	Low-income (<= $ 40,000)	0.39	0.28	0.44	0.22	0.48	0.49	0.32	0.29	0.53	0.48	0.47	0.26	0.78
	Middle-income ($ 40,000 - $ 100,000)	0.53	0.60	0.46	0.68	0.45	0.45	0.55	0.64	0.42	0.46	0.49	0.59	0.19
	High-income (> $ 100,000)	0.09	0.12	0.10	0.10	0.07	0.07	0.13	0.08	0.05	0.06	0.04	0.15	0.03

Table 6.2 Summary statistics for twelve respondent clusters: membership analysis

	Cluster number											
	#1	*#2*	*#3*	*#4*	*#5*	*#6*	*#7*	*#8*	*#9*	*#10*	*#11*	*#12*
Total cluster membership	137	225	401	238	419	171	229	247	262	53	348	48
Percentage in total (number of person-days)	4.93	8.10	14.43	8.57	15.08	6.16	8.24	8.89	9.43	1.91	12.53	1.73
Home chores (%)	25.09	34.73	27.81	41.39	43.10	33.78	30.14	40.54	40.12	35.01	29.47	35.52
Home leisure (%)	12.04	20.26	16.57	14.42	19.17	16.22	17.73	19.96	19.18	13.61	15.32	13.81
Night sleep (%)	62.88	45.01	55.62	44.19	37.73	50.00	52.13	39.50	40.70	51.38	55.21	50.67
Total in-home (%)	*100.0*	*100.0*	*100.0*	*100.0*	*100.0*	*100.0*	*100.0*	*100.0*	*100.0*	*100.0*	*100.0*	*100.0*
Workplace (%)	**89.26**	4.76	**82.61**	6.41	8.16	**79.53**	**89.31**	10.53	13.12	**90.90**	**86.56**	5.77
Shopping & services (%)	1.61	27.27	3.16	15.31	30.19	5.56	2.41	**32.26**	**38.27**	2.90	2.98	3.38
School/college (%)	0.00	1.00	0.14	1.20	0.42	0.40	0.29	3.16	1.28	0.66	0.15	**69.29**
Organizational/hobbies (%)	2.34	**29.57**	3.24	**31.28**	19.90	5.11	1.76	21.11	21.94	1.57	2.62	11.14
Entertainment (%)	4.67	17.49	6.92	**33.73**	10.35	6.48	3.73	8.75	11.66	1.93	4.30	4.53
Sports (%)	2.12	19.90	3.93	12.06	**30.98**	2.93	2.50	24.20	13.74	2.04	3.40	5.89
Total out-of-home (%)	*100.0*	*100.0*	*100.0*	*100.0*	*100.0*	*100.0*	*100.0*	*100.0*	*100.0*	*100.0*	*100.0*	*100.0*

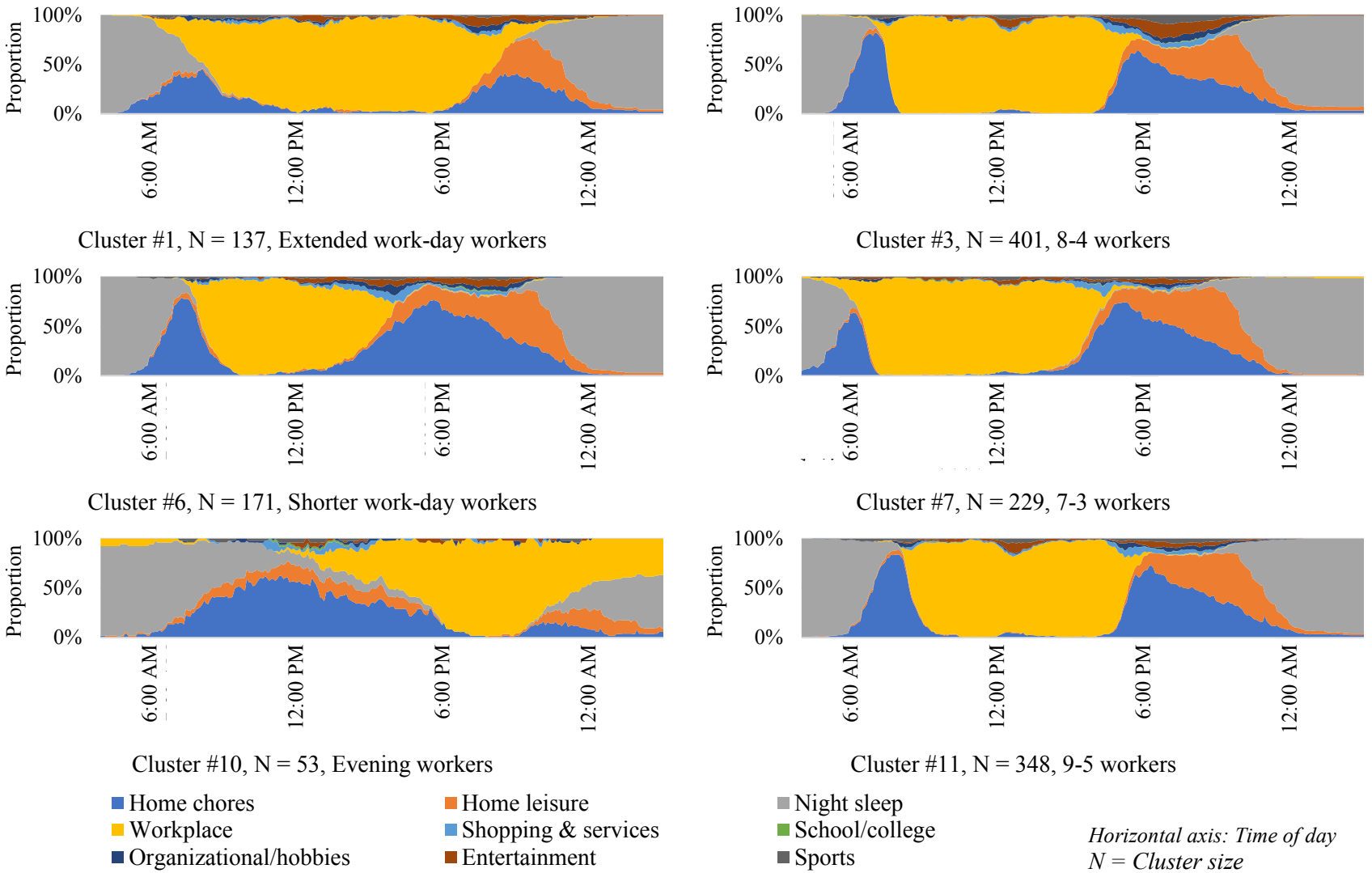


Figure 6.1 Observed temporal pattern of individual activities for six identified worker clusters

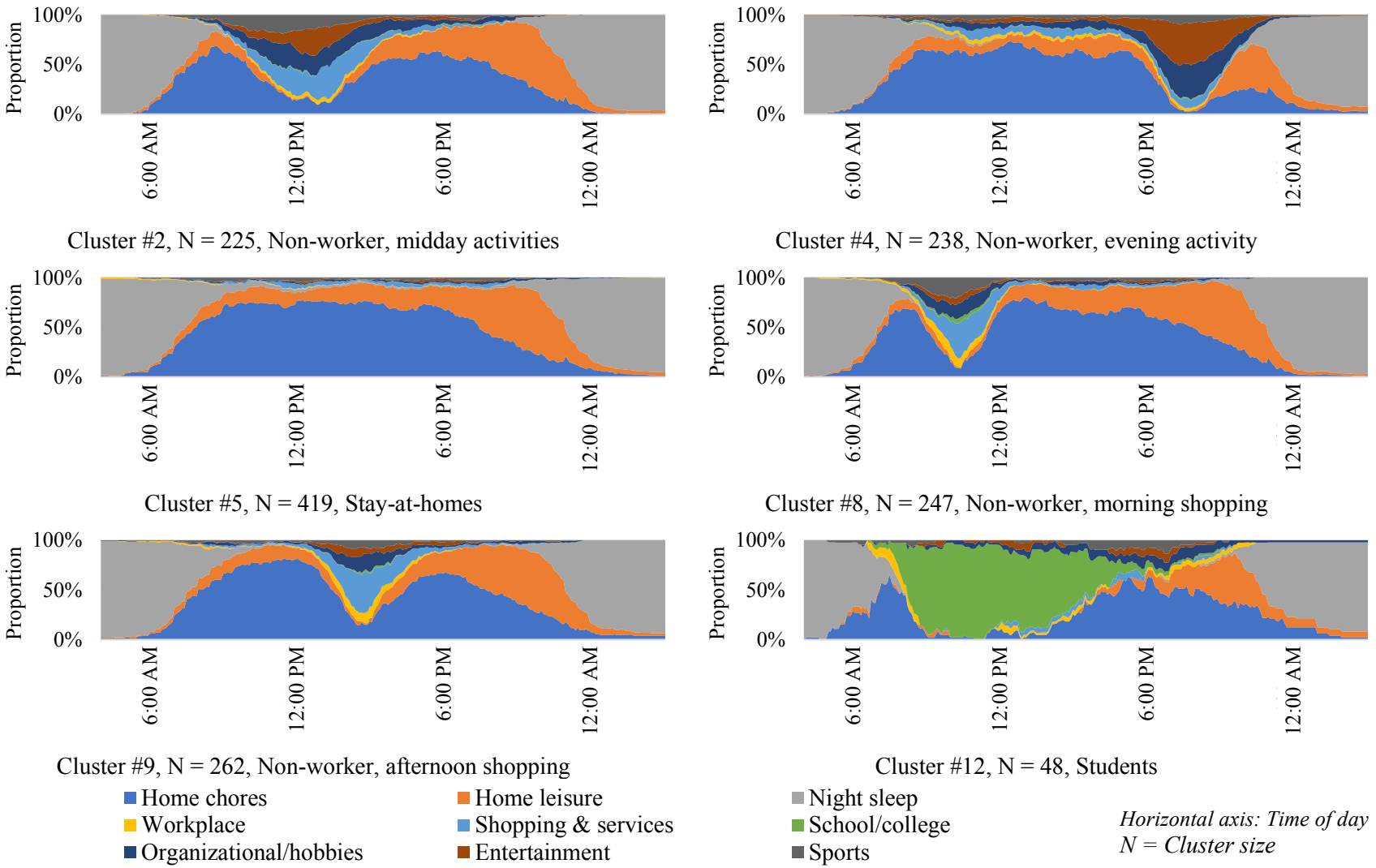


Figure 6.2 Observed temporal pattern of individual activities for six identified non-worker clusters

6.4 Methods

The proposed modeling framework for scheduling travelers' activities in this study consists of two steps, as follows. First, temporal information for the set of activities in the agenda is predicted using the Random Forest (RF) model. Details of the methodology for inferring activity types and undertaking rigorous validation can be found in (Hafezi, Liu and Millward 2018a). Given a predicted set of activities in the traveler's agenda, in the next phase we predict start time and activity duration of each activity in the agenda, and schedule them. Predicted activities are inserted into a skeleton schedule through a heuristic decision rule-based technique, and are scheduled with respect to two-tier constraints. Random Forest theory is based on the use of numerous decision trees that have been grown using the bagging technique (Breiman 2001; Suthaharan 2015). Each tree acts as a weak learner in the algorithm and the aggregation of these weak inputs provides a powerful ensemble learning model. Outcomes achieved from the RF model are based on the majority votes in the ensemble models. Although the RF method is used in other transportation fields such as traffic incident detection and transport mode recognition, to the best of our knowledge RF models using the CART classifier have not previously been employed for predicting start time and activity duration in travel behavior analysis.

6.4.1 The Random Forest (RF) Model

The Random Forest (RF) structure for predicting start time and activity duration is shown in Figure 6.3. The RF theory is based on an ensemble of many decision trees (Breiman 2001; Suthaharan 2015). Each tree acts as a weak learner and makes a prediction $\{\hat{P}_1, \hat{P}_2, \hat{P}_3, \cdots, \hat{P}_M\}$. The eventual prediction outcome gained is based on the majority votes

for $\hat{P}$. The start time and activity duration for each activity type based on their duration interval are transformed to a set of bins and shown in Table 6.3. The predictor variables Y_n are socio-demographic characteristics of travelers and the corresponding start time or duration bin numbers for each activity type in the agenda (selected from Table 6.4). The response variables X_i are defined as one of the activity start time /activity duration bin numbers in the model (selected from Table 6.4).

Generally, the RF model takes the following steps. First, test dataset (T_e) and training dataset $D_r = ((X_i, Y_1), \dots, (X_i, Y_N))$ are drawn from the primary dataset using a cross-validation partition process. At each node in the decision tree a square number of total existing predictor variables ($\sqrt{Q}$) are randomly selected from all the standing predictor variables (Q). Next, $E_n(C)$ observations are randomly drawn from the sampled data points and used for building each decision tree m in the RF model. We undertake $E_n(C)$ as a subset of training data D_r where C cut in C is as pair (u, w). u is the randomly nominated predictor variables. w is the place of the split along the u-th correspondent, within the limits of C. We hypothesize the set of all such possible cuts in C as T_C. The split condition $S_{clas,n}(u, w)$ is computed as follows (Breiman 2001; Biau and Scornet 2016):

$$S_{clas,n}(u,w) = \frac{1}{E_n(C)}\sum_{i=1}^{n}(Y_i - \bar{Y}_C)^2 \exists_{X_i \in C} - \frac{1}{E_n(C)}\sum_{i=1}^{n}(Y_i - \bar{Y}_{C_L} \exists_{X_i^{(u)} < w} - \bar{Y}_{C_R} \exists_{X_i^{(u)} \geq w})^2 \exists_{X_i \in C} \quad (1)$$

$$X_i = \left(X_i^{(1)}, \dots, X_i^{(p)}\right) \forall (u,w) \in T_C \quad (2)$$

$$C_L = \{x \in C : x^{(u)} < w\} \quad (3)$$

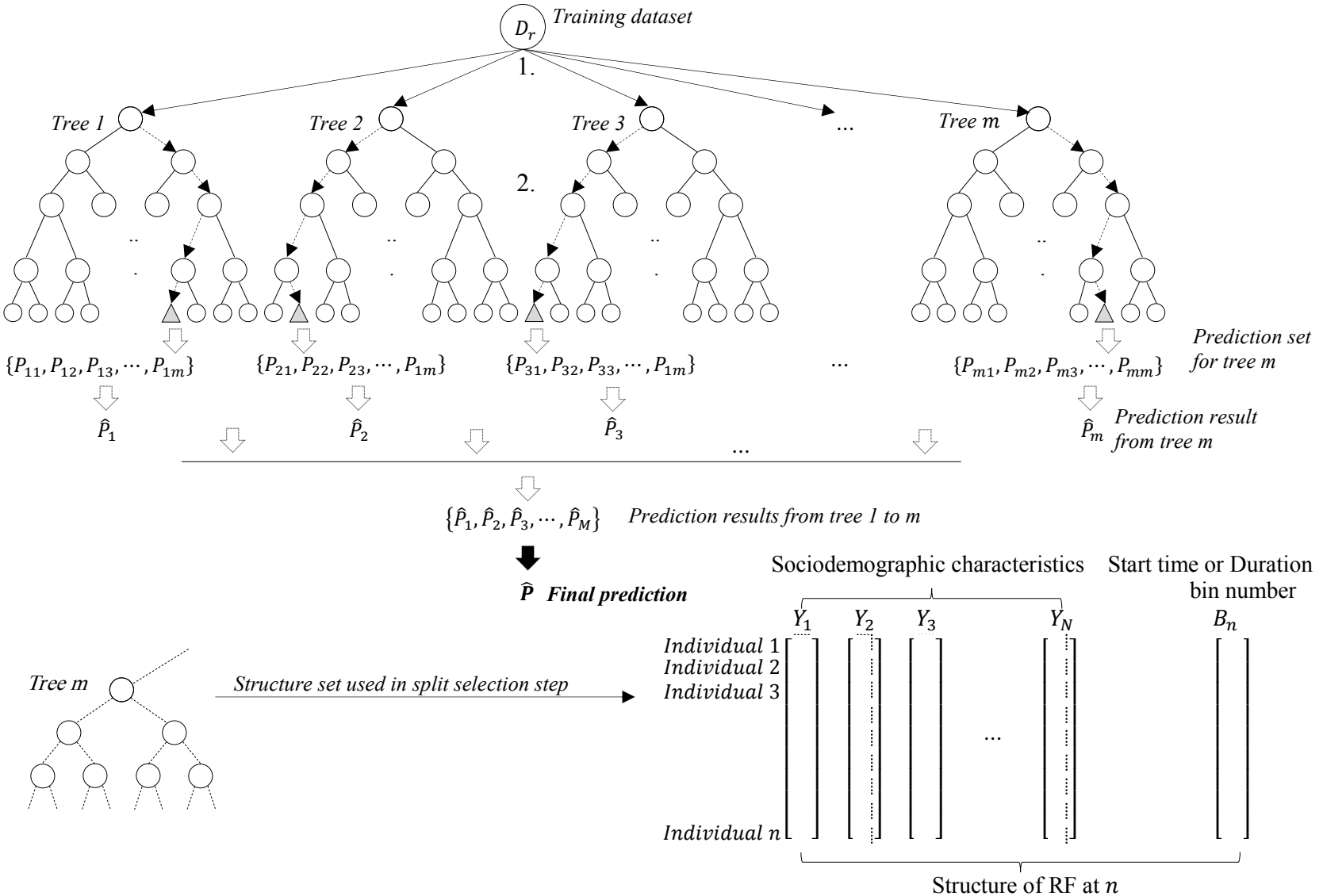


Figure 6.3 Random forest structure for predicting temporal information associated with the traveler's daily activity

$$C_R = \{x \in C : x^{(u)} \geq w\} \quad (4)$$

where $\bar{Y}_C$ is the average of the Y_i such that X_i belongs to C.

Table 6.3 Activity start time and activity duration bin structure

	Setting type	No. of bins	Interval duration	Pattern
Start time	I	144	10	Bin 1: [4:00-4:10], bin 2: [4:10-4:20], bin 3: [4:20-4:30], …, bin 48: [3:50-4:00].
	II	96	15	Bin 1: [4:00-4:15], bin 2: [4:15-4:30], bin 3: [4:30-4:45], …, bin 48: [3:45-4:00].
	III	48	30	Bin 1: [4:00-4:30], bin 2: [4:30-5:00], bin 3: [5:00-5:30], …, bin 48: [3:30-4:00].
	IV	8	180	Bin 1: [4:00-7:00], bin 2: [7:00-10:00], bin 3: [10:00-13:00], …, bin 8: [1:00-4:00].
Activity duration	Î	96	15	Bin 1: [less than 15 min], bin 2: [between 15 and 30 min], bin 3: [between 30 and 45 min], …, bin 24: [more than 1425 min].
	ÎÎ	48	30	Bin 1: [less than 30 min], bin 2: [between 30 min and 1 hour], bin 3: [between 1 and 1.5 hours], …, bin 24: [more than 1410 min].
	ÎÎÎ	24	60	Bin 1: [less than 1 hour], bin 2: [between 1 and 2 hours], bin 3: [between 2 and 3 hours], …, bin 24: [more than 1380 min].
	ÎV	4	360	Bin 1: [less than 6 hours], bin 2: [between 6 and 12 hours], bin 3: [between 12 and 18 hours], bin 4: [more than 1080 min].

**Modeling period is from 4 a.m. until 3:55 a.m. next day*

Table 6.4 Proposed predictor variables for predicting temporal information associated with the traveler's daily activity

Predictor	Subcategories
Gender	Male, female
Age	15 to 19, 20 to 24, 25 to 29, 30 to 34, 35 to 39, 40 to 44, 45 to 49, 50 to 54, 55 to 59, 60 to 64, 65 to 69, 70 to 74, 75 to 79, 80 to 84, 85+
Marital status	Married, living common-law, widowed, separated, divorced, single-never married,
Household size	1,2,3,4,5,6
Highest education level	Masters or earned doctorate, bachelor or undergraduate degree, diploma or certificate, some university, some community college, trade, technical or business college, high school/secondary, other
Full/part-time student	Full-time student, part-time student
Paid/self employed	A paid worker, self-employed, an unpaid family worker
Flexible work schedule	No, yes
Work at home	No, yes
Total personal income	Under $20,000, $20,000–$39,999, $40,000–$59,999, $60,000–$79,999, $80,000–$99,999, $100,000 or more
Total household income	Under $20,000, $20,000 - $39,999, $40,000 - $59,999, $60,000 - $79,999, $80,000 - $99,999, $100,000 or more
Dwelling type	Single unit residential, duplex or semi-detached, townhouse, multi-unit residential-less than 6 stories, multi-unit residential-6 or more stories, mobile dwelling, other-specify
Dwelling owned/rented	Owned, rented
Valid driver's license	No, yes
Buss pass	No, yes
Number household vehicles	0,1,2,3,4,5,6
Number household motorcycles	0,1,2,3,4,5,6,7,8,9,10
Number household bicycles	0,1,2,3,4,5,6,7,8,9,10
Usually mode to work	Car, truck or van - as driver, car, truck or van - as passenger, public transit, walk to work, bicycle, motorcycle, taxicab, other method
Usual mode to school	Car, truck or van - as driver, car, truck or van - as passenger, public transit, walk to work, bicycle, motorcycle, taxicab, other method
State of health	Excellent, very good, good, fair, poor
Activity start time	Corresponding duration bin number for each activity type in the agenda
Activity duration	Corresponding start time bin number for each activity type in the agenda

One of the main challenges in most decision tree models is to select the best split predictor method (Breiman 2001; Biau and Scornet 2016). In this study, we used the CART algorithm for decision splits in the RF model. The best cut (u_n^*, w_n^*) is computed by maximizing $S_{clas,n}(u, w)$ over B_{lit} and T_C.:

$$(u_n^*, w_n^*) \in \underset{\substack{u \in B_{lit} \\ (u,w) \in T_C}}{argmax} S_{clas,n}(u, w) \tag{5}$$

This process is terminated when all the information is saturated. The new input data at the testing stage $t_n \in \{1, \dots, e\}$ is propagated down to all of the trees in the RF model and each tree makes a prediction. The ultimate prediction result gained is based on the majority votes for $\hat{P}$.

6.4.2 Variable Importance Measures: Mean Decrease Accuracy (MDA)

One of the advantages of the RF model is its ability to measure the importance of variables and subsequently rank them in order to guide a decision split algorithm for finding the best cut points. This process in the RF model can be performed through Mean Decrease Accuracy (MDA) or Mean Decrease Impurity (MDI) computations. In this study, we measured the importance level of predictor variables using the Mean Decrease Accuracy (MDA) method that builds on the out-of-bag (OBB) error estimate (Breiman 2001; Biau and Scornet 2016). The MDA of the variable $X^{(u)}$ is computed by balancing the difference in OOB error (O_n) prior and subsequent to the permutation over all trees. The primary OOB error of each tree is calculated by testing the RF model using OOB data. The later OOB error of each tree is computed by adding noise to the sample data of the feature

randomly and retesting the OOB error. The MDA for randomly nominated variable $X^{(u)}$ is computed as follows:

$$\text{MDA}(X^{(u)}) = \frac{1}{m}\sum_{l=1}^{m}\left[O_n\left[d_n(.;\Theta_m), I_{m,n}^{u}\right] - O_n\left[d_n(.;\Theta_m), I_{m,n}\right]\right] \quad (6)$$

$$O_n[m_n(.;\Theta_l), I] = \frac{1}{|I|}\sum_{i:(X_i,Y_i)\in I}(Y_i - m_n(X_i;\Theta_l))^2 \quad (7)$$

$I_{m,n}$ is the out-of-bag data set of the m-tree, $I = I_{m,n}^{u}$ is the permutated data for variable u and $O_n(.;\Theta_m)$ is the estimation for the m-th tree. In total 70% of the dataset was used for training the model and 30% for testing model performance.

6.4.3 Model Calibration and Validation

In the RF algorithm several parameters need to be initialized and calibrated. These are the initial number of trees, the node split principle, and the cutoff vector. The initial number of trees m is set to 1000, and was calibrated by out-of-bag error estimation to verify if the model can be converged within this value. In this study, the best split at each node B_{lit} is determined with the CART algorithm. Another alternative approach to find the best split in the RF model is the Curvature Search technique. The cutoff vector T_C is a vector of length equivalent to the number of classes. T_C includes three sub-parameters ($c1$, $c2$, $c3$) that are originally each randomly set in the range between [0,1] with the requirement that the total sum is equal to 1. These parameters were cross-validated through the OOB error rate in order to achieve the best parameter values for the proposed RF model.

6.4.4 Decision Rule-Based Algorithm

Having predicted activity agendas and activity sequences of the traveler along with predicted start time and activity duration bin numbers for each activity type in the traveler's agenda for a 24-hour period, in this step activities are inserted into the skeleton schedule through a rule-based algorithm, and a 24 hours schedule is constructed. A conceptual framework of the scheduling model is shown in Figure 6.4.

The algorithm is started by predicting cluster membership for the selected individual. This process is done through a Classification and Regression Tree (CART) developed based on the socio-demographic characteristics of individuals in each cluster. Based on the random generated number and cumulative probability functions, the CART algorithm can find specific clusters for particular leaf nodes based on the high probability that an individual belongs to it. Interested readers can refer to earlier work by the authors (Hafezi, Liu and Millward 2017b; Hafezi, Liu and Millward 2017c) for more details on this step. Next, activity agendas and activity sequences of the traveler are predicted using the advanced RF model (Hafezi, Liu and Millward 2018a). For each activity type, start time and activity duration are generated from a uniform distribution within their interval time range as defined in Table 6.3. For instance, if bin number 1 from setting type III is predicted for start time, the algorithm will generate a random start time from its interval [4:00-4:30]. Then, a rule-based algorithm is used to insert activities into the skeleton schedule based on two conditions. First, the importance level of the activity is determined based on the cluster's representative pattern characteristics. For example, work activity has the highest priority compared to other activities for individuals in the worker clusters. Second, longer

non-mandatory activities (i.e. shopping and hobbies) have the higher rank compared to shorter activities.

A major difficulty with most activity-based models is that travel times are unknown since the activity location, transport modes, and trip-chaining characteristics are not yet determined. Thus, travel times are crucial for determining whether an activity fits in a given time slot or not. In this study, we used the median value for travel times. A conceivable further step of this work is to determine the above-mentioned characteristics and update more realistic travel times in the scheduling process. When there is a discrepancy between start time and activity duration of inserted activities, the rule-based algorithm makes a decision based on the prediction results of activity durations and correct activity start times.

Ultimately, a heuristic advance adjustment algorithm will be applied to predicted start times and durations made in the scheduling step where the activities are inserted in the schedule. At present, however, this phase of the scheduling process is limited to resolving time conflicts and making the schedule consistent. For this purpose, the representative pattern in each cluster is targeted as a benchmark and subsequently, the heuristic algorithm will rearrange the conflicted episodes by adding, removing, or truncating them in a way that the time-use aggregation of all travelers in the cluster is as close as possible to the cluster's representative pattern benchmark. The algorithm is terminated when there are no time overlaps between activities and travel episodes, and the episodes sum to 1440 minutes (24 hours).

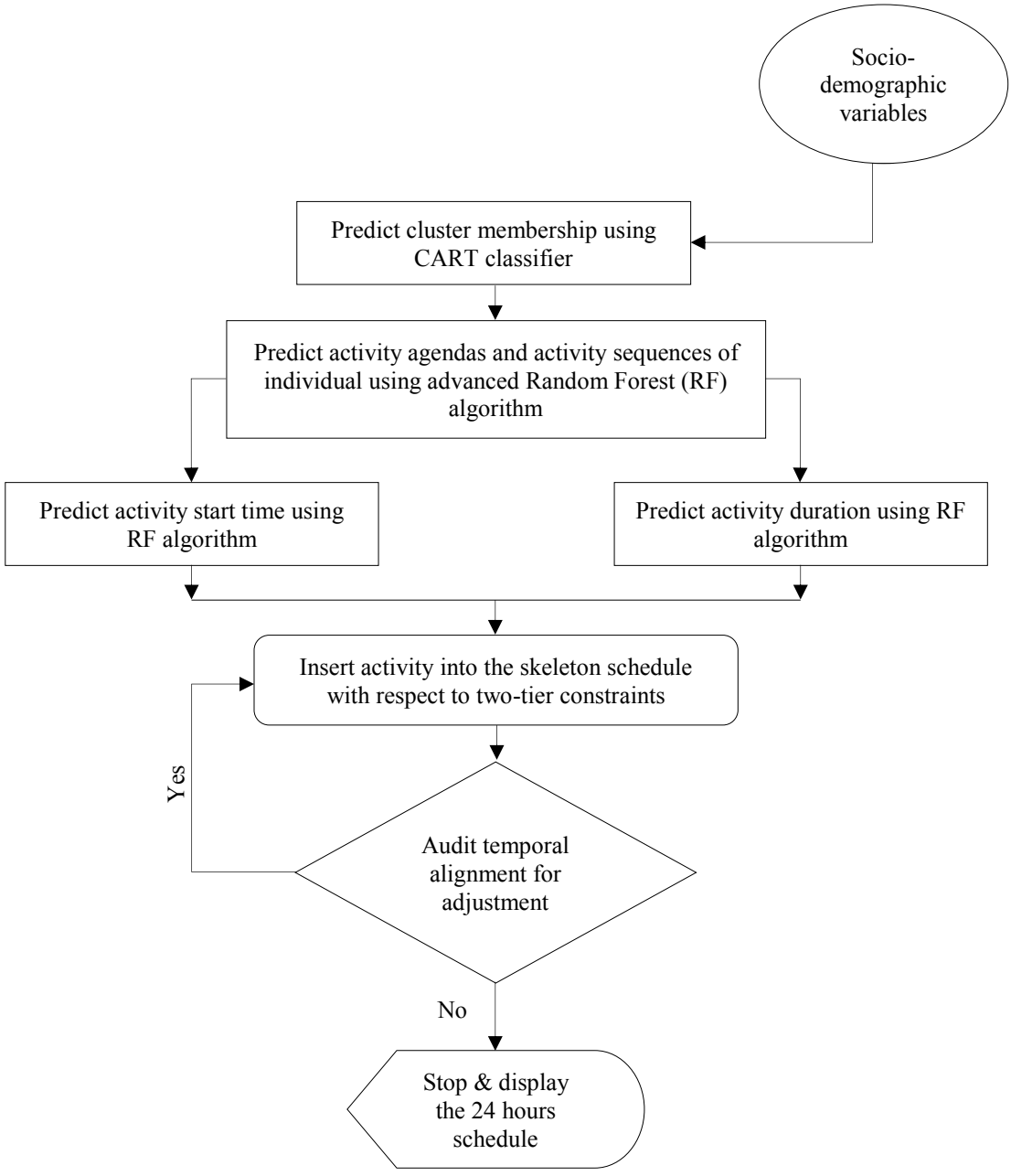


Figure 6.4 Conceptual framework for the scheduling model

6.5 Discussion of Results

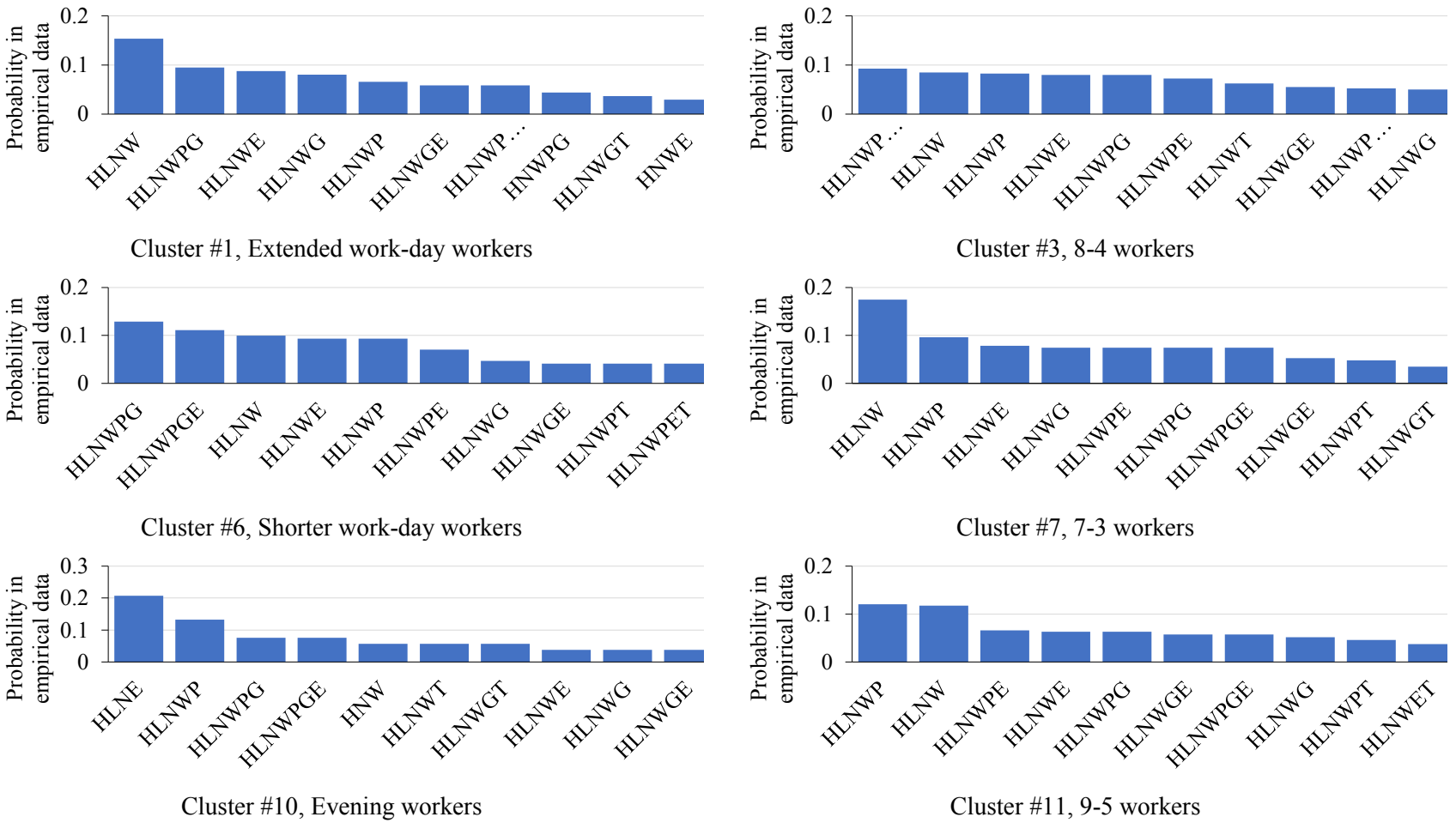
In order to analyze the efficiency and performance of RF model under different conditions and complex activity patterns, we applied the models to twelve clusters drawn from the Space-Time Activity Research (STAR) survey. In total, six clusters were recognized for

out-of-home workers (cluster #1: extended work-day workers; cluster #3: 8-4 workers; cluster #6: shorter work-day workers; cluster #7: 7-3 workers; cluster #10: evening workers; cluster #11: 9-5 workers), four clusters for non-worker non-student (cluster #2: non-worker, midday activities; cluster #4: non-worker, evening activity; cluster #8: non-worker, morning shopping; cluster #9: non-worker, afternoon shopping), and separate clusters for students (cluster #12: students) and individuals who mostly spend their time at home (cluster #5: stay-at-homes). The size of cluster varied in the range of 48 to 419. Further cluster analysis and statistical tests showed that there is heterogeneous diversity among clusters in terms of their distributions of start time, activity duration, activity type, and socio-demographic characteristics. Interested readers can refer to earlier work by the authors (Hafezi, Liu and Millward 2017b; Hafezi, Liu and Millward 2017c) for detailed description of each cluster.

In the next step, a set of activity types in travelers' agenda was predicted using the advanced RF model (Hafezi, Liu and Millward 2018a). The number of predicted activities in the cluster was assumed to be equal to the median stop number in the observed cluster. Figure 6.5 and Figure 6.6 illustrates the distribution of the ten most frequent combinations of agenda in the travelers' daily activity patterns for identified worker and non-worker clusters, respectively. As can be seen, the dominant activities for worker and student clusters are work and school activity, respectively. For non-worker non-student groups, non-mandatory activities such as shopping, entertainment, and organizational activities are highly important in travelers' out of home daily activity patterns.

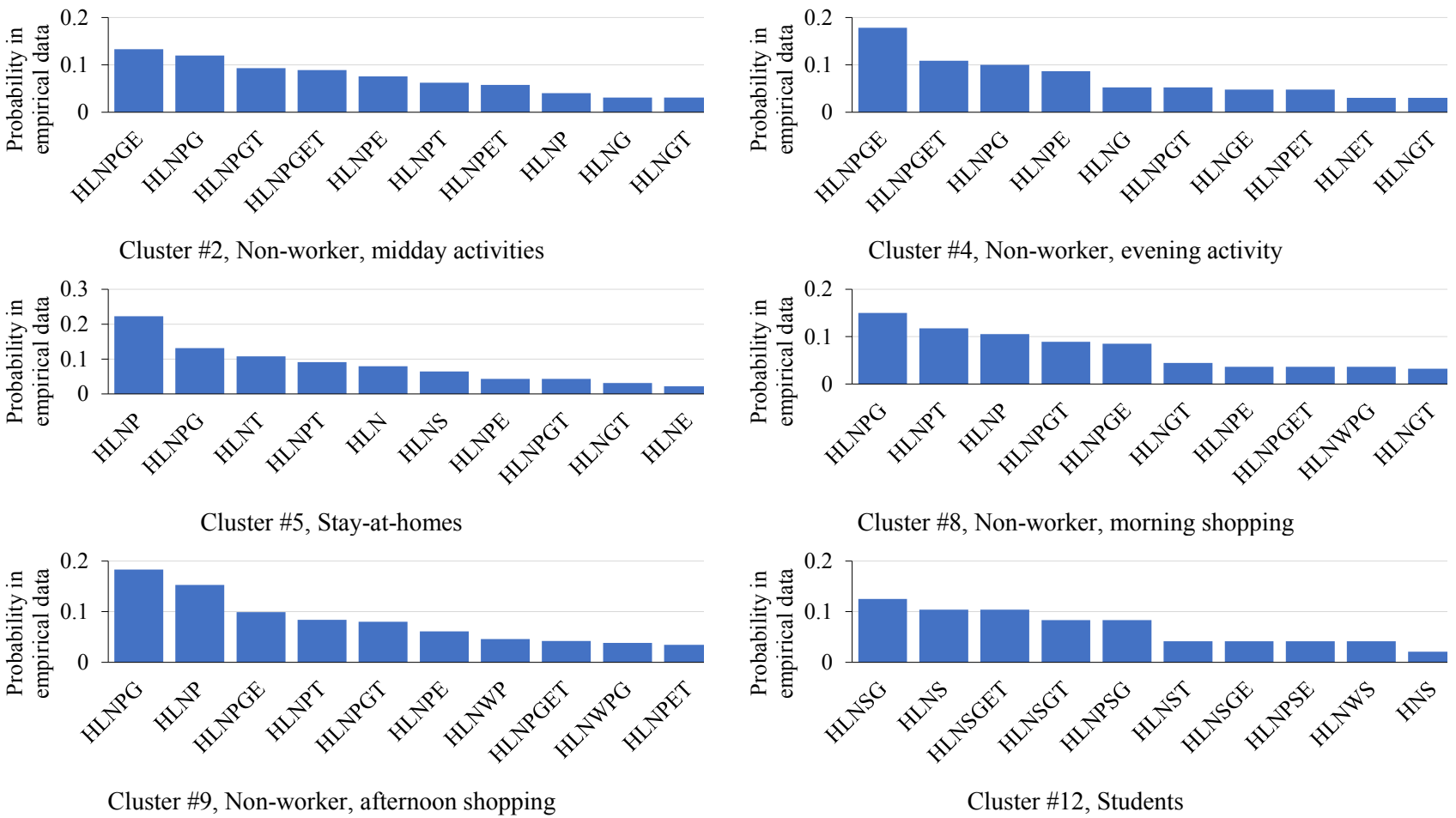
The RF model for predicting start time and activity duration was run under eight different bin settings for a twofold purpose: to test the efficiency level of the model and to compare

results with other alternative techniques. The OOB rate error cross-validation specifies that after $m > 850$ (m is the number of trees), the OBB error rate tends to be constant. Therefore, it is reasonable to accept m as 1000 at first. In view of this, we achieve the best optimal parameters for RF models as follows: $c1 = 0.25$, $c2 = 0.35$ and $c3 = 0.40$. For each cluster, the activity start time and activity duration (bin numbers) for all activity types in travelers' agenda are predicted. The model results for start time and activity duration prediction are presented in Table 6.5 and Table 6.6, respectively. The prediction accuracy is obtained from comparison of observed and predicted bin numbers for every activity type. The best result for predicting start time was obtained by setting IV (8 bins, each of duration 180 minutes), at 60.10% accuracy, followed by setting III (48 bins, each of duration 30 minutes), at 36.28%. Similarly, the best prediction result for activity duration was found under setting $\widehat{\text{IV}}$ (4 bins, each of duration 360 minutes), at 98.65%, followed by setting $\widehat{\text{III}}$ (24 bins, each of duration 60 minutes), at 67.32%. The empirical results show that with increases in the time interval (increasing bin numbers), the RF efficiency also increased. The lowest model result was reported for predicting start time with 144 bins, each of duration 10 minutes (setting I), at only 32.95% accuracy.



*Horizontal axis: Occasions (H=Home chores, L=Home leisure, N=Night sleep, W=Workplace, P=Shopping & services, S=School/college, G=Organizational/hobbies, E=Entertainment, T=Sports). **Agenda combinations shown in Figure 4 are regardless activity sequences*

Figure 6.5 Distribution of 10 most frequent combinations of agenda in the 24-hour day for six identified worker clusters



**Horizontal axis: Occasions (H=Home chores, L=Home leisure, N=Night sleep, W=Workplace, P=Shopping & services, S=School/college, G=Organizational/hobbies, E=Entertainment, T=Sports). **Agenda combinations shown in Figure 4 are regardless activity sequences*

Figure 6.6 Distribution of 10 most frequent combinations of agenda in the 24-hour day for six identified non-worker clusters

Table 6.5 Accuracy of activity start time estimation for test dataset

Activity type	Estimation accuracy for setting I (144 bins, each of duration 10 minutes)											
	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10	#11	#12
Home chores	7.98	27.40	25.01	7.53	7.94	10.77	30.46	7.65	26.00	37.96	12.26	33.78
Home leisure	14.81	16.91	12.01	5.56	53.57	16.47	41.92	9.62	27.19	-	7.42	-
Night sleep	48.41	39.70	36.69	56.25	55.00	44.91	37.67	40.74	36.21	85.00	33.08	34.30
Workplace	10.21	-	16.67	-	-	13.75	35.92	-	-	66.67	16.55	82.00
Shopping & services	-	16.67	32.00	13.35	-	22.86	25.00	9.30	13.69	-	10.00	33.33
School/college	-	-	-	-	-	-	-	-	-	-	-	31.25
Organizational /Hobbies	-	19.64	24.07	60.00	50.00	21.59	35.00	25.00	14.29	-	43.06	-
Entertainment	-	-	22.26	-	-	-	27.78	-	62.50	-	31.25	50.00
Sports	60.00	33.33	50.00	52.70	-	50.00	-	-	-	86.00	46.73	-
Activity type	**Estimation accuracy for setting II (96 bins, each of duration 15 minutes)**											
	#1	**#2**	**#3**	**#4**	**#5**	**#6**	**#7**	**#8**	**#9**	**#10**	**#11**	**#12**
Home chores	9.71	28.09	23.78	6.63	6.57	9.72	22.64	7.16	26.42	37.96	12.63	26.14
Home leisure	24.07	18.44	62.50	19.44	36.88	17.43	38.47	12.23	19.62	-	12.70	-
Night sleep	43.65	38.29	38.35	48.61	40.83	49.07	35.29	51.85	33.60	46.80	30.90	37.90
Workplace	8.27	-	26.67	-	-	18.75	35.67	-	-	66.67	21.04	46.30
Shopping & services	-	16.24	42.56	12.34	50.00	21.52	-	12.34	8.12	-	-	66.67
School/college	-	-	-	-	-	-	-	-	-	-	-	30.00
Organizational /Hobbies	14.29	18.65	21.30	60.00	50.00	27.27	27.50	11.11	19.64	-	66.67	33.33
Entertainment	16.67	11.11	20.03	25.00	65.70	49.80	33.33	-	62.50	-	37.50	50.00
Sports	78.60	23.81	66.67	85.60	-	-	33.33	33.33	-	67.30	47.86	50.00
Activity type	**Estimation accuracy for setting III (48 bins, each of duration 30 minutes)**											
	#1	**#2**	**#3**	**#4**	**#5**	**#6**	**#7**	**#8**	**#9**	**#10**	**#11**	**#12**
Home chores	19.96	21.91	25.16	10.90	9.99	11.73	25.70	12.99	19.82	37.96	20.39	40.13
Home leisure	25.00	16.78	30.80	14.33	25.62	19.91	28.50	17.90	25.62	62.50	26.39	-
Night sleep	47.02	39.55	56.25	62.50	41.88	51.39	45.29	41.67	50.30	88.00	35.30	76.00
Workplace	16.40	-	30.67	-	-	20.85	34.15	-	-	55.56	23.42	69.90
Shopping & services	25.00	20.89	34.46	15.99	27.04	18.05	25.00	16.47	19.18	-	14.29	66.67
School/college	-	-	-	-	-	-	-	-	-	-	-	42.50
Organizational /Hobbies	14.48	20.55	25.00	44.76	50.00	45.96	42.50	27.04	29.76	-	36.50	-
Entertainment	16.67	33.33	17.55	17.14	-	57.14	27.78	-	58.33	-	43.75	50.00
Sports	72.50	32.14	50.00	52.78	-	41.67	66.67	26.11	-	79.60	49.32	50.00
Activity type	**Estimation accuracy for setting IV (8 bins, each of duration 180 minutes)**											
	#1	**#2**	**#3**	**#4**	**#5**	**#6**	**#7**	**#8**	**#9**	**#10**	**#11**	**#12**
Home chores	43.90	43.39	51.53	37.73	36.27	41.52	52.04	38.03	47.72	45.12	55.33	62.05
Home leisure	43.33	33.73	56.14	36.14	39.28	48.21	59.67	49.15	46.02	43.75	44.68	77.78
Night sleep	68.06	67.64	48.10	43.75	42.76	57.78	49.66	63.89	64.48	76.67	44.83	80.00
Workplace	35.12	100.0	51.02	90.00	83.33	59.49	47.36	90.00	75.00	71.67	47.02	100.0
Shopping & services	44.58	40.86	53.35	35.30	33.75	49.98	51.19	54.49	49.46	52.38	54.56	86.11
School/college	-	-	-	-	-	100.0	-	100.0	100.0	-	100.0	57.92
Organizational /Hobbies	48.15	50.78	59.03	50.63	43.60	56.43	57.60	54.49	50.64	83.33	63.33	63.33
Entertainment	37.04	50.83	52.64	45.15	77.78	55.36	53.33	55.21	71.90	50.00	51.48	66.67
Sports	53.81	48.21	46.39	59.26	50.00	59.52	56.67	57.78	78.33	100.0	57.14	80.00

Mean accuracy 32.95%

Mean accuracy 33.25%

Mean accuracy 36.28%

Mean accuracy 60.10%

'-' indicates that algorithm didn't predict start time due to missing the particular activity type in model's input.

Table 6.6 Accuracy of activity duration estimation for test dataset

Activity type	Estimation accuracy for setting I (96 bins, each of duration 15 minutes)												
	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10	#11	#12	
Home chores	28.60	16.65	29.70	13.06	13.31	20.54	26.54	16.48	16.85	27.94	33.64	31.22	
Home leisure	29.05	19.47	20.58	15.92	11.15	31.30	38.36	16.29	15.71	41.67	20.85	61.11	
Night sleep	19.40	15.03	24.25	11.91	13.93	18.29	24.74	15.43	10.03	10.53	18.76	66.45	
Workplace	9.59	-	12.50	75.00	50.00	31.56	31.06	83.33	67.60	-	19.86	79.60	
Shopping & services	87.22	36.52	43.74	30.93	48.47	49.07	63.89	44.91	39.65	58.33	48.57	77.78	
School/college	-	85.00	-	75.00	-	-	-	57.60	-	-	-	12.50	
Organizational /Hobbies	39.32	43.23	51.04	45.32	35.60	55.12	52.74	48.62	36.35	50.00	58.61	61.11	
Entertainment	28.70	37.22	43.95	41.44	59.00	32.54	29.17	62.50	52.33	65.30	55.28	50.00	
Sports	58.89	49.40	77.78	58.33	79.17	-	88.89	75.00	30.67	75.90	42.86	76.80	
Activity type	**Estimation accuracy for setting II (48 bins, each of duration 30 minutes)**												
	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10	#11	#12	
Home chores	35.57	33.70	45.18	28.98	21.97	36.16	39.16	32.45	33.21	35.87	46.80	41.84	
Home leisure	35.83	36.13	32.65	29.36	29.52	32.94	41.29	35.57	27.82	52.08	42.65	73.96	
Night sleep	22.49	17.55	35.43	17.53	21.90	25.77	29.22	44.44	12.91	26.32	27.55	69.08	
Workplace	12.55	49.60	17.30	75.00	97.60	35.87	41.88	87.50	75.00	33.33	17.37	79.60	
Shopping & services	80.71	58.67	66.00	55.08	70.02	65.73	70.83	61.01	61.23	66.67	72.17	83.33	
School/college	-	68.70	76.60	-	-	76.80	-	72.30	-	-	-	12.50	
Organizational /Hobbies	49.79	53.37	55.56	48.07	68.27	65.76	76.02	69.77	53.05	83.33	74.69	64.29	
Entertainment	36.57	26.48	43.81	38.70	80.56	49.05	61.90	61.11	57.71	84.60	66.46	75.00	
Sports	72.78	40.48	68.33	47.92	80.00	68.75	66.67	48.67	55.74	75.60	60.54	89.60	
Activity type	**Estimation accuracy for setting III (24 bins, each of duration 60 minutes)**												
	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10	#11	#12	
Home chores	59.05	50.12	68.60	50.33	42.55	53.40	63.89	49.65	50.96	57.47	64.19	50.56	
Home leisure	68.64	56.03	64.91	54.43	62.38	54.64	62.27	51.95	58.07	56.25	54.33	77.27	
Night sleep	20.19	23.73	39.83	21.93	35.75	32.41	46.42	43.67	52.82	35.53	39.23	69.08	
Workplace	17.47	90.00	22.91	95.83	76.80	42.58	40.02	79.00	88.89	66.67	22.15	96.00	
Shopping & services	85.00	83.53	82.29	86.53	87.79	76.82	96.88	76.04	89.29	79.17	84.05	94.44	
School/college	-	89.60	75.60	97.60	-	73.60	88.30	76.80	-	-	94.00	25.00	
Organizational /Hobbies	65.19	63.76	62.04	65.08	77.46	84.48	93.00	85.88	67.96	92.86	93.00	75.00	
Entertainment	68.65	76.11	67.42	65.55	79.30	63.69	70.37	72.22	77.22	90.00	80.00	66.67	
Sports	89.52	64.29	68.70	58.33	80.56	80.00	90.48	75.83	80.19	87.60	71.31	87.50	
Activity type	**Estimation accuracy for setting IV (4 bins, each of duration 360 minutes)**												
	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10	#11	#12	
Home chores	99.15	100.0	100.0	97.75	95.60	99.26	96.15	100.0	100.0	100.0	100.0	100.0	
Home leisure	98.77	100.0	98.70	99.57	100.0	98.38	99.29	100.0	100.0	100.0	100.0	100.0	
Night sleep	89.65	100.0	87.90	96.47	91.42	92.53	68.38	100.0	100.0	100.0	100.0	100.0	
Workplace	76.11	100.0	91.85	100.0	100.0	93.40	90.71	100.0	100.0	100.0	100.0	100.0	
Shopping & services	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	
School/college	-	100.0	100.0	100.0	-	100.0	100.0	100.0	100.0	100.0	100.0	100.0	
Organizational /Hobbies	96.88	100.0	100.0	99.04	100.0	100.0	97.50	100.0	100.0	100.0	100.0	100.0	
Entertainment	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	
Sports	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	

Mean accuracy 42.86%

Mean accuracy 52.33%

Mean accuracy 67.32%

Mean accuracy 98.65%

'-' indicates that algorithm didn't predict activity duration due to missing the particular activity type in model's input.

For comparison, using the AdaBoost algorithm (Allahviranloo 2016), accuracy of prediction for the start time in the out-of-sample observations with durations of 30 and 180 minutes was obtained as 14.41% and 47.13% (compared to 36.28% and 60.1% in the RF model). Furthermore, the model accuracy for activity duration with durations of 60 and 180 minutes was obtained as 61.03% and 92.45% (compared to 67.32% and 98.65% in the RF model). The prediction results for start time in the ALBATROSS model (Arentze and Timmermans 2004) using the CHAID algorithm with duration of 180 minutes was 35.4% (compared to 60.1% in the RF model) and activity duration with duration of 360 minutes was 38.8% (compared to 98.8% in the RF model). Although the data conditions and other settings of the models are different, a proper comparison can be made on the level of the algorithm/method used where the CHAID and AdaBoost algorithms are used instead of the RF algorithm on the same dataset. A conceivable further step of this work is to evaluate the performance of the RF model in such a comparison.

As was discussed earlier, in total 70% of the dataset was used for training the model and 30% for testing model performance. In order to evaluate the performance of the heuristic rule-based algorithm, the estimation errors in minutes on a continuous scale and in percentage were computed to show the duration of misclassification, and are shown in Table 6.7. For every activity type in each cluster, the error is estimated by calculating the edit-distance between the observed activity pattern and projected temporal pattern in the test set. Results show that the highest misclassification error in each cluster is for those activities with a shorter duration in the traveler's daily activity patterns. For example, in extended work-day workers cluster, entertainment activity has the highest misclassified error by 32.74 minutes. Further studies to overcome this limitation associated with activity

types with shorter duration are recommended. The total error in each cluster was estimated based on the summation of all misclassification errors over 24-hours of projected traveler's activity. The highest error percentage was found for the students cluster, at 36.71%, followed by the evening worker cluster, at 25.50%. Compared to other clusters, students and evening worker clusters had the lowest sample size in our model. Empirical results therefore reveal that the RF model can predict response variables with more precision when trained with a larger dataset. The mean estimation error for all twelve clusters in our model is 18.38% in the 24-hour period.

Figure 6.7 and Figure 6.8 shows the scheduled temporal pattern of traveler activities at the aggregate level obtained from the rule-based algorithm for identified worker and non-worker clusters, respectively. Compared to the observed temporal patterns (Figure 6.1 and Figure 6.2), the algorithm could mostly re-assemble different activities of travelers in each cluster. However, in cases where the dominant activity (e.g. work) occupied a large portion of the traveler's daily activity pattern, the accuracy level for scheduling activities with smaller durations decreased. Improving the performance of the heuristic rule-based algorithm with more insertion and adjustment constraints for scheduling activities with shorter durations is recommended for future studies.

Table 6.7 Mean scheduling error for test dataset (duration of misclassification*)

Activity type	Mean estimation error (minutes)											
	Cluster 1	**Cluster 2**	**Cluster 3**	**Cluster 4**	**Cluster 5**	**Cluster 6**	**Cluster 7**	**Cluster 8**	**Cluster 9**	**Cluster 10**	**Cluster 11**	**Cluster 12**
Home chores	24.44	56.30	48.02	86.96	105.79	32.03	4.69	86.36	61.81	61.50	32.91	45.64
Home leisure	118.24	31.15	29.09	15.57	69.49	70.54	53.38	2.48	52.30	15.20	50.81	231.62
Night sleep	49.31	54.92	20.76	14.93	30.02	44.04	55.30	28.28	31.18	174.83	28.22	186.85
Workplace	18.08	2.09	164.05	10.43	0.36	58.99	134.91	3.38	1.21	101.59	60.35	8.47
Shopping & services	11.29	48.63	12.28	8.46	2.21	0.35	7.60	20.01	3.44	4.11	0.40	10.02
School/college	-	0.61	1.13	0.52	0.04	4.16	1.69	4.61	5.60	3.81	0.56	13.71
Organizational/hobbies	16.40	26.55	7.85	36.26	2.44	3.71	2.39	9.30	7.77	0.87	2.24	13.31
Entertainment	32.72	1.69	28.10	54.13	2.28	3.65	7.19	5.50	6.39	1.21	3.89	18.51
Sports	14.89	3.28	19.08	7.65	0.24	4.21	12.05	12.79	2.74	4.12	10.03	0.46
Total error in 24-h	285.38	225.22	330.37	234.91	212.86	221.69	279.20	172.71	172.43	367.23	189.43	528.60
Mean estimation error (%)	19.82	15.64	22.94	16.31	14.78	15.40	19.39	11.99	11.97	25.50	13.15	36.71

**Each cell with the length of 5 minutes*

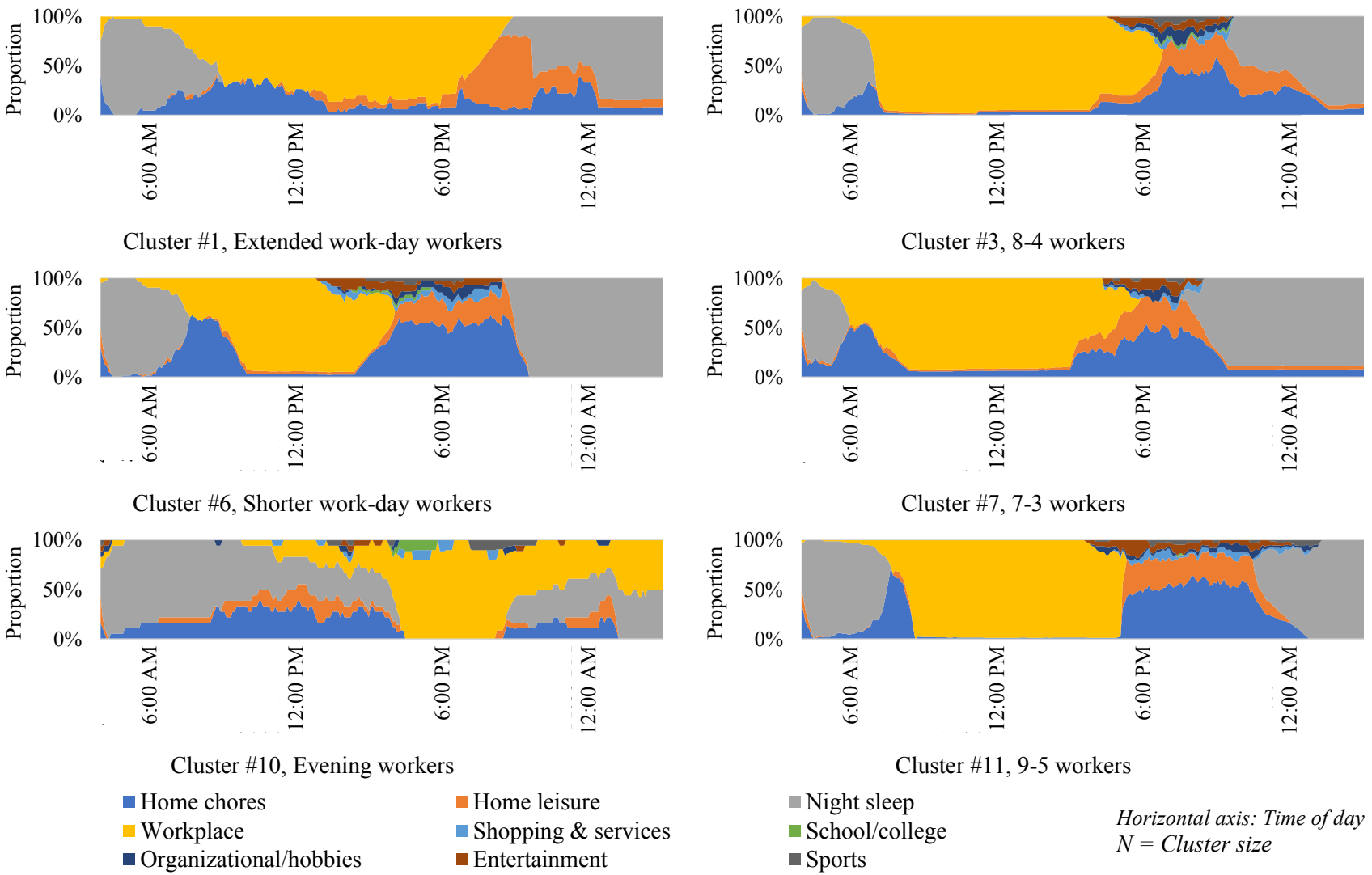


Figure 6.7 Scheduled temporal pattern of individual activities for six identified worker clusters

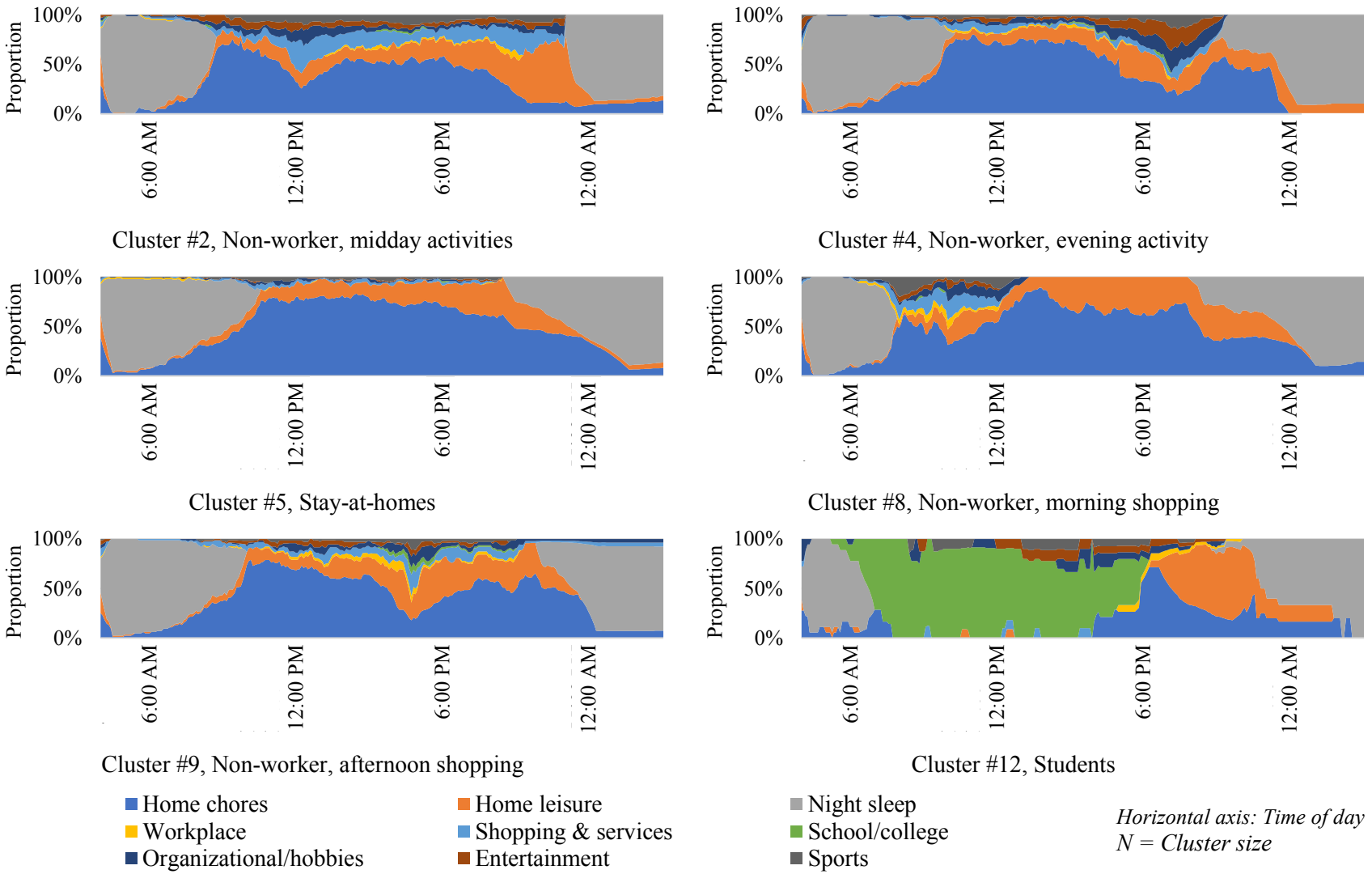


Figure 6.8 Scheduled temporal pattern of individual activities for six identified non-worker clusters

6.6 Conclusions

Complexities in activity-travel behavior of population groups in the study region vary according to their socio-demographic and socio-economic characteristics. For instance, homemakers and retirees have lower variances in time expenditure choices compared to worker and student groups. Accordingly, the best policy is to predict or model travel behavior for a representative set of model individuals, who represent homogeneous cohorts. The significant original contribution of this study is to develop a new modeling framework that is able to learn and predict temporal attributes of activities for use in activity-based travel demand models. We modeled the in-home and out-of-home activity temporal features of twelve clusters containing individuals with homogeneous activity patterns drawn from the large Halifax STAR travel diary survey. Activity start time and activity duration for every activity type were allocated to a set of bins. With respect to the pattern complexity of activity sequences and sample size of person-days in the clusters, eight different bin structures, varying in the time interval, were designed. The model was trained with 70% of the dataset and the remaining 30% was used for testing the model performance.

The modeling framework proposed in this study comprises numerous prediction decision trees developed using the Random Forest (RF) algorithm. Each tree plays as a weak learner in the algorithm and is able to make a prediction. The aggregation of these weak inputs provides a powerful ensemble learning model. A final prediction result is obtained from the majority votes obtained from each ensemble tree. Independent variables were selected from the socio-demographic characteristics of travelers and the corresponding start times or duration bin numbers for each activity type in the agenda. Bin numbers were defined as

the response variable in the RF model. Activity agendas and activity sequences of travelers are predicted using an advanced Random Forest (RF) algorithm. Consequently, for each activity type, activity start time and activity duration were generated from a uniform distribution within their interval time range of predicted bin numbers. In the next step, activities were inserted into the skeleton schedule using a heuristic decision rule-based algorithm, and a 24 hours schedule was constructed with respect to two-tier constraints: the importance level of the activity established from the cluster's representative pattern characteristics, and the duration of non-mandatory activities.

In this study, we demonstrated the utility of the RF machine learning technique for modeling the temporal attributes of activities, an application not previously employed in travel behavior analysis. The estimation accuracy of the proposed RF model was examined under different bin settings. The best estimation results for predicting activity start time were found for setting IV (8 bins, each of duration 180 minutes), 60.10%, followed by setting III (48 bins, each of duration 30 minutes), 36.28%. Similarly, the best model estimation results for predicting activity durations were found for setting $\widehat{\text{IV}}$ (4 bins, each of duration 360 minutes), 98.65%, followed by setting $\widehat{\text{III}}$ (24 bins, each of duration 60 minutes), 67.32%. Results show that the proposed model is able to assemble the traveler's schedule with an average 81.62% accuracy in the 24-hour period. By comparison of Figure 6.7 and Figure 6.8 (predicted temporal patterns) and Figure 6.1 and Figure 6.2 (observed temporal patterns) a visual impression of the goodness-of-fit can be obtained. The empirical results from comparison of prediction accuracy among twelve different clusters reveal that with increases in the time interval (decreasing number of bins), the RF

efficiency also increased. When the RF model is trained with a larger dataset, it is expected to predict response variables with more precision.

Numerous aspects of temporal information on activities, such as activity start time, activity duration, and activity end time, can be predicted for various population groups with various activity sequence patterns. Such precise information is essential for the scheduling phase of activity-based travel demand modeling. The proposed method improves on previous methods, and provides more accurate temporal information especially for individuals with high pattern complexity of activity sequences. For instance, predicted changes in travel over time compared to actual changes in travel over time can be accomplished in a short period of time using the proposed algorithm in this study.

Compared to other decision tree techniques such as boosted decision trees, the proposed RF model in this study is able to automatically handle missing values in the algorithm. Furthermore, variables do not need to be transformed, very few parameters need to be adjusted, and the algorithm does not overfit easily. However, compared to conventional hypothesis driven approaches used in transportation such as multinomial logit regression, which yields interpretable coefficients, most machine learning based approaches are designed as a black box. The model is trained to evaluate the labels of the data and its results are a set of support data-points and their respective weights. This is potentially problematic if the intention is to understand how elements of the activity-travel system interact, but it is not an issue if the purpose is simply accurate prediction. In addition, machine learning based approaches are very efficient in computational time with high degree of reproducibility. The methods employed in this study can also be adapted for

modeling other components of activity-based travel demand models, such as transport mode, and work and residential location choice models.

To build on this study and further demonstrate the potential of our proposed method, we are proposing several avenues of research. Firstly, it is possible to integrate the proposed model with a dynamic traffic assignment model and increase the model's ability for use in rescheduling activities. This will require updating the algorithm with new data on congested travel times. Secondly, and in line with growing worldwide interest in developing activity-based travel demand model at the household level, we aim to explicitly model intra-household interactions using the proposed modeling framework in this study. Thirdly, the large STAR survey data that has been used for building the RF model in this study includes business hours survey data. Therefore, one potential extension would be to include operations hours in the modeling process. The RF model presented in this study predicted activities with smaller durations with lower estimation accuracy compared to activities with larger durations. Therefore, improvement in the model structure for predicting and scheduling activities with smaller durations remains an area of future investigation.

In summary, the modeling framework presented in this study yields a straightforward and easy-to-implement tool for urban and transport modelers to predict and model activity temporal information of various population groups within a region. The results of this study are expected to be implemented within the activity-based travel demand model for Halifax, Nova Scotia, Scheduler for Activities, Locations, and Travel (SALT).

Chapter 7 Population Synthesis for Activity-Based Travel Demand Model Systems[5]

7.1 Introduction

In the past decade, interest has grown significantly to utilize disaggregate data in response to the need for complex systems modeling and policy analysis that require a higher level of disaggregate representation, spatially, temporally and socio-economically (Long and Shen 2013; Jordan, Birkin and Evans 2014). Numerous studies are available that use microsimulation platforms including activity generation and scheduling models such as ALBATROSS (2004) and TASHA (2008), and, network models such as MATSIM (2008) and Open Traffic (2014), and, integrated urban system models such as METROSIM (1994), TRESIS (2002) and ILUTE (2005).

One of the most important components of developing an integrated urban system model is the extensive amount of detailed information during the model building process. The initial step of most of the agent-based microsimulation models is to produce micro-data that includes details of the household and individual attributes (Wu, Birkin and Rees 2011). Typically, this step is known as population synthesis. Micro-data are usually a sample of thematically disaggregate data at individual or household level which could be spatially aggregated to a certain extent. Due to the lack of completeness and availability of the micro-data, population synthesis is an imperative step of the modeling process to generate synthetic virtual population data for the base year of the simulation. This study focuses on

[5] A condensed portion of this chapter has been published:
Hafezi, M. H., N. S. Daisy., L. Liu., and H. Millward. (2018). "Emissions analysis for the synthetic baseline population of a large Canadian university". Peer reviewed proceedings of the 97th Annual Meeting of Transportation Research Board (TRB), Washington, D.C., USA.

developing a methodology that is able to produce a set of synthesized population at both individual and household levels. It is expected that the synthesized population will be for microsimulation travel behavior.

The aim of population synthesis is to use the public random sample data (such as Public-Use Microdata Samples of the U.S. and Anonymized Records of the U.K.) and expand it to mirror known aggregate information (such as Summary Files of the U.S. and the Small Area Statistics file of the U.K.). Essentially, population synthesis requires two types of data: disaggregate sample data and spatial unit level aggregate totals. The first set of input data, called the seed data, is generated by the joint distribution of household and individual micro-sample data. The second set of input data is called the control tables or the marginal tables, for distributions of single variables in smaller spatial units (Ye et al. 2009). In this study, we present a new functional form for a fitness based synthesis algorithm (Ma 2011) that resulted in improved computational efficiency and model accuracy. The performance of the algorithm for each proposed model is validated using error-percentages and goodness-of-fit. In addition, to measure dispersion, results for 5% and 10% samples are compared.

The remainder of the study is structured as follows; first, the study provides a review of relevant past research concerning population synthesis and techniques. Following the literature review, a discussion of the data used in the generation of the synthetic population for Halifax is presented. The proposed population synthesizer approach used is described in the next section, followed by a discussion of model results. The study concludes by providing a summary of contributions and future research directions.

7.2 Literature Review

Synthesizing population at different spatial unit levels is a long-standing problem in numerous disciplines, including: transportation, industrial engineering, applied and survey statistics (Estevao and Särndal 2006). Several statistical techniques such as survey sampling, weighting method and maximum entropy modeling are developed to synthesize population at different spatial unit levels for microsimulation modeling (Lemaître and Dufour 1987; Deville, Sarndal and Sautory 1993; Malouf 2002; Estevao and Särndal 2006). An overview of literature on existing population synthesis approaches is presented in this section. Maximum Entropy (ME) models are constructed based on a set of probability distributions and leads to a sort of statistical conjecture. They use different types of algorithms such as conjugate gradient, iterative scaling, quasi-newton and gradient ascent for estimating the parameters (Malouf 2002; Bar-Gera et al. 2009). Lee and Fu (2011) proposed a cross-entropy optimization model utilizing a quasi-newton algorithm to synthesize a desired amount of completely identified individual activity patterns. Results of the model show that the proposed cross-entropy optimization can generate a realistic synthetic population in different geographic areas.

Nagle et al. (2013) introduced the Penalized Maximum Entropy Dasymetric model (P-MEDM) that is able to produce spatial micro-data at the household level. Christakos in 2000 studied the Bayesian Maximum Entropy (BME) method for examining spatiotemporal distributions of natural variables. In another study, Kyriakidis (2004) developed a geostatistical framework for the spatial estimation of point values from areal data. The suggested framework predicted each point with high reliability. Moreover, several current methods for area-to-point interpolation can be embedded within the

suggested framework by Kyriakidis. Wu and Murray (2005) developed the Co-kriging method to interpolate residential population density at the Thematic Mapper (TM) level (30 by 30 meters) of an urban region. They modeled the spatial interrelationship and cross-correlation of population using census count data. In comparison with other interpolation methods, their proposed model provides estimation variance that allows the assessment of estimated population at TM level without the need of aggregation to the census reporting zone.

Numerous methods have also been developed in the transportation literature to synthesize population at different spatial unit levels. Common approaches include fitness-based methods (such as iterative proportional fitting and fitness based synthesis) and combinatorial optimization. One of the main differences in the population synthesizer method is the capability to simultaneously control both the household and individual characteristics (Hermes and Poulsen 2012). With the exception of the traditional Iterative Proportional Fitting (IPF), other methods namely Combinatorial Optimization (CO), Iterative Proportional Updating (IPU) and Fitness-Based Synthesis (FBS) are capable of controlling both the household and individual characteristics in the process. IPF was one of the earliest population synthesis methods developed by Beckman et al. (1996) that used an iterative fitting process. IPF utilized sample and census aggregate data of the U.S. The standard IPF method comprises two phases: fitting un-adjusted cell data (i.e. seed data), and generating the synthesized households. The seed data should be adjusted to a known margin for both the horizontal and vertical dimensions of the table (i.e. control tables). This approach uses a Monte Carlo Simulation (MCS) method, generating the synthesized households by drawing household and individual records from the seed data. Some

limitations of the original IPF method are: discrete control for individual and household level attributes, zero cell problem, high-dimensional (memory) problems and rounding cell values. These limitations cause a reduction in the accuracy and validity of the synthesized population. Subsequently, several studies have improved and expanded the original IPF method, such as Pritchard and Miller (2009); Guo and Bhat (2007); Ye et al. (2009); Auld and Mohammadian (2009), and Arentze and Timmermans (2004). The majority of the aforementioned research has focused on improving the efficiency of the original IPF method, and on issues such as the limitation of memory and joint distributions of the control variables at both household and individual levels.

Pritchard and Miller (2009) explored the synthesis of large-scale attributes per agent by using a list-based method, in which household and individual level attributes were fitted simultaneously. The use of sparse matrices increased the capability of controls and categories resulting in decreased memory use and computational time. Guo and Bhat (2007) studied the 'zero cell' issue and controlling individual level attributes. They defined a certain tolerance for individual attribute levels in the initial steps involving the seed data from which household data ensured that individual level restrictions were not disrupted. This method also allows manipulation of data from different data set sources, which solves another limitation of the original IPF method. Arentze and Timmermans (2004) used a two-step model on the relevant attributes between the household and individual levels. The individual level attributes were aggregated to the household level, then the results of the marginal table from the first phase were used to synthesize the population. Additionally, this study addressed the differences in populations relating to locational characteristics using sample segmentation technique.

Muller et al. (2010) studied the issue of synthetic reconstruction, that reweights the household and individual attribute levels to allow the algorithm to efficiently permit for simultaneous controls at numerous levels. This new method is called Hierarchical Iterative Proportional Fitting (HIPF). An entropy-optimizing fitting step was added to simulate the individual attribute levels into the household level, based on situational specific constraints. The synthetic population is drawn from the individual/household group data (Muller and Axhausen 2011). Generally, this heuristic method is similar to the combinatorial optimization (CO) method. Lovelace et al. (2015) evaluated the performance of IPF for spatial microsimulation with the objective of generating spatial micro-data. Barthelemy et al. (2015) introduced the Multidimensional Iterative Proportional Fitting Procedure (MIPFP). MIPFP addressed the zero- cell and non- integer weight problems of the original IPF method. Furthermore, several alternative estimation methods such as minimum chi-squared (CHI2), maximum likelihood (ML) and weighted least squares (WLSQ) are included in the MIPFP with the goal of updating the N-dimensional array with respect to certain control tables.

The combinatorial optimization (CO) method is structured differently, and addresses some of the restrictions of the previous approaches. Seed data and control tables similar to the previous method are also required in the CO method. The CO method utilizes the integer reweighting technique. The CO method uses optimization techniques including genetic algorithm, hill-climbing, or simulated annealing to optimize the weighting process. The CO method minimizes the difference between the synthesized population and marginal tables. Additionally, variance values and memory usage in the CO method are lower than in the IPF method. However, the optimization process of the CO method has more

attributes, which is computationally time consuming. An exhaustive review of the combinatorial optimization (CO) method can be found in Voas and Williamson (2000, 2001), and Huang and Williamson (2001).

Similar to the CO method, Iterative Population Updating (IPU) is another population synthesis method that is able to control both household and individual level attributes. IPU, introduced by Ye et al. (2009) matches household and individual level attribute distributions with high precision. The IPU method comprises three main steps. The first step is to obtain household and person level constraints by selecting 5% of individual and household level attributes from PUMS seed data. There is a pre-treatment procedure in the first step which corresponds to correcting the zero-cell problem in the seed data and the zero-marginal problem in the marginal tables. The second step is to estimate weights of the individual and household level joint distributions in the way that both individual and household distributions can be closely matched. The last step consists of drawing household data from the procedure in the prior phase in order to generate the synthetic population for the region. Generally speaking, matching household and individual level attributes using the IPU method is better than the IPF method. However, there are still some issues of discrepancy observed in matching seed data to marginal tables (Ye et al. 2009; Lim and Gargett 2013). Table 7.1 shows a summary of several population synthesizer methods with selective advantages and limitations.

Most population synthesizer methods use a joint multi-way distribution. Ma (2011) presented a Fitness-Based Synthesis (FBS) algorithm that sequentially matches multilevel marginal tables. The Fitness measure for each sample household is calculated during the process to verify the match for both individual and household level distributions. In each

iteration, the highest fitness value will be chosen and the resulting households including all of their individual members will be added to the synthetic population list. The termination criterion of the iterative method is an absence of positive fitness values. The FBS algorithm performs well in terms of simultaneously controlling both individual and household attributes of interest. Several studies suggest that despite improvements in population synthesizer methods, there are still unresolved issues, such as high-dimensionality problems (i.e. association structure of the households and individuals), memory use and processing time, discrepancies in matching the seed level data to marginal data, and generation of similar fitness values in the case of few control tables.

A common challenge among most of the population synthesizer methods is how to evaluate the validity of the synthesized population. Accuracy and precision of the synthesis often are validated by comparing the true and synthetic populations (Oketch and Carrick 2005). However, difficulties in obtaining the true population creates a critical challenge in the validation of any population synthesis procedure (Edwards and Clarke 2009). Apart from Anderson et al. (2014), who had access to spatially-referenced disaggregated micro-data in Switzerland, other researchers proposed different approaches for the validation of the synthetic population. The validation procedures are still evolving in the existing literature. Huang and Williamson (2001) used the total absolute error, the Z score, and the standardized absolute error to validate the synthetic population. Schroeder (2007) attempted to validate spatial allocation of micro-data by utilizing the finest resolution US census data. The proposed method involves hypothesized absolute bounds on interpolation inaccuracy that enable calculation of upper and lower prediction bounds for the population.

Table 7.1 Advantages and limitations of four population synthesizer methods

Population synthesizer method	**Advantages**	**Limitations**
Traditional Iterative Proportional Fitting (IPF)	- Combination of probability table - Zone-by-zone versus multi-zone - The format of the seed is remembered	- High-dimensional problems - Zero cell problems - Rounding of the cell values in joint distribution - Variance value is high - Association structure for household and individuals - The method is unable to control for individual attributes. - The method ignores differences in household type among households grouped within a cell - Only observed and simulated control variables are matched - Long running time
Combinatorial Optimization (CO)	- Variance value is small - Continuous/discrete characteristics - Proficient memory - Changing the number of layers is insignificant	- PUMA level data is required to be applied to the procedure due to correlation structure within the zonal population
Iterative Proportional Updating (IPU)	- Controls for both household and individual attributes simultaneously - Matches seed level data and marginal control characteristics on several analysis levels - Less running time - Extra corrections to match individual and household level data are not required	- Discrepancies in matching of seed level data to marginal data - Rounding of values
Fitness-Based Synthesis (FBS)	- Generates a list of households to match numerous multilevel controls - Determining a joint multi-way distribution is not required	- Creates the same fitness value iteration in the case of few-control-tables

Ruther et al. (2013) proposed a validation method that examines spatial allocation of census micro-data to census tracts. The allocation is determined by a series of known aggregate census tract population distributions. The evaluation of the model is performed with an assessment of the estimation error utilizing maximum entropy imputation and a spatial allocation model. The study showed that the addition of constraining variables can

improve model fit for both constraining variables and correlated variables to the constraining variables. Furthermore, Ballas et al. (2007) used a regression analysis and an R-square measure to validate synthetic population. Reflecting on all those challenges in the existing literature, in this study we propose a population synthesizer approach that can address most of the above-mentioned limitations. Population is synthesized for individuals and households at the regional and dissemination area levels using the proposed approach. To illustrate the accuracy of the synthesized population using different levels of control tables and spatial unit levels, the proposed algorithm is examined by three sub models: using only the household level control tables (HL model); using both individual and household level control tables (HPL model); and, with weights added to both individual and household level control tables (WHPL model). Note that WHPL is a new weighted model proposed in this study to improve the efficiency of the synthesizer.

7.3 Data Used in the Generating a Synthetic Population

The generation of a synthetic population requires two major datasets: 1) micro sample data (i.e. a sample of thematically disaggregate data at individual and household level) and 2) aggregate totals of individuals and households at the spatial level of interest for the study area. The first data source employed in this study is the 2006 Hierarchical Public Use Microdata File (PUMF) obtained from the Statistics Canada. The second data source is the 2006 Canadian Census, which is used to synthesize population for the base year of the proposed integrated urban model of Halifax. At first, this study aimed to synthesize population for the Halifax region directly. However, it appears that the 2006 PUMF hierarchical file (i.e. both individual and household information) is available for large Census Metropolitan Areas (CMAs) only (i.e. Montreal, Toronto, Calgary, Edmonton, and

Vancouver). Information for the Halifax CMA is only available in the 2006 PUMF individual file. The individual PUMF is not adequate to generate household-level information. Therefore, this study uses the Atlantic Canada region (which comprises New Brunswick, Newfoundland and Labrador, Nova Scotia, and Prince Edward Island) for this empirical application. Control tables for synthesizing population at the regional level (RL) are extracted from four regions of Atlantic Canada and synthesized population for Atlantic Canada is used as seed data.

The spatial unit considered for synthesizing the Halifax population is the Dissemination Area (DA). Control tables for Halifax DAs are extracted from Census tabulations. Initially, synthesizing the population at the regional level reproduces the appropriate amounts of different types of household, which are represented in both the control table and seed data from the corresponding DA. In other words, the 5% PUMF data might not contain all types of households that are present in the DAs for the specific census tract. The process this study took essentially waives a requirement for zero cell treatment.

Reviewing the PUMF data source revealed that the individual, household, and family levels were not linked together until the 2006 census. Essentially, there was a separate PUMF for the individual file, household file, and families file. For this reason, researchers used several variables to obtain the household composition and linked separate PUMFs together in synthesizing the population (Beckman, Baggerly and McKay 1996; Pritchard and Miller 2009). However, since 2006, fortunately PUMF data for the aforementioned levels are linked together (Statistic Canada 2011). Each individual's record includes their household and family identification. Hence, a complete set of information for each individual in each household offers an advantage to explore essential details of the

composition of households during the population synthesis process. Furthermore, the micro-sample of Nova Scotia is available in the 2011 PUMF data source. However, reviewing the 2011 census at dissemination area level indicated several missing information that are required for synthesizing population. This issue is known as zero cell problems in population synthesis studies. Therefore, prior to use the 2011 data source in population synthesis, it is highly recommended to solve zero cell issue first in order to obtain valid synthetic population results. Moreover, due to lack of data availability particularly for Nova Scotia, the most appropriate geographical level for synthesizing population in Halifax and Nova Scotia are Census Tract (CT) and Dissemination Area (DA).

7.3.1 Attributes Considered for Synthesizing Population

In this study, six household attributes and five individual attributes are considered for population synthesis. The household attributes are: household size, type of household, tenure, household income, structural type of dwelling, and labor force activity. Household size is categorized as 1, 2, 3, 4, 5 and 6+ residents. The type of household is categorized into four groups: married, single parent, parents with children, and single occupant. Tenure is classified by owned and rented properties. Household income is divided as follows: under $19,999; between $ 20,000 and $ 24,999; between $ 25,000 and $ 29,999; between $ 30,000 and $ 34,999; between $ 35,000 and $ 39,999; between $ 40,000 and $ 44,999; between $ 45,000 and $ 49,999; between $ 50,000 and $ 59,999; and $ 60,000+. Dwelling type has eight classifications: single-detached house, semi-detached or double house, row house, apartment/flat in a duplex, apartment in a building that has five or more stories, apartment in a building that has fewer than five stories, other single-attached house and

mobile home, and other movable dwelling. Labor force activity is classified as employed and un-employed.

Variables associated with individual attributes are: age, education level, ethnicity, legal marital status and gender. Age was defined in the following 11 categories: ≤19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-64, 65-74 and 75+. Education level is defined as: high school graduation diploma or equivalent certificate, apprenticeship or trades certificate or diploma, college, CEGEP or other non-university certificate or diploma, university certificate or diploma below bachelor level, bachelor's degree, university certificate or diploma above bachelor level, degree in medicine, dentistry, veterinary medicine or optometry, master's degree, and, earned doctorate degree. Ethnicity has seven classifications: British Isles origins, French origins, Aboriginal origins, Canadian, European origins, Asian origins and other origins. Legal marital status is defined by divorced; legally married; single; and widowed. Gender is identified by either female or male.

7.3.2 Data Preparation for Synthetic Population

The data preparation for a synthetic population involved multiple stages. First, a database with all explanatory variables was created. These data were derived from the hierarchical and individual PUMF and the Canadian Census. The second step is cleaning the data set of any missing values for validity, consistency, and uniformity. An example of 10 households for the seed data for variables of interest (under control tables) is demonstrated in Table 7.2.

Table 7.2 Sample seed data used in the empirical application

HH_ID*	PP_ID	AGEG	MARS	GEN	HDGR	LFAC	TOTIN	DTYP	TENU	ETHDE	CFST
12	121101	7	1	1	1	1	10	6	2	1	6
28	281101	5	2	1	3	1	6	1	1	4	1
28	281102	4	2	2	1	1	10	1	1	4	1
32	321101	10	2	1	5	2	9	1	1	7	1
32	321102	10	2	2	1	2	13	1	1	7	1
37	371101	2	4	1	1	1	7	6	2	7	3
48	481101	11	4	1	4	2	9	6	2	7	6
72	721101	10	2	1	2	2	4	1	1	4	1
72	721102	10	2	2	3	1	10	1	1	7	1
73	731101	1	4	1	1	2	3	1	1	4	4
73	731102	6	2	1	4	1	13	1	1	4	1
73	731103	7	2	2	2	1	14	1	1	4	1
97	971101	6	4	2	3	1	8	1	2	1	6
100	1001101	4	4	1	6	1	12	1	1	7	3
100	1001102	9	2	2	4	2	13	1	1	7	1
100	1001203	9	2	1	3	2	15	1	1	7	1
107	1071101	7	1	1	3	1	11	1	1	4	2
107	1071102	8	1	2	3	1	11	1	1	4	2

**HH_ID = household identifier, PP_ID = person identifier, AGEG = age groups, MARS = legal marital status, GEN = gender, HDGR = highest certificate, diploma or degree, LFAC = labor force activity, TOTIN = total income of individual, DTYP = structural type of dwelling, TENU = tenure, ETHDE = ethnicity, and, CFST = household type, detailed census family status and household living arrangements*

The next step is to create the count table in which all values are initialized to zero as a default. The count tables are structured similarly to the control table. During the computational procedure, count tables are updated. Cross tabulations are derived from the corresponding variables of interest. Table 7.3 shows an example of a control table (age and gender) used in the empirical application.

Table 7.3 Sample control table used in the empirical application

Gender \ Age group*	1	2	3	4	5	6	7	8	9	10	11
Male	88	19	19	19	34	29	24	34	39	24	19
Female	93	15	15	26	21	46	31	31	46	10	15

**1 = ≤19, 2 = 20-24, 3 = 25-29, 4 = 30-34, 5 = 35-39, 6 = 40-44, 7 = 45-49, 8 = 50-54, 9 = 55-64, 10 = 65-74 and, 11 = 75+*

The last step is to create the MH-table. The MH-table includes values of 0 and 1. If a household does not exhibit the cell classification value in the seed data then the value will be zero. If cell classification is evident there is an initial value of one. This study synthesizes 5%, 10% and 100% samples of the population for the regional and dissemination area levels, respectively.

7.4 Methodology

In this study, we present a new functional form for a fitness based synthesis algorithm (Ma 2011) that resulted in improved computational efficiency and model accuracy. According to the calculated fitness value, a set of households in each iteration are selected and added to the count table. For each control table in the algorithm, there is a unique count table that is given an initial value of zero. As the procedure progresses, households are added or removed to this table. Adding or removing households from the count tables is decided based on the fitness value. The algorithm procedure works in an iterative fashion that continues until the count table reproduces the control tables as closely as possible. However, the values may not be an exact match for all count tables. Therefore, the iterative process is terminated when there is an absence of positive fitness values.

In this study, two types of the fitness measures for determination of adding or removing the selected households are considered. Briefly, only those households that have a positive fitness value are selected as potential candidates to be added into the synthesized list. The first fitness measures (type 1) corresponds to the error if the selected household is added to the count table in the current iteration. The second fitness measures (type 2) corresponds with the error if the selected household is eliminated from the count table in the current

iteration. In each iteration, only one type of fitness value can be positive; both fitness values cannot be positive at the same time. The average distribution of the control table in the seed data is used as a new weight to calculate the fitness values. These measures ensure that the performance and precision of the synthetic population created has been enhanced. The optimized type one and type two fitness values are given by the following:

$$F_{1Q}^{xy} = n_q * [\left(G_{qw}^{y-1}\right)^2 - \left(G_{qw}^{y-1} - MH_{qw}^{x}\right)^2] \quad (1)$$

$$F_{2Q}^{xy} = n_q * [\left(G_{qw}^{y-1}\right)^2 - \left(G_{qw}^{y-1} + MH_{qw}^{x}\right)^2] \quad (2)$$

$$G_{qw}^{y-1} = H_{qw} - AH_{qw}^{y-1} \quad (3)$$

$$m^q \in \{F_{1Q}^{xy} > 0\} \quad (4)$$

$$\neg \mathrm{A}\ m^q = (\,) \quad (5)$$

$$m^q \in \{F_{2Q}^{xy} > 0\} \quad (6)$$

$$\neg \mathrm{A}\ m^q = (\,) \quad (7)$$

$$rm_y^1 \in \{m^1(:,:,y) >= 0\} \quad (8)$$

$$rm_y^2 \in \{m^2(:,:,y) >= 0\}$$

$$rm_y^3 \in \{m^3(:,:,y) >= 0\}$$

$$\vdots$$

$$rm_y^{qw} \in \{m^{qw}(:,:,y) >= 0\}$$

$$mm_y \in \{rm_y^1 \cap rm_y^2 \cap rm_y^3 \cap \cdots \cap rm_y^{qw}\} \quad (9)$$

$$\neg A\ mm_y = ( ) \quad (10)$$

$$mm_y \in \{rm_y^1 \cap rm_y^2 \cap rm_y^3 \cap \cdots \cap rm_y^{qw-1}\} \quad (11)$$

$$\neg A\ mm_y = ( ) \quad (12)$$

Where F_{1Q}^{xy} is the fitness value type 1 for control table Q, F_{2Q}^{xy} is the fitness value type 2 for control table Q, x is the selected household, y is the iteration number, q is the index representing for the both count and control tables, Q is the total number of both count/control tables, w is the index representing the various cells in the count table, W is the index representing the various cells in the control table, H_{qw} represents the amount of cell w in control table q, n_q represents the average distribution of control table q in the seed data, AH_{qw}^{y-1} represents the value of cell w in the count table q, G_{qw}^{y-1} is the difference value between control and count tables for cell w in control table q, MH_{qw}^{x} is the contribution of the x^{th} household in the seed data to the w^{th} cell in control table q, m^q is the selected household type 1 or 2 according to the fitness value, rm_y^{qw} is the selected household for the cell w in the count table q in the iteration y, mm_y is a set of selected households for adding into the count tables and synthesized population list. Moreover, $G_{qw}^{y-1} - MH_{qw}^{x}$ and $G_{qw}^{y-1} + MH_{qw}^{x}$ are the number of households which need to be added or removed, respectively for the cell w in the count table q in the iteration y.

Fitness value for each household are calculated according to formula 1 and 2. Positive type 1 or 2 fitness values are examined as the candidates for addition or removal from the list

of updated synthetic population. The traditional FBS algorithm utilized a random selection of one of the households from the candidate sets of households. In contrast, this study proposes a new selection approach that improves computational time and increases algorithm efficiency. Instead of one selection per iteration, the proposed algorithm selects all of the positive fitness values for each of the variables, with some conditions. First, all of the positive fitness values for cell w in count table q in iteration y are selected. Then, the same household that is repeated in all of the selected sets, will be considered as a set of households for adding into the count tables and synthesized population list. Note here that with increase in the number of iterations, subsequently the number of selected households is decreased. This practice is defended by reason that the algorithm is trying to expand the seed data according to the defined control tables, and the sequence of household selection is not important in the procedure. Hence, adding all of the potentially repeated households at once is reasonable and logical. For example, if household id number z is selected u times in y iterations, the new selection method allows household id number z to be added u times in one iteration to the final synthesized population list and count table. Following the selection process, all values of all dimensions in the count tables are updated. Again, all of the prior steps are repeated in the next iteration. Algorithm is terminated when all of the fitness values are negative.

The population synthesis approach described in this study is comparable to work done by Ma (2011) in that it produces a list of potential households to match numerous multilevel marginal tables. The fitness-based synthesis (FBS) algorithm solves several problems of previous population synthesizer approaches, including zero cell problems and computational resources (memory). Furthermore, the Monte Carlo Simulation (MCS) step

in the household selection procedure is skipped in the FBS algorithm through generation of integer fitness values. The population synthesis approach proposed in this study has some notable advantages in comparison with the FBS algorithm. The proposed approach can address most of the limitations of previous population synthesizer methods mentioned in the literature, and improves the efficiency of the algorithm, both in computational time and distribution of seed data presented in the final synthesized population list. One of the strength aspects of the proposed approach is in the household selection step. The new selection procedure of the population synthesizer approach creates a synthesis of the complete set of potential individual and households members in a sole iteration, and thereby greatly improves the efficiency of the algorithm.

7.5 Discussion of Results

The performance of the proposed population synthesis approach is examined by three sub models. First, only using the household level control tables (HL model) and no control for individual level attributes. Second, using both individual and household level control tables (HPL model), and third, with weights added to both individual and household level control tables (WHPL model). The n_q value in fitness calculation formula is considered as 1 for the HL and HPL models, based on the distribution of household attributes in the seed data from the corresponding control table in the WHPL model. Then, n_q is calculated by the average of all household attributes in one control table category. These measures ensure that the performance and precision of the synthetic population produced through the WHPL model has been improved in comparison with the HL and HPL models. Numerous techniques are available for calculating the precision level of synthetic populations (Ballas et al. 2007; Schroeder 2007; Edwards and Clarke 2009; Ruther et al.

2013; Anderson et al. 2014). In this study, the error-percentage and goodness-of-fit are computed to measure the accuracy of synthesized populations. Error-percentage measures are calculated as follows:

$$EP_q = \frac{\sum_{q=1}^{Q} |H_{qw} - AH_{qw}|}{\sum_{q=1}^{Q} H_{qw}} \quad (13)$$

Where:

EP_q represents the error-percentage for control table q,

H_{qw} represents the amount of cell w in control table q,

AH_{qw} represents the amount of cell w in the count table q.

7.5.1 Synthetic Baseline Population at the Regional Level

For each of the control tables, in each of the three models, error percentages are summarized in Table 7.4. Error percentages represent the differences between target population (control table from census data) and synthesized population.

Table 7.4 Error percentages of three models (regional level)

Explanatory variables	HL model	HPL model	WHPL model
Individual level			
Age	6.4%	0.9%	0.3%
Education level	8.3%	1.4%	1.1%
Ethnicity	18.9%	0.8%	0.1%
Legal marital status	7.9%	0.2%	0.1%
Gender	8.8%	0.1%	0.1%
Household level			
Household size	0.9%	1.1%	1.1%
Type of household	0.5%	0.8%	0.8%
Tenure	0.9%	1.9%	1.4%
Household income	0.4%	0.0%	0.9%
Dwelling type	0.9%	1.6%	1.3%
Labor force activity	2.4%	1.2%	1.1%

Two major differences exist in the error percentages of the three sub models in Table 7.4. In the case of individual level control tables, error percentages in the HPL and WHPL models are significantly smaller than that of the HL model. This difference demonstrates that use of both individual and household level control tables increases the precision of the resulting synthetic population. Possibly, this is due to the improved precision of the fitness value when additional control tables are included in each iteration. Secondly, in the case of household level control tables, error percentages in the HPL and WHPL models are slightly larger than those of the HL model. This is likely due to a growth in the total number of marginal tables in the model. When the number of control tables is increased the capability to replicate each count table is decreased.

Additionally, goodness-of-fit comparison is performed to compare the multivariate joint distributions. The synthesized population has a good fit with the control table when a trend-line slope is close to 1 and there is a high R-square value. Figure 7.1, Figure 7.2 and Figure 7.3 show a comparison of the goodness-of-fit of individual (age x gender) and

household (household size x tenure, household type x tenure) level attributes for the regional model, with three sub models consisting of the HL, HPL and WHPL methods (for three samples of results).

Figure 7.1 demonstrates that a synthesized population produced by the HL method does not provide a close match to the real population (control table). Figure 7.2 and Figure 7.3 demonstrate that the goodness-of-fit of the synthesized population produced by the HPL and WHPL methods is better than the synthesized population produced by the HL method. All three sub models for the household level attributes produced synthesized populations with a close match to the actual population (control table). However, the results of the comparison between different models show that the goodness-of-fit for populations synthesized by fewer control tables (Figure 7.1) produce a slightly better result than those produced using more control tables (Figure 7.2 and Figure 7.3) in terms of household level attributes. Model overfitting could be another reason for this conclusion. This result is consistent with the result of the percentage error calculation. The goodness-of-fit outcomes also demonstrate that the attributes with fewer categorizations display better fitting outcomes.

Additionally, through comparison of the results of the synthesized population between the HPL and WHPL models, the WHPL model's repetition of larger household selections has declined. Smaller households in the WHPL model are more prone to accidental selection and addition to the population synthesis list in comparison to those of the HPL model. Presumably, this might be due to using additional weight in the calculation of the fitness value in the WHPL model. Smaller household size has bigger weight in comparison with

larger household size. This prevents the selection of larger households in the early stage of the iteration so that the synthesized population better matches the target population.

In general, the proposed population synthesizer approach better preserves attributes in the control tables in comparison with non-controlled attributes. Future research should evaluate the performance of the proposed approach for attributes of non-controlled tables. Generally, the total number of iterations in the WHPL model has marginally increased. Synthesizing population at both individual and household level attributes is essential for microsimulation that will incorporate intra-household interactions. For instance, joint activities such as drop-off/pick-up or joint discretionary activities among all household members can be simulated within agent-based activity travel models.

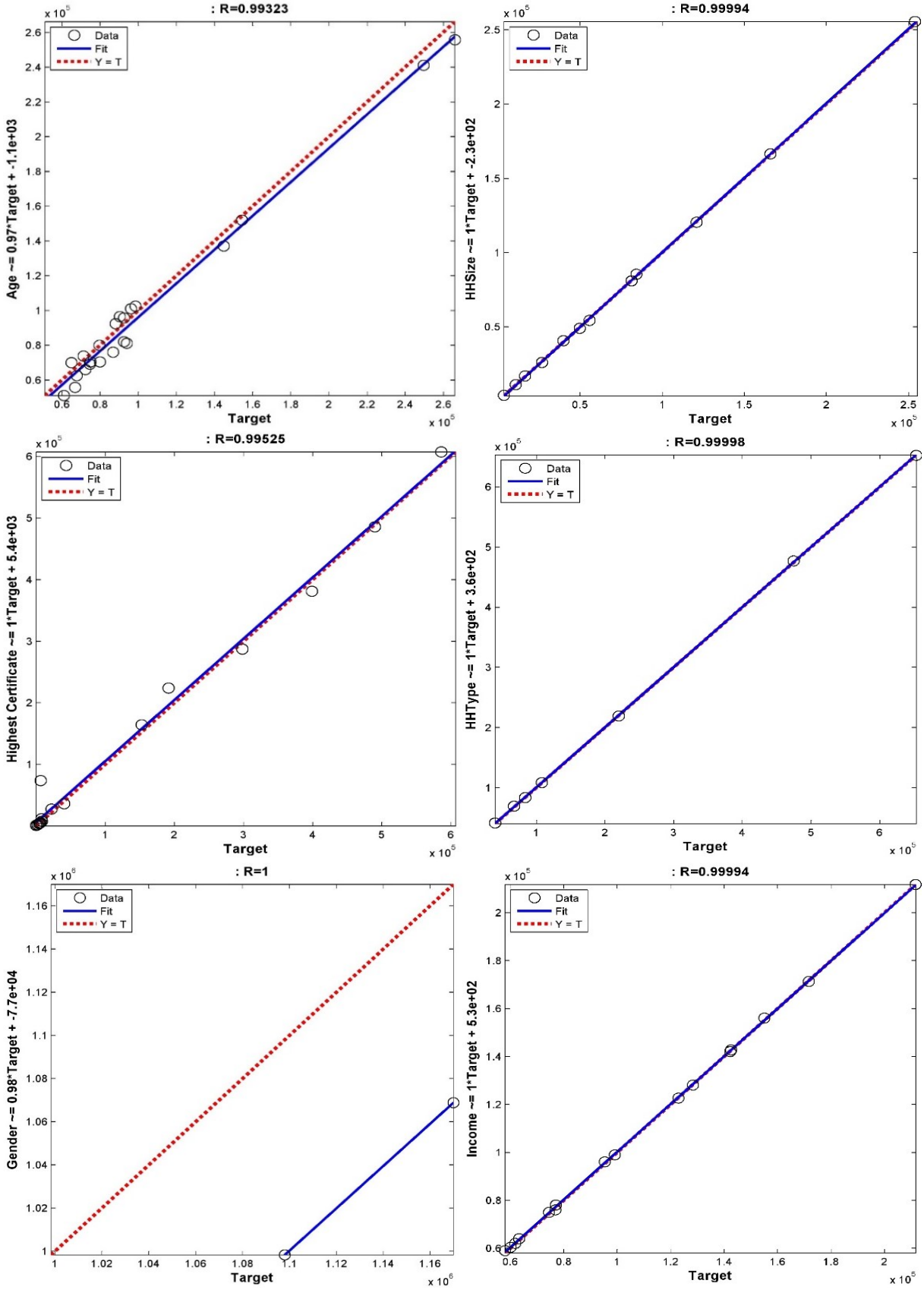


Figure 7.1 HL model: individual and household level's goodness-of-fit comparison

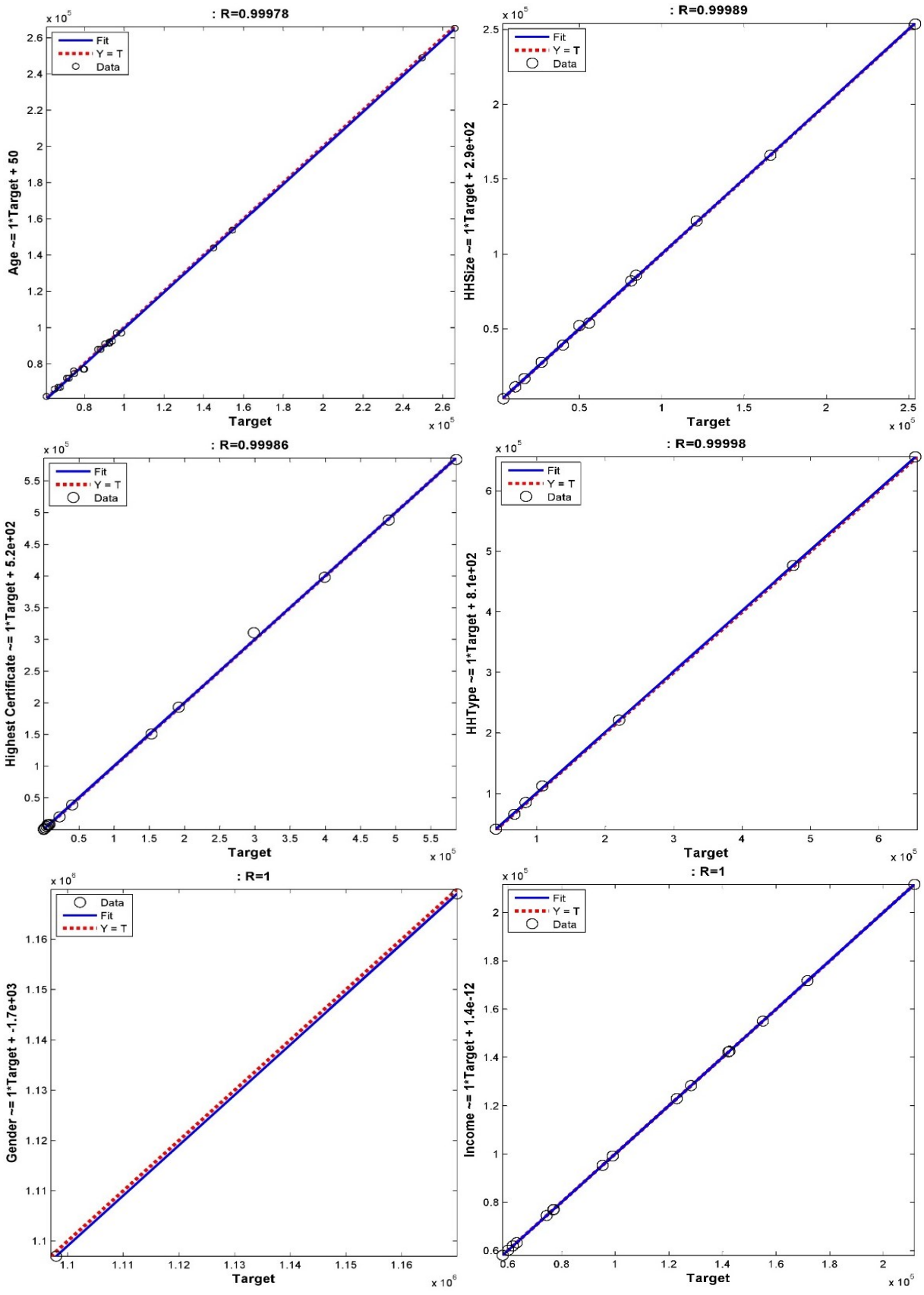


Figure 7.2 HPL model: individual and household level's goodness-of-fit comparison

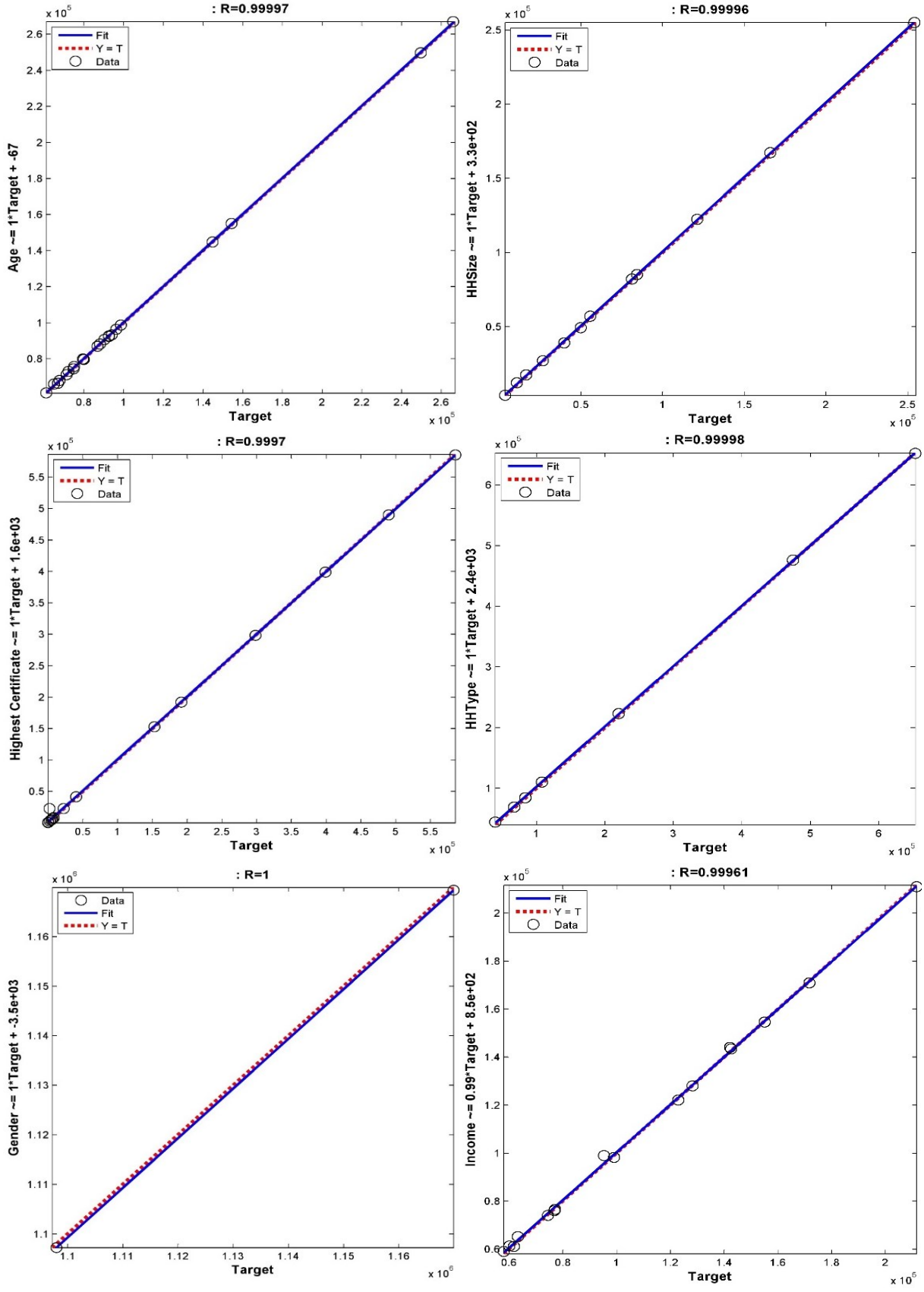
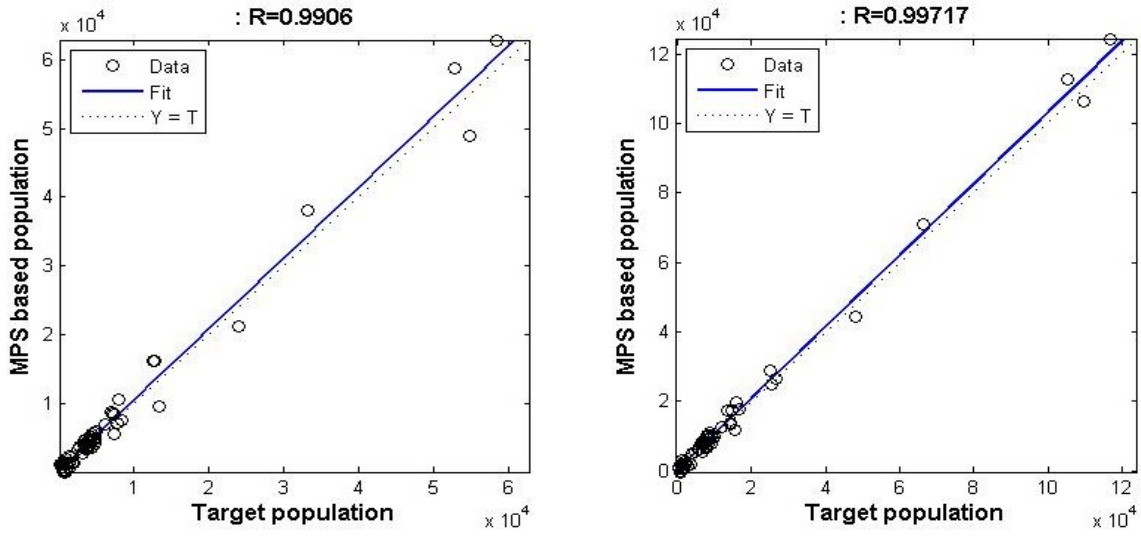


Figure 7.3 WHPL model: individual and household level's goodness-of-fit comparison

To measure the model dispersion, results for 5% and 10% sample of the synthesized population are compared. These results are examined at the regional level (RL) with the best fit sub-model (WHPL) and are shown in Figure 7.4. It is clear that the fit between the target population and the synthesized population in the 10% sample is better than in the 5% sample.



5% sample 10% sample

Figure 7.4 Dispersion comparison between 5% and 10% sample of RL model

Comparison between number of households' selection in the RL model for both the proposed population synthesis model and FBS algorithm are shown in Figure 7.5 and Figure 7.6. To better illustrate the difference in the two households' selection modes, Figure 7.6 shows only the first 100 iterations in the procedure. Comparison between the two models shows that in the first 100 iterations of the procedure, in the population synthesizer procedure approximately 582,839 individuals and 380,658 households are selected and added to the final synthesized population list while the FBS algorithm selects just 100 individuals and 39 households during first 100 iterations. The direct effect on the improvement in computation time is over 99.98%. Moreover, Figure 7.5 highlights that

the population synthesizer model is terminated after 4,794 iterations, while the FBS algorithm has not completed the computational process.

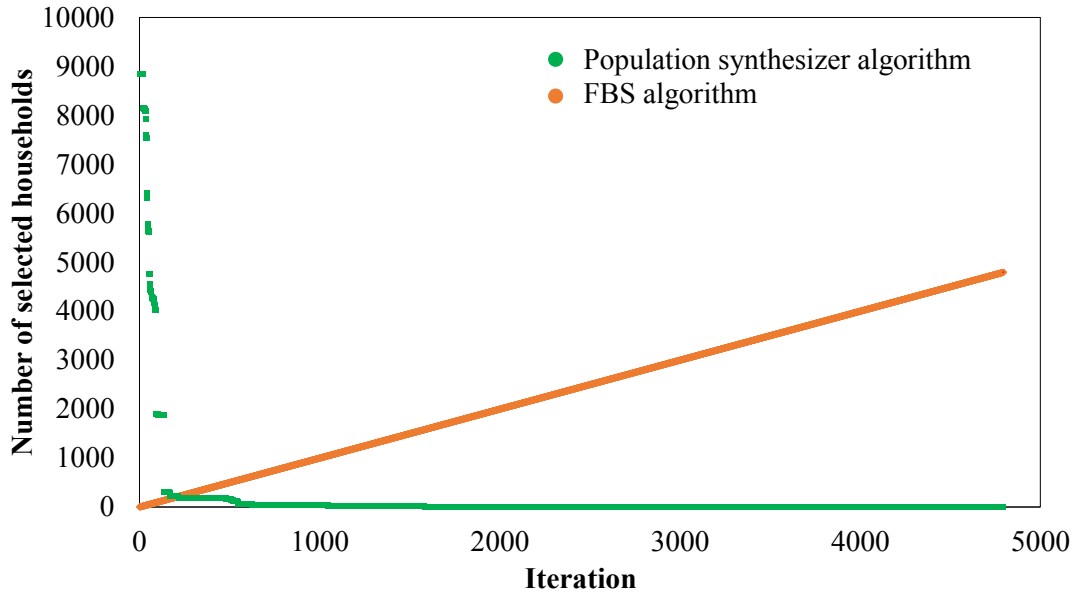


Figure 7.5 Comparison between numbers of households' selection in the RL model

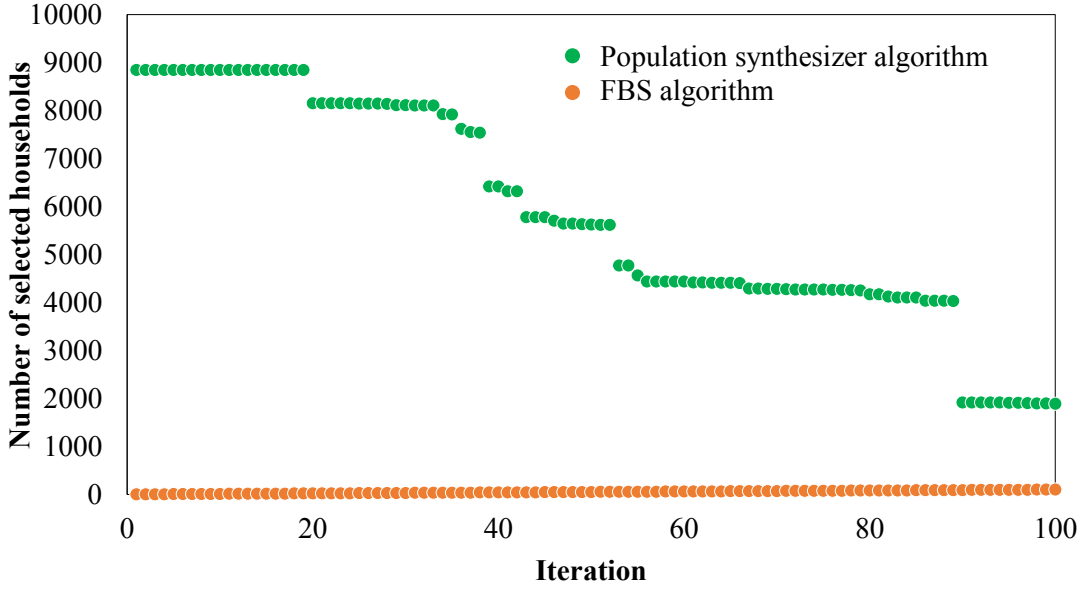


Figure 7.6 Comparison between numbers of households' selection in the RL model (first 100 iterations)

7.5.2 Synthetic Baseline Population at the Dissemination Area (DA) Level

Additionally, error percentages were calculated for the dissemination area (DA) model with the WHPL model that has the highest accuracy in comparison with other models. Table 7.5 presents the error percentages for the randomly selected 10 DAs.

Table 7.5 Error percentages of base year synthesized population (DA level)

Explanatory variables	DA ID									
	12090751	**12090672**	**12090873**	**12090871**	**12090869**	**12090660**	**12090879**	**12090876**	**12090875**	**12090874**
Individual level										
Age	0.2%	0.3%	0.3%	0.4%	0.4%	0.2%	0.4%	0.2%	0.3%	0.2%
Legal marital status	0.2%	0.3%	0.3%	0.4%	0.4%	0.2%	0.4%	0.2%	0.3%	0.2%
Gender	0.1%	0.2%	0.4%	0.5%	0.3%	0.0%	0.0%	0.2%	0.4%	0.3%
Household level										
Household size	0.2%	0.1%	0.3%	0.3%	0.3%	0.2%	0.1%	0.2%	0.4%	0.2%
Type of household	0.1%	0.3%	0.2%	0.2%	0.3%	0.1%	0.3%	0.1%	0.1%	0.2%
Tenure	0.2%	0.3%	0.2%	0.2%	0.2%	0.3%	0.1%	0.1%	0.2%	0.2%
Dwelling type	0.2%	0.3%	0.2%	0.2%	0.2%	0.3%	0.1%	0.1%	0.2%	0.2%
Labor force activity	0.0%	0.1%	0.2%	0.3%	0.3%	0.1%	0.4%	0.1%	0.2%	0.2%

Figure 7.7 demonstrates summary statistics of absolute difference by DA size for the age attribute only (for illustration purposes). Based on the DA size (measured in terms of total population), absolute discrepancy between synthesized and target population for nine groups with respect to age attribute is calculated that yields min, max, median, quartile one, and quartile three. The resulting box-plot charts shows that the proposed population synthesizer approach produced better results for smaller sized DAs in comparison to larger sized DAs. It is interesting to note that with the increase in the size of DA, there is an increase in the absolute discrepancy. In contrast, due to the higher number of larger DAs

and their relatively smaller absolute discrepancies, error percentage is small among larger size DAs.

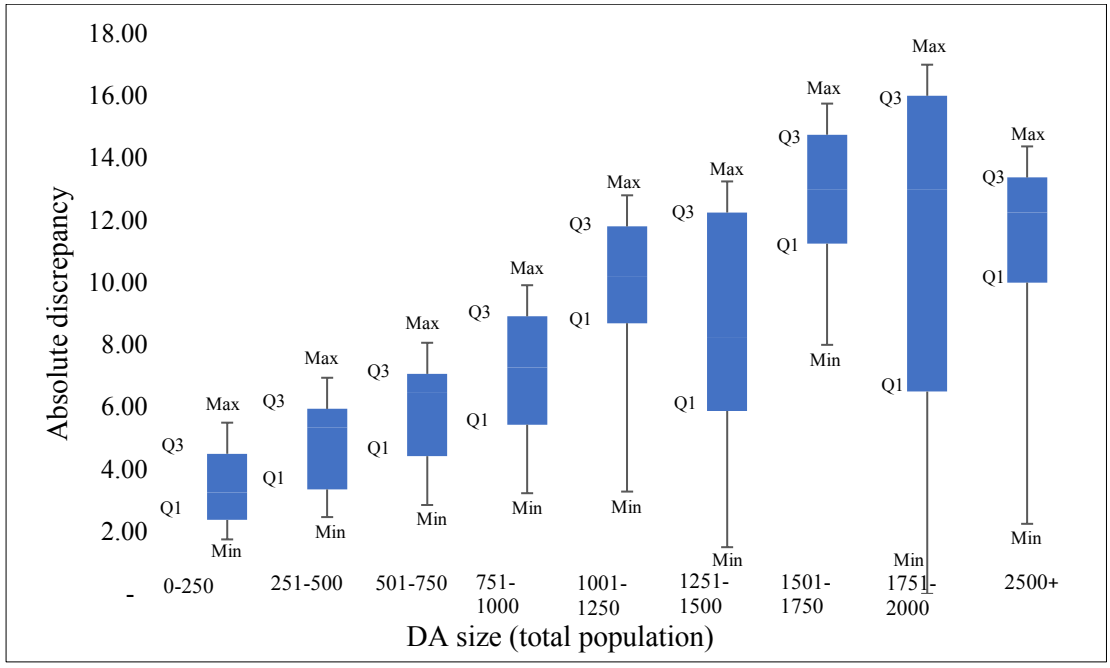


Figure 7.7 Summary statistics of absolute difference for age attribute by DA size

7.5.3 Synthetic Baseline Population for the University Community

A population synthesis technique was employed to match the survey sample size with the actual university population and produce a 100% synthetic population for all four commuter groups of Dalhousie city campuses. We used EnACT survey sample data as seed data and control tables were obtained from the specific Dalhousie population characteristics derived from the 2016 Dalhousie Analytics Data. The 100% synthetic population of the base year for all Dalhousie university groups (undergraduate students, graduate students, faculty members, and staff) was generated. The comparison between observed population and synthetic population, based on the population characteristics (gender, age and household income), is shown in Figure 7.8. The population synthesis

algorithm represents the seed data (observed population) for most of the count and corresponding control tables within an error percentage of -1.0% to +1.0%.

The statistical measurement of goodness-of-fit suggests an overall r-squared value of 0.978, indicating a very close level of fit. Figure 7.9 illustrates the distribution of transport mode, vehicle availability, and living arrangement between the observed population and synthetic population. As can be seen, in most cases the population synthesis algorithm accurately replicated attribute distributions in the sample population with respect to the observed attributes. Furthermore, the slight differences between synthetic and observed population can be handled through applying the normalization factors to the synthetic trips in the quantification of emission reduction benefits (Mitloehner 2016).

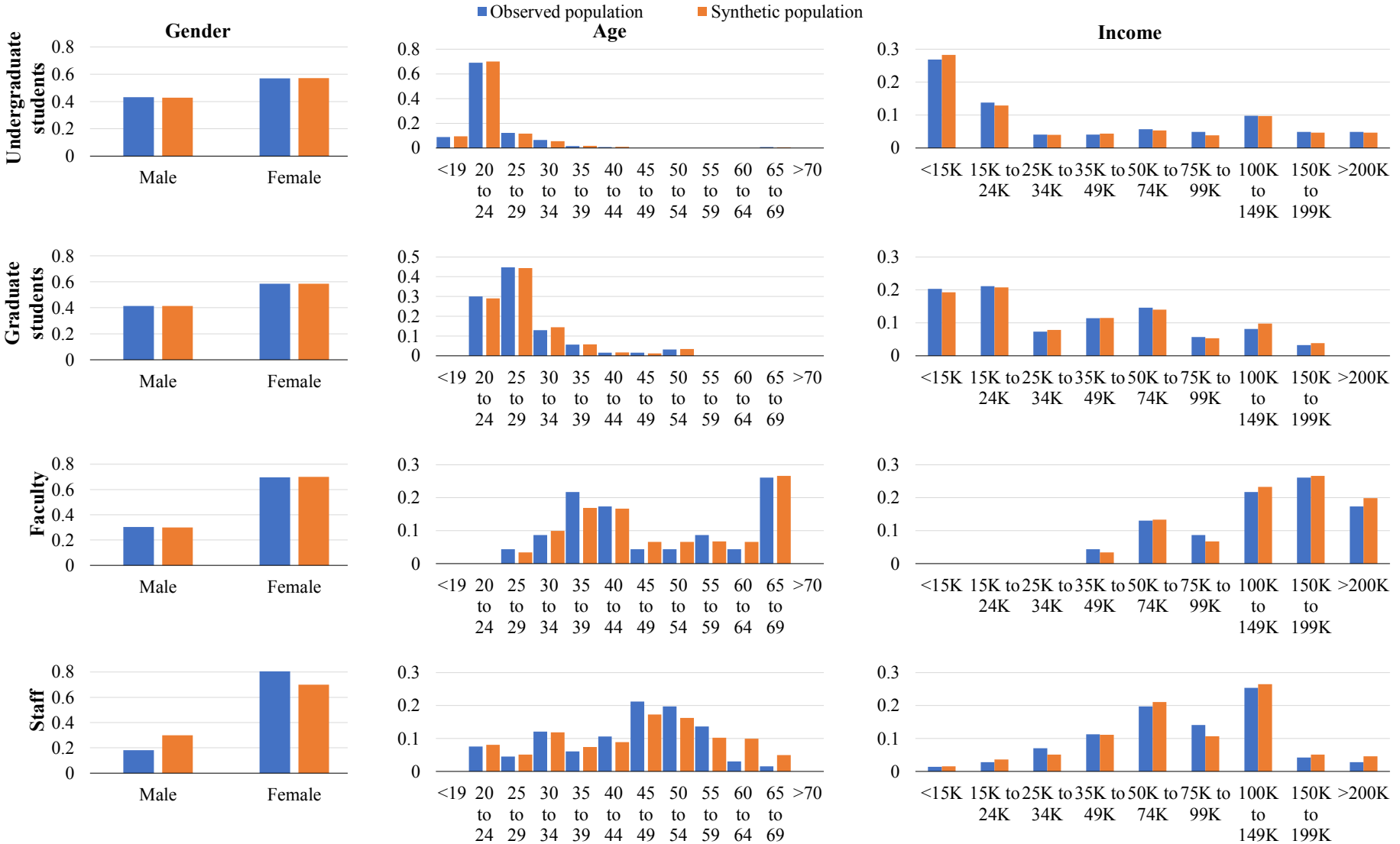


Figure 7.8 Comparison of gender, age and household income attributes between observed population and synthetic population

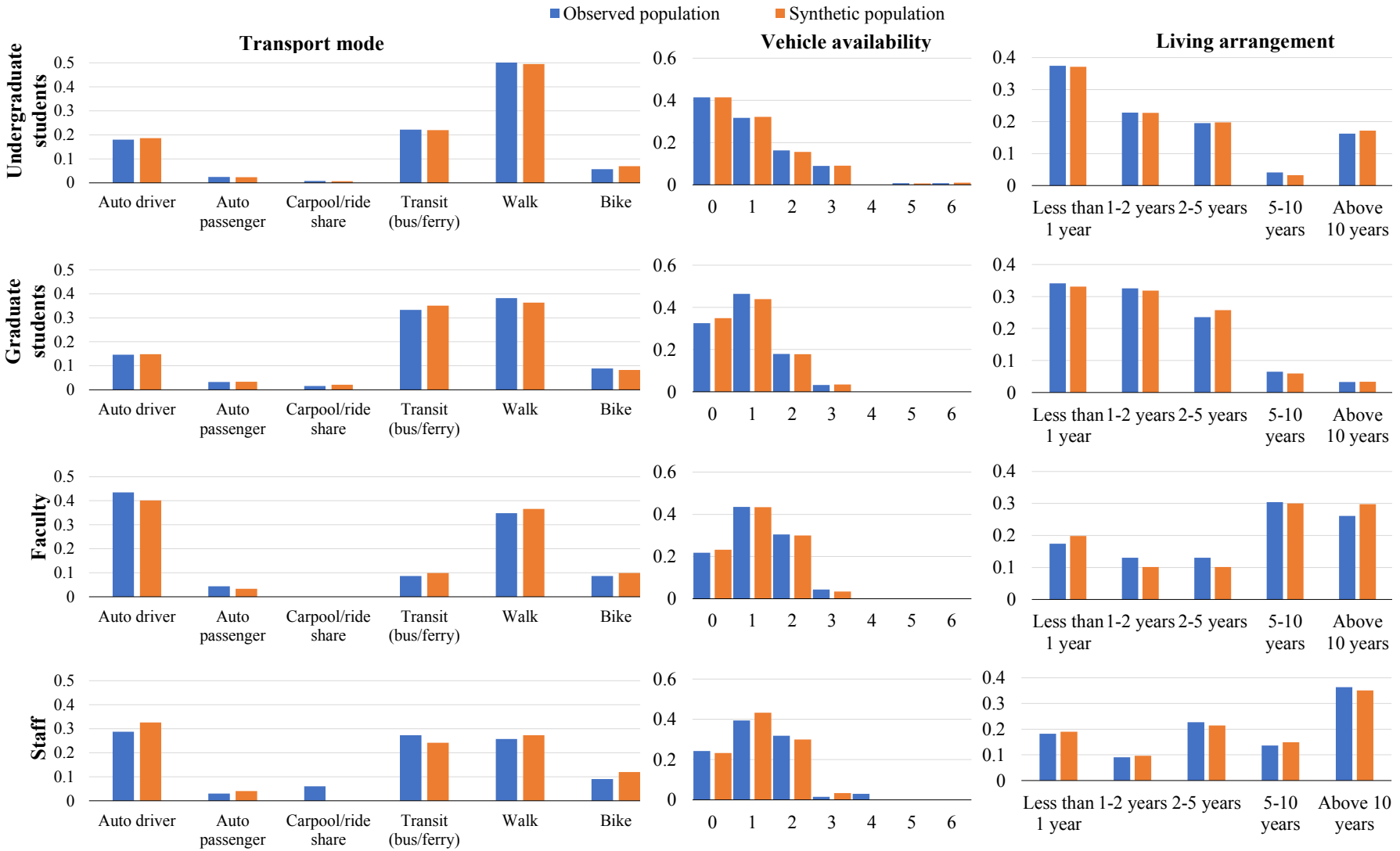


Figure 7.9 Distribution of transport mode, vehicle availability and living arrangement between observed population and synthetic population

7.6 Conclusions

In this study, we proposed a population synthesizer approach to synthesize population for activity-based travel demand model systems. This technique is able to synthesize population at both the regional level (RL) and the dissemination area (DA) level. Population is synthesized first at the regional level. Different types of households which are represented in the control table and seed data from the corresponding DA are reproduced. Additionally, utilizing larger scale empirical testing under different sets of control tables achieved the best set of control tables for synthesis of population at the smaller spatial unit. Moreover, results at this stage of modeling show that due to model overfitting, more control tables cannot increase the goodness-of-fit of synthesized population. Likewise, three different sub models, the HL model (only using the household level control tables), the HPL model (using both individual and household level control tables), and the WHPL model (weighted individual and household level control tables) were tested to assess the performance of the algorithm. This test is accomplished to show the performance and accuracy of the algorithm under different control tables.

The new selection procedure of the population synthesizer approach caused a synthesis of the complete set of potential individual and households members in a fewer run. Traditionally, one individual or household was selected and synthesized per one iteration. The new approach resulted in improved efficiency of the algorithm, both in computational time and distribution of seed data presented in the final synthesized population list. However, it is observed that absolute discrepancy increases with increase in the size of the DA. Future studies will further evaluate the performance of the proposed approach for other attributes and population sizes. As demonstrated, the proposed approach can

efficiently achieve an acceptable outcome using both individual and household level control tables. To measure the model dispersion, results for 5% and 10% samples are compared.

Future work includes examining the performance of the proposed approach for target year population synthesis, evaluating the performance of the algorithm in the case of uncontrolled attributes, and further tests of homogeneity. The results of this study are expected to be implemented within the activity-based travel demand model for Halifax, Nova Scotia, Scheduler for Activities, Locations, and Travel (SALT).

Chapter 8 Daily Activity and Travel Sequences of Students, Faculty, and Staff at a Large Canadian University[6]

8.1 Introduction

Recent transportation planning research suggests that commuters to large universities should be considered as a sizeable sub-population that needs special consideration in regional travel demand models. This is partly due to this sub-group's unique accessibility, mixed-use, higher density, and alternative mode friendly environment. In addition, this sub-group is typically under-represented in regional travel surveys: survey are not well able to target this sub group's student members for a variety of reasons, such as using random-digit dialing of landlines to reach individuals. Previous studies show that a high proportion of university populations are mobile-phone-only users (Wang, Khattak and Son 2012; Hafezi et al. 2017). As a result, the travel behavior of this sub-group is not suitably modeled or well understood in regional travel demand models. The few regional travel demand models that have modeled university travel behavior consider this sub-group as a special generator attractor that employed external trip rates (Eom, Stone and Ghosh 2009; Hafezi, Liu and Millward 2018b).

Many universities in the United States and Canada use university-based travel demand management strategies to find alternative solutions for parking and fleet management, car sharing, and to promote active transportation, etc. (Axhausen 1996; Black, Mason and Stanley 1999; Daisy et al. 2018a). However, there are few studies (mostly in the United

[6] A version of this chapter has been published:
Hafezi, M. H., N. S. Daisy., L. Liu., and H. Millward. (2018). "Daily activity and travel sequences of students, faculty, and staff at a large Canadian university". Transportation Planning and Technology.

States) that conducted a university-based travel diary survey and modeled the travel behavior of a university population (Eom, Stone and Ghosh 2009; Wang, Khattak and Son 2012; Volosin et al. 2014). The current study is also unique in terms of collecting the first university-based travel diary survey across Canada. The student sample population for Nova Scotia in the 2010 Canadian General Social Survey (GSS) is only 26 individuals (Statistics Canada 2010, Hafezi et al. 2017b), which is not representative. Also, the student cluster identified in the Halifax Space-Time Activity Research survey (STAR) has small sample size and is not specific to a university population (Hafezi, Liu and Millward 2017c).

To this end, this study aims to fill the gap by exploring the travel behavior of the university population at Dalhousie University by modeling daily activity patterns (DAP) of undergraduate students, graduate students, faculty, and staff. The data used in this study were obtained through a unique online travel diary survey conducted at Dalhousie University campuses in spring 2016. The results of this study are expected to be incorporated within the activity-based travel demand model, Scheduler for Activities, Locations, and Travel (SALT) for Halifax Regional Municipality (HRM), Nova Scotia (Daisy et al. 2017; Hafezi, Liu and Millward 2017c). The findings of this study also provide deeper insights for modeling the travel behavior of North American university populations in general.

The remainder of this study is structured as follows: first, the study provides a review of relevant past research concerning university population segments and student travel. Secondly, we discuss the travel diary survey methods. Then, we present methods to explore and predict the daily activity travel patterns of the four university population

segments, followed by presentation and discussion of survey and modeling results. The study concludes with a summary of contributions and brief discussion of future research directions.

8.2 Literature Review

Transportation engineers and metropolitan planning organizations suggest that large university populations should be considered as a sub-population with special travel behavioral characteristics in the regional travel demand models (Daisy et al. 2018b). Nevertheless, previous studies reveal that university population samples are under-represented in regional travel surveys. As a result, the travel behavior characteristics of this group are not well recognized in regional travel demand models. In the latest public use micro data of GSS for Nova Scotia, only 26 individuals with full-time student status are recorded. This sample size is not representative of student groups and it does not include other university populations such as faculty and staff. Balsas (2003) compared eight pre-selected commuting surveys among different American universities. Balsas concluded that the travel behavior of university communities needs special attention, since there are distinguishable differences between the travel behavior of university populations and general populations. Balsas argued that these differences exist for a variety of reasons, such as mixed land-use, livable environments, higher density, and the university's friendly setting for alternative travel modes.

In the past decade, transportation engineers and metropolitan planning organizations have begun to model the travel behavior of university communities as a special trip generator in regional travel demand models. Given the importance of studying the university

population, there exist only a limited number of peer-reviewed literature on university travel behavior. Table 8.1 presents a summary of these existing studies.

There is considerable variety in case studies by both location (urban to sub-urban universities) and survey period (travel diary recording from one-day to one-week). The Institute of Transportation Studies at University of California at Davis conducted a yearly travel survey which provides a longitudinal perspective for university travel demand (Popovich 2014). In most cases, university campus based travel surveys considered the travel behavior of student populations only, and excluded faculty and staff. There have been a variety of research focuses related to these surveys, such as commuting mode choices, the effect of built environment on travel behavior, commuting patterns, and active transportation. Rodriguez and Joo (2004) investigated the linkage between built environment characteristics and mode choice using activity travel data from the University of North Carolina-Chapel Hill campus. Shannon et al. (2006) explored commuting patterns of student and staff groups using one-week travel survey data from the University of Western Australia (UWA). They found that the percentage of transit usage was greater in home locations farther from campuses, whereas walking and bicycling percentages were higher in areas close to campus.

In another study, Kamruzzaman et al. (2011) investigated the university students' out-of-home travel and activities characteristics. They conducted a two-days trip diary survey at the University of Ulster and Queen's University, Belfast. They found that female students visited more unique locations in comparison to male students. They reported that students traveled to 3.59 unique locations, on average. Eom et al. (2009) used one-day activity travel survey data of North Carolina State University (NCSU) to explore the activity-travel

characteristics of university students. In comparison with other studies, they developed a temporal and spatial activity-based model for university commuters. They found that undergraduate students and students living on campus participate in higher number of activities compared to graduate students. Also, they compared survey results with the Triangle Regional Model household travel survey in order to examine the differences of travel behavior between university students and the general population. Consistent with previous studies, survey results revealed that student trip rate is notably higher than the trip rate used in the regional model.

In another study, Khattak et al. (2011) conducted a one-day travel survey of university students at Old Dominion University (ODU) and Virginia Tech (VT). Similar to Eom et al. (2009), they found that undergraduate and on-campus students made more trips per day than graduate and off-campus students. Chen (2012) developed a statistical and activity-based model of university students by using a one-day travel survey at Virginia Commonwealth University (VCU). Consistent with previous studies, he found that mode choice, trip frequency, and activity participation of university students are different from those of the general population.

Table 8.1 Summary of existing university and student travel studies

Study	Case study	Target population	Study duration	Motivations/goals
Balsas, 2003	-	-	-	Develop a sustainable transport for US university campuses
Rodriguez and Joo, 2004	University of North Carolina-Chapel Hill (UNC)	Students and staff	One-day	Investigation of linkage between built environment characteristics and mode choice
Ubillos and Sainz, 2004	University of Bilbao (UPV/EHU)	Students	One-day	Investigation of the effect of quality and price on the demand for urban transport
Heung et al., 2006	Hong Kong University (HKU)	Students	One-day	Investigation of the travel behavior of university students
Shannon et al., 2006	University of Western Australia (UWA)	Students and staff	One-week	Investigation of the commuting patterns of students and staff
Villanueva et al., 2008	University of Western Australia (UWA)	Students	One-week	Investigation of the linkage between active transportation and university students
Eom et al., 2009	North Carolina State University (NCSU)	Students	One-day	Investigation of the activity/travel characteristics of university students
Kamruzzaman et al., 2011	University of Ulster (UU) - Queen's University	Students	Two-days	Investigation of university students' out-of-home travel and activities characteristics
Limanond et al., 2011	University of Thailand (SUT)	Students	One-week	Investigation the travel patterns of on-campus students
Khattak et al., 2011	University of Virginia (UVA) - Virginia Tech (VT) - ODU - VCU	Students	One-day	Investigation of the travel behavior of university students
Akar et al., 2012	Ohio State University (OSU)	Students, faculty and staff	One-day	Investigation of the mode choice for university commuting
Chen, 2012	Virginia Commonwealth University (VCU)	Students	One-day	Investigation of the travel behavior of university students
Rissel et al., 2013	Sydney University (USYD)	Students and staff	One-day	Investigation of the travel mode and physical activity of university populations
Popovich, N., 2014	University of California at Davis (UCD)	Students, faculty and staff	One-day (annual travel survey)	Investigation of a longitudinal perspective for university travel demand
Volosin et al., 2014	Arizona State University (ASU)	Students, faculty and staff	One-day	Investigation of the activity travel characteristics of university populations

Our literature review suggests that modeling of university travel behavior needs special attention in regional travel demand models. To the best of our knowledge, the activity travel behavior in the context of Canadian universities has not yet been explored. This study aims to fill the gap by exploring and evaluating empirical data on travel behavior for a large university community. The results of this study are expected to be implemented within the activity-based travel demand model, Scheduler for Activities, Locations, and Travel (SALT).

8.3 Survey and Data Description

The data used for this study are derived from the Environmentally Aware Commuter Travel Diary (EnACT) Survey, an online web-based survey conducted by Dalhousie Transportation and Environmental Simulation Studies (TESS) group in spring 2016 at Dalhousie University, Nova Scotia. Dalhousie University is the largest university in the Maritime Provinces of Canada, with four campuses spread across Nova Scotia province (three urban campuses in the city of Halifax and one in the town of Truro). In this study, we focused on the city campuses only. The entire university community, including undergraduate students, graduate students, faculty, and staff, were considered as the target population of the survey. After a pilot study, through the cooperation of the university administration the survey link was sent to the entire university community. Respondents were asked to complete a 24-hour travel log (using 5-minute time segments) and also provide detailed individual and household information. The survey was dynamically designed (include branching and piping options) in order to reduce response burden. Respondents who completed the survey were entered into a random draw for $300 and $250 gift cards. The EnACT survey includes six sections: (1) household information, (2)

individual information, (3) environmental attitudes and behavior, (4) attitudes toward transportation, (5) information and communications technology (ICT) related information, and (6) a 24-hour travel log. Previous studies reveal that appropriate representation of university communities as a large sub-population is essential in regional travel demand models (Black, Mason and Stanley 1999; Eom, Stone and Ghosh 2009; Wang, Khattak and Son 2012). Therefore, the questions and activity log were designed to be consistent with the Canadian GSS instrument and the Halifax Space-Time Activity Research survey (STAR) which represented the world's largest deployment of global positioning system (GPS) technology for a household activity survey (TURP 2008; Millward and Spinney 2011), so that the sub-models of this survey can be utilized to improve the accuracy of regional travel demand models. However, sections 3, 4 and 5 were not included in the traditional GSS. Moreover, the traditional GSS cannot entirely capture trip generation of university community. Therefore, the EnACT survey was redesigned and adjusted to ensure that the survey was able to efficiently capture university community travel behavior. In comparison with other similar university travel diary surveys, the EnACT survey is unique in many aspects, including the consideration of simultaneous activities, ICT usage for trip purpose, and environmental attitudes.

Following survey data collection, and after rigorous error-checking, cleaning, and geo-coding, the survey yielded a sample of 346 entirely completed 24-hour travel logs for the city campuses. Descriptive statistics of respondents' individual and socio-demographic characteristics are presented in Table 8.2.

Table 8.2 Descriptive statistics of respondents characteristics

Variable	Description	Coding	Statistics	
Gender	Male, Female	Male: 0	130 (37.58%)	
		Female: 1	216 (62.42%)	
Age	Average age	Continuous	30.96	
License	Having a valid driver license	Have: 0	307 (88.73%)	
		Don't have: 1	39 (11.27%)	
Vehicle	Average number of vehicle available in household	Discrete	1.02	
Bicycle	Average number of bicycle available in household	Discrete	0.96	
Respondent status	Student, non-student	Undergraduate Student: 1	129 (37.28%)	
		Graduate Student: 2	126 (36.42%)	
		Faculty member: 3	24 (6.94%)	
		Staff: 4	67 (19.36%)	
Employment/student status	Full-time, not-full-time	Full-time (+30 hours per week): 1	111 (32.09%)	
		Part-time (<30 hours per week): 2	31 (8.96%)	
		Student full-time: 3	180 (52.02%)	
		Student part-time: 4	16 (4.62%)	
		Volunteer work: 5	8 (2.31%)	
Vehicle (person/veh)	Number of person per vehicle	Discrete	1.48/veh	
Mode	Travelling mode	Car (driver): 1	23.7%	20.00 (min)*
		Car (passenger): 2	3.7%	27.50 (min)*
		Walk: 3	36.35%	15.00 (min)*
		Bus: 4	25.9%	25.00 (min)*
		Bike: 5	10.0%	15.00 (min)*
		Boat/Ferry: 6	0.35%	22.5 (min)*

**Travel time (median)*

Consistent with other university travel diary surveys, the proportion of full-time students/workers is higher than part-time students/workers. The average age of respondents is around 30.96 years. Furthermore, around 88.73% of respondents have a valid driving license. Average number of vehicles available per household is 1.02, while the average number of bicycles is around 0.96. Across all travelling modes, the walk mode has the highest percentage of 36.35% trips, followed by auto drive and transit. Respondents who walk or bike have the smallest median time per trip (15 min). On the other hand, the auto passenger mode has the highest travel time (27.5 min).

A comparison of the sample distribution of EnACT respondents' key characteristics compared to those of the Dalhousie University population (using Dalhousie Analytics Data and Dalhousie Commuter Survey) showed that the sample is roughly representative with respect to gender, age, and employment status, commute time and travel mode. However, female individuals are slightly over-represented and older age individuals are slightly under-represented. As it was out of the scope of the current study, the samples used in this study were used directly from the survey without considering sample weighting or population synthesis techniques to match the sample distribution. A comprehensive descriptive analysis of all six sections of the EnACT survey can be found in Liu et al. 2016.

8.4 Methods

8.4.1 Daily Activity Travel Patterns

In In this study, we employed a series of statistical and rule based methods to explore and model the daily activity travel patterns of a university population. The frequency and time allocations of daily activities, the overall sequencing of activity episodes (through implementation of a transition matrix) and the daily activity patterns (both at the aggregated and disaggregated levels) by university community groups were investigated. Furthermore, the Kolmogorov-Smirnov statistical test was performed in order to examine the similarities and dissimilarities of temporal distributions associated with the aggregated activity categories for each of the university groups. The summary statistics presented in this study provide detailed guidelines for developing a statistical/econometric model capable of predicting the activity travel behavior of the university population.

In order to predict daily activity patterns (DAP), the rule-based decision tree clustering analysis method was utilized by employing the classification and regression tree (CART) classifier algorithm. The CART algorithm is an accurate decision tree algorithm that can construct the best tree that comprises the most information. The decision tree discerns the association between predictor (DAP) and response variables and classifies those predictor values that have the most notable associations to the response value. The response values used in the proposed model include socio-demographic characteristics such as age, gender, income level, education level, and car ownership, and travel attributes such as travel mode and travel distance. Consistent with other decision tree algorithms such as CHAID, C4.5, and ID3, in the CART algorithm the impurity measure is a decision maker for searching leaf nodes, and subsequently building the best fitting decision tree (Tan, Steinbach and Kumar 2005). Each leaf node includes all the identified predictor values that have high associations to the response values. The impurity measure in the CART algorithm is completed by estimating the Gini index. The Gini index is computed for every predictor variable at each node and the one that has the lowest value is chosen. Furthermore, the CART algorithm utilizes cross-validation as an additional measurement to choose the optimal decision tree. The Gini index is calculated as follows:

$$G(S) = 1 - \sum_{i=1}^{m} p_i^2 \quad (1)$$

Where:

m is the number of classes,

p_i is the relative frequency of class i in the data set S.

The dataset was divided into two sub-datasets. The first part provided training samples that comprised 75 percent of the observations of the travel diary survey data. Trees were built using the training samples. The second part provided testing samples that comprised the remaining 25 percent of observations and were used to check the precision level of proposed prediction models. An exception was made for the faculty group, where training and testing samples size were set to 60 and 40 percent due to their small share in the dataset. The class label (DAP) was predicted using the following method (Hand, Mannila and Smyth 2001):

$$\hat{d} = \underset{d = 1, \dots n}{arg} min \sum_{n=1}^{n} \hat{d}(n|m)\, T(d|n) \quad (2)$$

Where:

$\hat{d}$ is the predicted classification,

n is the number of classes,

$\hat{d}(n|m)$ is the posterior probability of class n for observation m,

$T(d|n)$ is the classification cost of an observation as d when its true class is n.

8.4.2 Transport-related GHS Emissions

Transport-related GHG emissions in this study are estimated based on the emission estimation framework developed in MOVES 2014a (Koupal et al. 2002; Hafezi et al. 2018b). The major air pollutants, including carbon dioxide (CO_2), carbon monoxide (CO), nitrogen oxide (NO_x), particulate matter under 10 micron diameter (PM_{10}), particulate matter under 2.5 microns diameter ($PM_{2.5}$), total-hydrocarbon (THC) and volatile-organic-

compounds (VOC), were derived from the MOVES model, and adjusted with input data representing the Halifax city context. The project scale analysis in MOVES 2014a is used for estimating the transport-related GHG emissions. Vehicular age distribution and fuel characteristics are drawn from the EnACT survey data and Canadian Vehicle Survey (Statistics Canada 2015). Using the information derived from the Halifax Regional Municipality database (HRM 2016), transit passenger kilometer estimates were calculated (total transit kilometers divided by the average transit occupancy) and vehicle speed information was obtained. The road network database was developed in ArcGIS 10.2.2 using GeoBase - National Road Network (NRN) - Nova Scotia, available at the Natural Resources Canada database (NRN 2016).

In this study, auto trips were all assumed to take the shortest distance. Transit routes are modeled based on the existing Halifax transit network map and transit distances were estimated on the street network. Meteorological input data such as humidity and temperature data for the study area were derived from the historical climate data at the Environment Canada archive (ENR 2016). Other required information (e.g. fuel supply specifications, fuel formulations, etc.) are drawn from the data of Cumberland county in Maine, United States, as a representation of the Halifax study area. Emission factors were estimated for each travel mode and university group, in grams per person per day. Home postal code and work (destination) postal code of all the respondents were geocoded in ArcGIS 10.2.2 and network commuting distances were computed using the network analyst tool.

In order to understand and explore the feasible emission reduction scenarios, five zones were utilized in this study (Millward and Spinney 2011). The on-campus zone (ONC) is

defined as the area within 1 km travel distance from centroids of any of the three Dalhousie city campuses. The inner city (INC) is defined as the area between 1 and 5 km from the campuses, and the suburban area (SUB) is defined as the area between 5 and 10 km from Dalhousie city campuses. The inner commuter belt (ICB) is defined as the area between 10 and 25 km from Dalhousie city campuses, and the outer commuter belt (OCB) is defined as the area between 25 and 50 km from Dalhousie city campuses.

Proportions of rural restricted, rural unrestricted, urban restricted, and urban unrestricted road types (as defined in MOVES) were fitted to these zones according to information obtained from the Halifax Regional Municipality. Respondents who usually choose walk or bike modes for commuting to campuses mostly fall in the ONC and INC zones. Respondents who choose the transit mode for commute trips mostly fall in the ONC, INC and SUB zones. Finally, respondents who usually choose motorized vehicles for commute trips are distributed in the ONC, INC, SUB and ICB zones, with a few in the OCB zone.

In this study, two alternative scenarios were examined, based on the spatial distribution of respondents, distance zone, and available travel mode. Respondents from the ONC zone were excluded in testing of both scenarios since the majority of respondents stated walk or bike modes as primary traveling modes for commuting to and from Dalhousie campuses.

Scenario 1 represents the total amount of emissions produced if 25%, 50%, 75% or all commuters living within INC and SUB zones use transit for commuting to Dalhousie city campuses. Respondents from the ICB and OCB zones were excluded due to the lack of availability of transit service in these areas (HRM 2016).

Scenario 2 represents the total amount of emissions produced if 25%, 50%, 75% or all commuters use a motorized vehicle for commuting to Dalhousie city campuses, and there is no ride sharing. In this scenario, we assumed car mode for all zones except the ONC zone, and mode as reported for the ONC zone. Adequate transit service and parking capacity were assumed to be available in both scenarios and new traveling distances are recalculated according to the changes in travel mode. The purpose of scenario testing is to exemplify how changing the primary travel mode can impact emissions volume.

8.5 Discussion of Results

This section focuses on exploring how the overall university community group organized their weekday activities, followed by an examination of how each sub-group sequenced the numerous activities in their daily activity schedule. Activity profiles for each population group are drawn in order to understand the daily activity sequencing, timing, and activity types. Next, results of Kolmogorov-Smirnov tests on activity start time distributions are presented, testing the similarity of activity profiles by respondent group. Finally, the results of the model prediction for the daily activity pattern (DAP) by university community groups are presented.

We first present the results of frequency and time allocations for daily activities. It should be noted that in this study, for the sake of brevity, we don't provide analysis for mode share and trip frequency. The main purpose of this study is to explore the similarity and dissimilarity in daily activity patterns of different university groups.

8.5.1 Frequency and Time Allocations for Daily Activities

Activities are classified into nine groups as follow: household related works, school/work, entertainment related activities, media and communication, organizational, voluntary and religious activity, personal care related activities (including sleep), shopping activities, care giving activities, and sports and hobbies. Table 8.3 presents frequency of activity types by trip purpose for all respondents.

In total there are 2,969 activity episodes undertaken by 346 individuals during the day. As expected, school/work related episodes were the most frequent activities in the day, with a proportion of 38.9%. Other high frequencies were personal care related activities, household related activities, and shopping. In contrast, organizational, voluntary and religious activity, and care giving activities had the smallest frequencies. Table 8.4 outlines activities by average episodes per day, median time allocations by different university groups and median travel time dedicated to the activity.

Table 8.3 Frequencies of major activities

Activity	Descriptions	Frequency of Episodes	Percent (%) of Episodes
Household related works	Cooking, cleaning, laundry, gardening, care for plants, other household related activities.	521	17.5
School/work	Full-time classes/research, work for pay at main job, other school/work related activities.	1154	38.9
Entertainment related activities	Socializing with friends/relatives at bars, clubs, movies/films at a theatre/cinema, art films, art exhibition, other entertainment related activities.	102	3.4
Media and Communication	Communicating over e-mail, Facebook, skype, twitter, Instagram etc., reading books/magazines/newspapers, other media and communication related activities.	87	2.9
Organizational, voluntary and religious activity	Organizational, voluntary & religious activity political, civic activities, other related activities.	45	1.5
Personal care related activities	Meals/snacks/coffee at home, private prayer, mediation, other spiritual activity, night sleep/essential sleep, other personal care related activities	721	24.3
Shopping activities	Shopping for goods & services, other related shopping activities	161	5.4
Care giving activities	Child care, personal care/help/medical care for household adults, other related care giving activities	44	1.5
Sports and hobbies	Other outdoor activities, walking, hiking, jogging, running etc., other related sports and hobbies activities	134	4.5

As expected, the most frequent out-of-home activity episodes are school and education related activities (1.85) and paid work (1.33), demonstrating that university community members typically participated in school/work activity (related to the university context) more than once a day. In terms of out-of-home daily time allocation percentages, school and work have the highest proportions. In general, faculty members allocated more time to in-home activities in comparison with staff. Also, undergraduate students allocated more time to in-home activities in comparison with graduate students. Interestingly, the results show that faculty members spent least time for shopping activities while undergraduate students spent the highest time for shopping activities among the four

groups. Finally, graduate students allocated more time to organizational and voluntary activities in comparison with undergraduate students, graduate students, and staff.

Table 8.4 Time allocations for daily activities

Activity	Average number of activity episodes[1]	Average duration (min)[2]				Percentage of the day spent				Median total travel time (min)[5]
		UnderGs[3]	Grads[4]	Faculty	Staff	UnderGs	Grads	Faculty	Staff	
Household related works	1.44	262.31	230.50	279.62	268.58	18.22	16.01	19.42	18.65	20.0
Paid work	1.33	90.19	22.91	424.04	525.68	6.26	1.59	29.45	36.51	22.5
Entertainment related activities	0.28	37.05	33.06	32.88	10.95	2.57	2.30	2.28	0.76	20.0
Media & Communication	0.24	23.81	33.76	20.96	37.91	1.65	2.34	1.46	2.63	30.0
Organizational, voluntary and religious activity	0.12	8.36	10.43	7.50	8.51	0.58	0.72	0.52	0.59	18.8
Personal care related activities	1.99	571.31	569.30	572.12	505.20	39.67	39.53	39.73	35.08	30.0
School & education related activities	1.85	392.91	475.62	27.12	3.18	27.29	33.03	1.88	0.22	20.0
Shopping activities	0.44	33.73	22.91	6.73	24.39	2.34	1.59	0.47	1.69	20.0
Care giving activities	0.12	0.45	11.16	25.77	20.61	0.03	0.78	1.79	1.43	15.0
Sports and hobbies	0.37	19.89	30.35	43.27	35.00	1.38	2.11	3.00	2.43	25.0
All Trips	8.18*	114**	93.4**	94.8**	76.7**	9.1**	8.8**	8.9**	7.1**	22.1*

[1]*Number of activities per individual per day*

[2]*Average duration per activity (∑ activity time / ∑ frequency of activity)*

[3]*Undergraduate students*

[4]*Graduate students*

[5]*Median minutes consumed on travel for each activity per day; *Average values, **Median values*

8.5.2 Overall Sequencing of Activity Episodes by University Community Groups

The sequencing of episodes for each university community sub-group, including both in-home and out-of-home activity episodes, was examined to better understand the

differences between activity profiles of each group. A transition matrix between consecutive activity episodes was created, to show the likelihood that a successive episode of a certain category will occur, given an episode of a current category (Lockwood, Srinivasan and Bhat 2005). Table 8.5 presents a transition matrix of all surveyed 2,969 activity episodes.

It presents detailed information on trip chaining patterns of different market segments in a more compact yet compelling way. The rows in Table 8.5 represent the category of the current activity episode, while the columns represent the category of the subsequent activity episode. The value calculated in each cell indicates the percentage of occurrence of a subsequent activity episode of a certain category after the current episode; the values in each row sum to 100. Many interesting findings from the transition matrix can be observed.

Table 8.5 Activity episode transitions (in percentage) matrix

			Subsequent activity episode									
			School/class	Work/paid job	Household work	Meals*	Organizational/ volunteer	Shopping	Sports/hobbies	Entertainment	Media/ communication	Personal care
Undergraduate	Current activity episode	School/class	*0.0*	5.4	10.5	25.9	6.7	13.1	4.1	5.4	6.7	22.0
		Work/paid job	2.7	*0.0*	12.3	26.5	7.5	2.7	12.2	2.7	12.1	21.7
		Household work	8.8	3.4	*0.0*	9.0	3.4	3.4	14.5	8.9	3.4	45.0
		Meals	40.9	17.5	0.3	*0.0*	0.3	6.6	1.9	5.0	5.0	22.2
		Organizational/volunteer	0.0	0.0	42.8	28.6	*0.0*	0.0	14.3	0.0	0.0	14.3
		Shopping	9.3	0.5	26.6	4.8	9.1	*0.0*	4.8	4.8	9.2	30.9
		Sports/hobbies	9.5	4.8	19.0	19.0	0.0	9.5	*0.0*	4.8	9.5	23.8
		Entertainment	7.7	1.4	26.4	20.2	1.4	7.7	7.7	*0.0*	13.9	13.9
		Media/communication	14.8	5.3	14.8	19.5	5.3	10.0	10.0	0.5	*0.0*	19.5
		Personal care	17.5	2.2	10.5	32.7	0.8	7.7	11.9	7.7	9.1	*0.0*
Graduate		School/class	*0.0*	2.1	26.1	16.5	3.0	13.7	13.5	5.9	5.9	13.6
		Work/paid job	10.0	*0.0*	33.3	0.0	35.7	0.0	15.0	0.0	0.0	6.0
		Household work	23.9	0.5	*0.0*	14.6	3.6	5.2	2.1	6.8	20.8	22.4
		Meals	39.2	0.2	8.7	*0.0*	3.6	3.6	5.3	1.9	8.7	29.0
		Organizational/volunteer	15.4	7.7	30.8	23.1	*0.0*	7.6	0.0	7.7	0.0	7.7
		Shopping	32.0	0.0	24.0	12.0	11.0	*0.0*	0.0	4.0	4.0	13.0
		Sports/hobbies	42.3	0.4	16.5	23.0	0.4	3.6	*0.0*	3.6	3.6	6.9
		Entertainment	29.6	1.8	12.9	7.4	1.8	7.4	18.5	*0.0*	7.4	12.9
		Media/communication	15.0	3.2	9.1	15.0	0.3	3.2	6.2	3.2	*0.0*	44.4
		Personal care	31.9	1.2	13.2	19.9	3.9	5.2	11.9	2.5	10.5	*0.0*
Faculty		School/class	*0.0*	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
		Work/paid job	0.0	*0.0*	38.7	9.3	5.3	8.3	21.3	5.3	1.3	10.3
		Household work	0.0	40.0	*0.0*	15.3	0.0	0.0	20.0	0.0	11.3	13.3
		Meals	0.0	47.0	8.5	*0.0*	1.6	0.8	16.2	9.5	7.5	8.5
		Organizational/volunteer	0.0	47.7	23.0	0.0	*0.0*	14.3	0.0	0.0	0.0	15.0
		Shopping	0.0	33.3	30.0	33.3	0.0	*0.0*	0.0	3.3	0.0	0.0
		Sports/hobbies	0.0	70.0	10.0	0.0	0.0	0.0	*0.0*	0.0	20.0	0.0
		Entertainment	0.0	0.0	33.3	33.3	0.0	33.3	0.0	*0.0*	0.0	0.0
		Media/communication	0.0	0.0	0.0	20.0	0.0	0.0	0.0	20.0	*0.0*	60.0
		Personal care	0.0	12.5	25.0	62.5	0.0	0.0	0.0	0.0	0.0	*0.0*
Staff		School/class	*0.0*	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
		Work/paid job	0.0	*0.0*	29.9	18.9	7.0	11.0	17.6	3.1	5.7	7.0
		Household work	0.0	22.7	*0.0*	11.6	7.2	12.6	9.4	0.5	25.4	10.6
		Meals	0.0	51.7	20.7	*0.0*	0.0	3.4	3.4	0.0	13.8	6.9
		Organizational/volunteer	0.0	49.9	0.0	14.7	*0.0*	0.0	18.7	0.0	0.0	16.7
		Shopping	0.0	38.4	23.1	7.7	0.0	*0.0*	15.4	7.7	7.7	0.0
		Sports/hobbies	0.0	53.4	24.3	14.8	0.5	0.5	*0.0*	0.5	0.5	5.3
		Entertainment	0.0	0.0	0.0	20.0	0.0	0.0	50.0	*0.0*	0.0	30.0
		Media/communication	0.0	16.7	27.8	27.7	0.0	5.6	0.0	0.0	*0.0*	22.2
		Personal care	0.0	43.3	32.8	11.7	1.2	2.3	1.2	6.5	1.2	*0.0*

**Meals category in this Table include meals/snacks/coffee break at home or school/work*

In general, across all four sub-groups, the most frequent subsequent activity episodes following the current activity episode are meals, personal care, and household work. For school/work activities, the most frequent subsequent activities are meals, personal care, and household work. Maintenance shopping episodes for undergraduate students and graduate students are very likely to be succeeded by an in-home activity episode. In contrast, maintenance shopping episodes for faculty and staff are very likely to be succeeded by work/paid job activity episodes. Presumably, this might be due to accompanying the spouse or children in shopping activities in case of faculty or staff, while students need to transport the purchased food/goods items back home and store/fridge them. Another interesting finding is that faculty and staff frequently have a sport activity episode preceded by a work activity episode. Finally, recreation activity episodes for undergraduate students and graduate students are very likely to be succeeded by an in-home activity episode. In contrast, recreation activity episodes for faculty and staff are very likely to be succeeded by an out-of-home activity episode (shopping or another recreation).

8.5.3 Daily Activity Patterns by University Community Groups

In order to understand and compare the daily temporal activity patterns of different university groups, an analysis on the sequence of activities, embedding the type of activities and their timing is performed. Figure 8.1 and Figure 8.2 shows the aggregated and disaggregated activity profile by different university groups. The horizontal axis covers a 24-hour time period beginning at 12:00 a.m. (midnight) and ending at 11:59 p.m. on the following day. The vertical axis indicates the proportion of individuals' engagement in each activity type according to time of day.

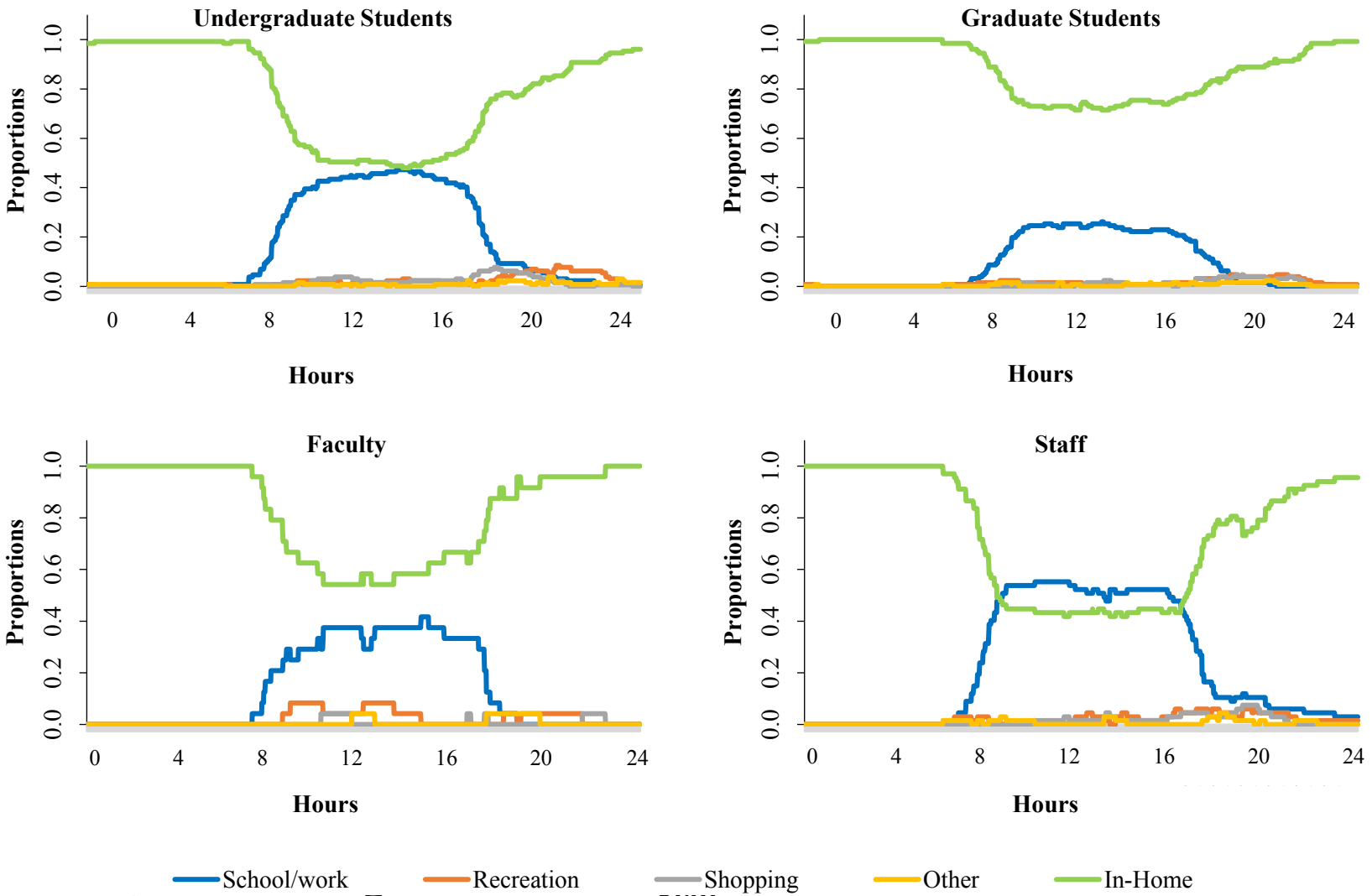


Figure 8.1 Aggregated activity profile by different university groups

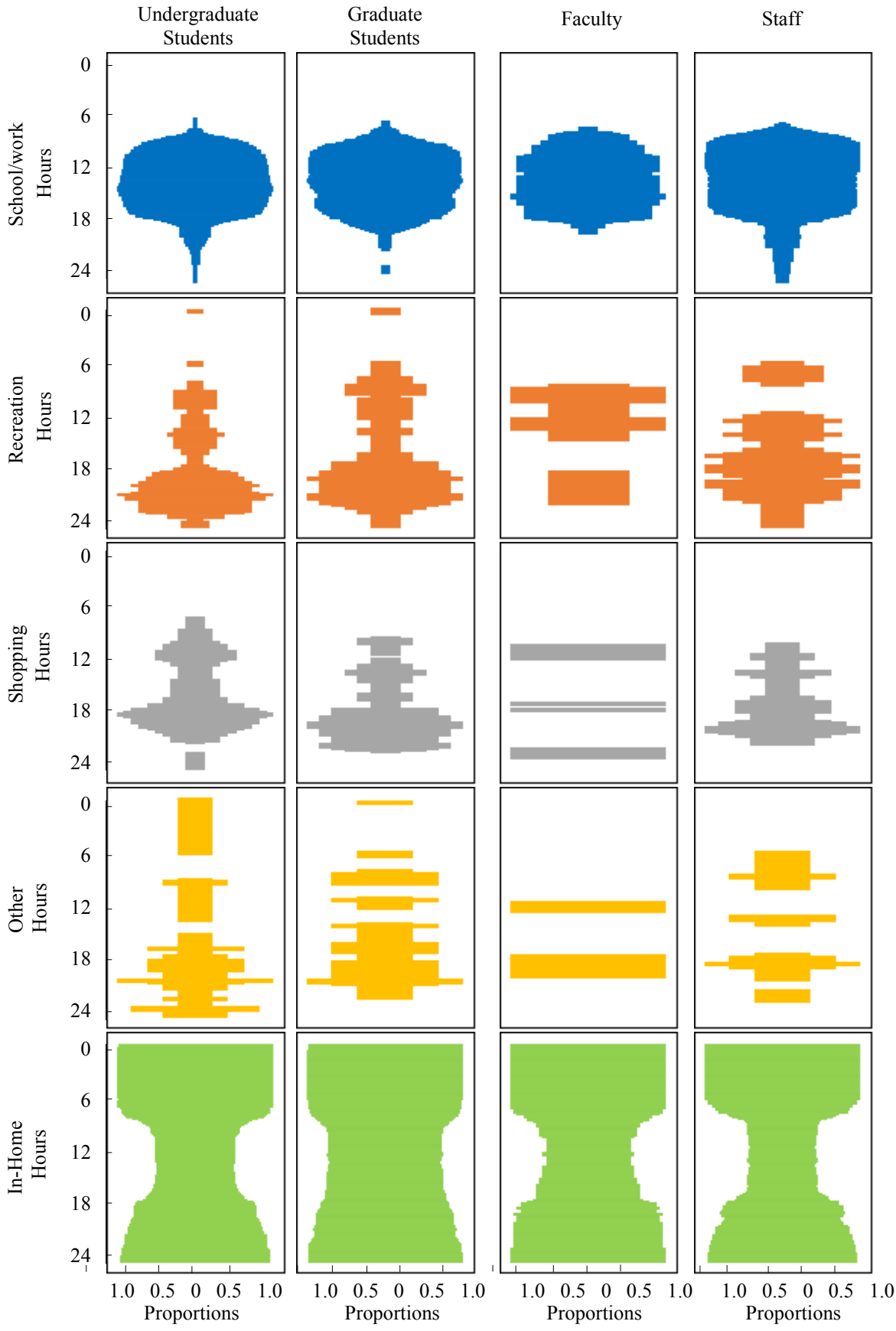


Figure 8.2 Disaggregated activity profile by different university groups

For illustration purposes, all activities are aggregated into five major activities: school/work (all school/work related activities), recreation (all eating out and leisure related activities), shopping (all shopping for goods and services related activities), other out-of-home activities, and in-home (all in-home related activities). One can clearly see the times of day when a large proportion of different university groups are at university between 8:30 a.m. to 4:30 p.m., though it appears that a small proportion of staff work the night shift until midnight. Moreover, staff seem to start their work earlier than other university groups, and their working hours of staffs are higher than those of faculty members.

Compared to students and staff, faculty members are found to engage in other non-work activities in between their regular working hours. This might be motivated by flexible working hours of faculty in comparison with other university groups. Graduate students spend more time in the school in comparison with undergraduate students. On the other hand, undergraduate students engaged more in recreational and shopping activities after school compared to graduate students. Most of the students undertake recreational and shopping activities in the evening. A small number of graduate students and undergraduate students participate in recreational and shopping activities, respectively, during the day. Presumably, students with flexible daytime schedules participate in discretionary activities during the day. This result is similar for other out-of-home activities conducted by both graduate and undergraduate students. Whereas faculty members are found to engage in fewer recreational and shopping activities than staff members, but with longer duration. Overall, in comparison to student groups, faculty and staff are found to engage in fewer in-home activities between 8:30 a.m. to 4:30 p.m. (working time at school). In part, this

reflects less discretionary time for these two groups, with their work requiring their presence on campus. These groups also tend to live further from campus, and thus tend to minimize trips to and from home.

8.5.4 Daily Time-Use Activity Patterns by University Community Groups

The best-fit decision tree model for the 10 most frequent daily activity patterns by university community groups is depicted in Figure 8.3. The identified daily activity patterns for each population group in every node of the tree model are drawn in order to better understand differences in daily activity patterns between different university community groups. Consistent with previous studies, tree model results revealed that the daily activity patterns with sequencing of home-school-home activities (H-S-H) and home-work-home activities (H-W-H) are the most frequent DAPs in the university travel behavior pattern. This result is highlighted in the first node of the tree model in Figure 8.3. The decision tree splits the daily activity pattern by employment status and results in classifying the data into two classes: students and non-students (faculty and staff).

The activity pattern with sequencing of home-school/work-shopping-home activities (H-S/W-G-H) is identified as the second most frequent DAP for both student and faculty/staff groups. Additionally, the activity pattern with sequencing of home-work-home activities (H-W-H), home-school-home-school-home activities (H-S-H-S-H), and home-school-home-shopping-home activities (H-S-H-G-H) are identified as the next most frequent activity patterns for students. In contrast, for faculty and staff groups the next most frequent DAPs include sport activity prior to or after work activity with one or two stops at home.

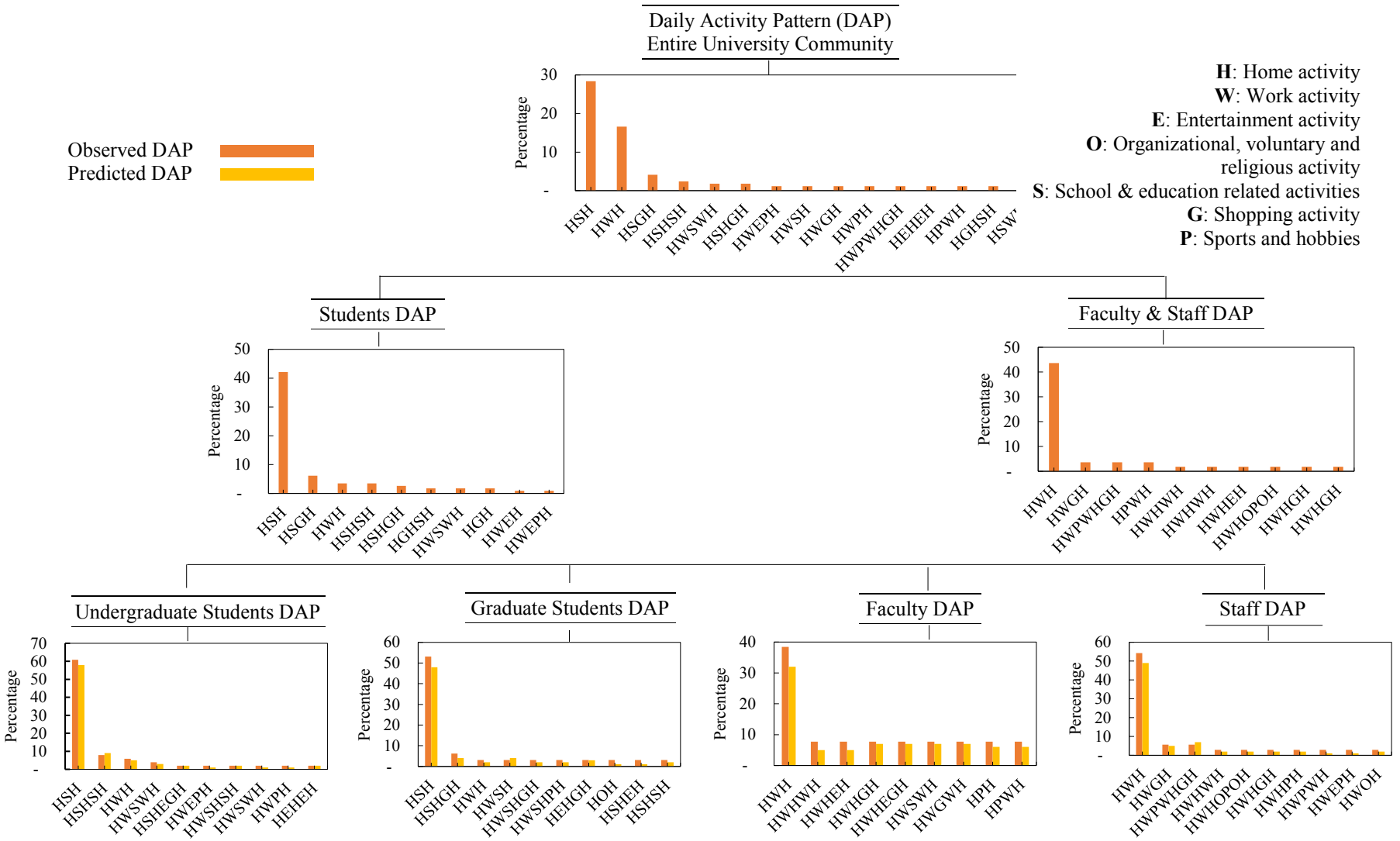


Figure 8.3 Predicted percentage of the most frequent daily activity patterns (DAP) by university community groups

In the next branch of the tree model, the decision tree splits the daily activity patterns by university community groups into four classes. The first identified class consists of undergraduate students' DAPs. Home-school/work-home activities with one or two stops at home (H-S-H, H-S-H-S-H, H-W-H, and H-W-S-W-H) are identified as the first four most frequent DAPs for undergraduate students. Additionally, activity patterns which include shopping, entertainment, and sports activities after school activity are identified as other frequent DAPs for this population group. The second identified class consists of graduate students' DAPs. The algorithm identified home-school-home activities sequence (H-S-H) as the most frequent DAP for graduate students, followed by home-school-home-shopping-home activities (H-S-H-G-H). The next most frequent DAPs for this group include shopping, sports, and organizational/voluntary activities after main activity (school/work activity). The third identified class consists of faculty DAPs. The most frequent DAPs for this group are sequences of home-work-home activities with one or two stops at home (H-W-H, H-W-H-W-H). The next most frequent activity sequences for this group include shopping and entertainment activities after main activity (work). Consistent with the findings of the previous section, an activity pattern with sequencing of home-sports-work-home activities (H-P-W-H) is also one of the most frequent DAPs for the faculty group. Lastly, the fourth identified class consists of staff DAPs. Similar to faculty DAPs, the most frequent sequence for this group is identified as home-work-home activities (H-W-H). In addition, the next most frequent activity patterns for staff members include shopping, organizational/voluntary, and sports activities after work activities.

As illustrated in Figure 8.3 predicted DAPs are closely similar to the observed DAPs in the travel diary dataset, particularly for the student and staff groups. However, the tree

model fails to predict faculty DAPs with high accuracy, due to their small share in dataset. Overall, however, the tree model is well able to predict the daily activity patterns of university community groups. It should be noted, too, that the decision tree model is expected to predict with higher accuracy when trained with a larger training sample.

8.5.5 Similarity Test of Activity Profiles by University Community Groups

In this subsection, the start time distributions for all aggregated activity categories for different university groups are examined using the Kolmogorov-Smirnov statistical test. The Kolmogorov-Smirnov statistical test can determine whether or not two non-parametric datasets are drawn from similar distributions (Sheskin 2003). In this study, we utilized the Kolmogorov-Smirnov statistical test in order to understand the differences between start time, duration, and end time distributions associated with all aggregated activity categories for each university group. For the sake of brevity, we state only results related to activity start times (though the same statistical test can be performed for duration and end time distributions). The Kolmogorov-Smirnov test is defined by:

$$T = \max_{1 \le i \le D} \left(CDF(Y_k) - \frac{i-1}{S}, \frac{i}{S} - F(Y_k) \right) \quad (3)$$

Where:

CDF is the cumulative distribution function,

i is the start time,

S is sample size.

As shown in Table 8.6, the null hypothesis associated with school/work activity for staff is rejected, indicating that this group has a significantly different start time of work activity in comparison with other university groups. Furthermore, graduate students, faculty, and staff have distinct start times in terms of recreation activity. Interestingly, it is found that different university groups have different start times for shopping activity. This difference might be motivated by varying levels of flexibility in study/work hours, specific needs of each group, and presence of accompanying persons.

Table 8.6 Summary of Kolmogorov-Smirnov test on activity start time distribution by different university groups (5% significance level)*

School/work activity	Group ID	UnderG[1]	Grads[2]	Faculty	Staff
	UnderG	0	0	0	1
	Graduate		0	0	1
	Faculty			0	1
	Staff				0

Recreation activity	Group ID	UnderG	Grads	Faculty	Staff
	UnderG	0	0	0	0
	Graduate		0	1	1
	Faculty			0	1
	Staff				0

Shopping activity	Group ID	UnderG	Grads	Faculty	Staff
	UnderG	0	1	1	1
	Graduate		0	1	0
	Faculty			0	1
	Staff				0

Other activity	Group ID	UnderG	Grads	Faculty	Staff
	UnderG	0	1	1	1
	Graduate		0	1	0
	Faculty			0	0
	Staff				0

**Values of 1 indicate the null hypothesis (H_1)*

[1]Undergraduate students

[2]Graduate students

8.5.6 Estimation of Emission Factors by Population Groups and Distance Zone

This section focuses on exploring who contributes the most vehicular emissions among the four specified groups commuting to Dalhousie city campuses, followed by exploring

the share of emissions attributable to each travel mode. In addition, the total amounts of emissions estimated for the defined geographic zones are presented. These estimations and analysis may help in generating feasible scenarios for university administrators that can be used for the emission reduction goal in short and long terms.

Table 8.7 outlines estimation results of emission factors by sample groups and distance zones. As can be seen, the staff group emitted the most (3707.23 CO_2 grams per person per day) among different commuter groups on Dalhousie city campuses, followed by the faculty group (3036.16 CO_2 grams per person per day). Furthermore, undergraduate students emitted more than graduate students. Staff members, who tend to live farther from Dalhousie city campuses, emitted nearly two and one-half times more than the graduate students, who live closer to Dalhousie city campuses. This result is consistent with the findings of previous university population based emission studies, which found that staff members emit more than three times as much as students (Mathez et al. 2013).

Table 8.7 Estimation of emission factors by population groups and distance zone

	Emission Factor (gram per person per day)							**Commuting Distance (km)**	
Sample group	**CO_2**	**CO**	**NO_x**	**PM_{10}**	**$PM_{2.5}$**	**THC**	**VOC**	**Average**	**Median**
Undergraduate students	2406.43	32.68	8.23	0.10	0.09	1.69	1.63	5.14	1.89
Graduate students	1512.69	20.54	5.17	0.06	0.05	1.06	1.03	8.27	2.36
Faculty	3036.16	41.24	10.38	0.12	0.11	2.13	2.06	7.67	3.18
Staff	3707.23	50.35	12.67	0.15	0.13	2.60	2.51	10.98	5.63
All groups	1953.38	26.53	6.68	0.08	0.07	1.37	1.32	8.02	2.77
Zone								**Primary Mode**	
INC	139.39	1.89	0.48	0.01	0.00	0.10	0.09	Walk	
SUB	1021.82	13.88	3.49	0.04	0.04	0.72	0.69	Transit	
ICB	2464.26	33.47	8.42	0.10	0.09	1.73	1.67	Auto Drive	
OCB	5397.24	73.30	18.45	0.22	0.19	3.79	3.66	Auto Drive	

Given that commuting distance and mode type directly influence vehicular emissions, results show that respondents commuting from the OCB to Dalhousie city campuses have the highest emissions, and those commuting from the ONC and INC have the lowest vehicle emissions. The estimated emission rates for all major air pollutants for respondents who live in the ICB (2464.26 CO_2 grams per person per day) zone (farther than 10 km from Dalhousie city campuses) is around 17 times greater than for those who live in the INC (139.39 CO_2 grams per person per day) zone (equal or less than 5 km from Dalhousie city campuses). Clearly, this reflects mode choices, since the great majority of OCB and ICB respondents have auto-drive as their primary mode, whereas active transportation (walk and bike) and transit are preferred modes in the ONC and SUB zone for students.

8.5.7 Estimation of Emission Factors by Scenario

Table 8.8 presents a summary of estimated emission factors for two defined scenarios. In Table 8.8 the base case represents the total amount of emissions that Dalhousie commuters produce on a typical weekday. All proposed scenarios are compared against the base scenario. Base case emission is estimated as 1953.38 CO_2 grams per person per day.

The emission factor variations by scenario are shown in Figure 8.4 and Figure 8.5.Scenario 1 (which assumes commuters switch their commuting mode into transit, where available) reduces the emissions of CO_2 for the minimum of 59.9% (assuming 25% switch to transit) and for the maximum of 71.4% (assuming all switch to transit). Scenario 1 emits around 784.23 and 668.98 CO_2 grams per person per day for 25% and 100% increase in transit ridership. Scenario 2, where we assume that commuters switch to auto-drive for campus commuting, has the worst outcome, increasing CO_2 emission factor by around 65.3%

(assuming all switch to auto drive) and 8.5% (assuming 25% switch to auto drive). Scenario 2 has an emission level of 3229.58 and 2118.81 CO_2 grams per person per day for 100% and 25% increases in auto drive.

Table 8.8 Estimation of emission factors by scenario

	Emission Factor (gram per person per day)						
	CO_2	**CO**	**NO_x**	**PM_{10}**	**$PM_{2.5}$**	**THC**	**VOC**
Base case	1953.38	26.53	6.68	0.08	0.07	1.37	1.32
Scenario 1							
25% increase in transit ridership	784.23	10.65	2.68	0.04	0.04	0.55	0.53
50% increase in transit ridership	745.82	10.13	2.55	0.03	0.03	0.52	0.51
75% increase in transit ridership	707.40	9.61	2.42	0.02	0.03	0.50	0.48
100% increase in transit ridership	668.98	9.09	2.29	0.02	0.02	0.47	0.45
Scenario 2							
25% increase in auto drive	2118.81	28.78	7.24	0.09	0.08	1.49	1.44
50% increase in auto drive	2759.51	37.48	9.43	0.11	0.10	1.94	1.87
75% increase in auto drive	2951.89	40.09	10.09	0.12	0.11	2.07	2.00
100% increase in auto drive	3229.58	43.86	11.04	0.13	0.11	2.27	2.19

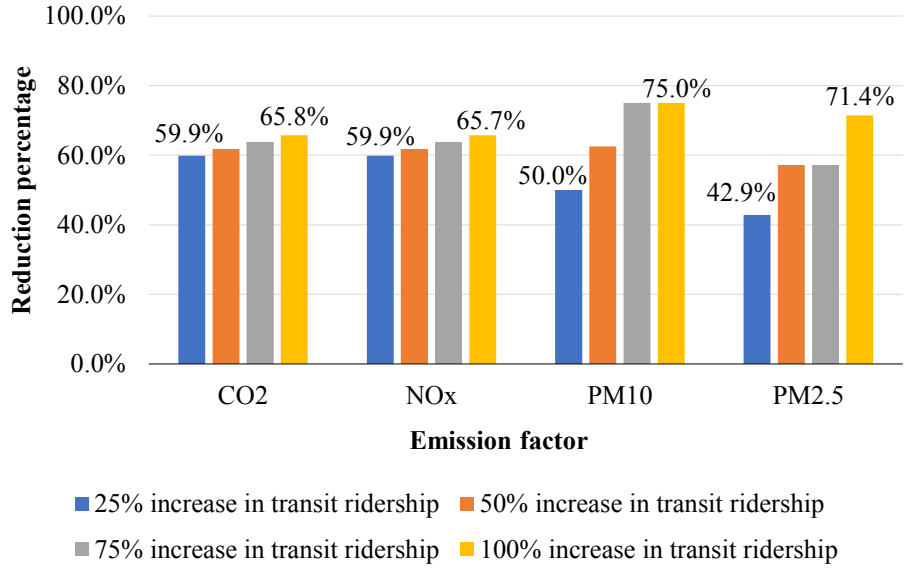


Figure 8.4 Emissions reduction (percentage) in scenario 1 from base case

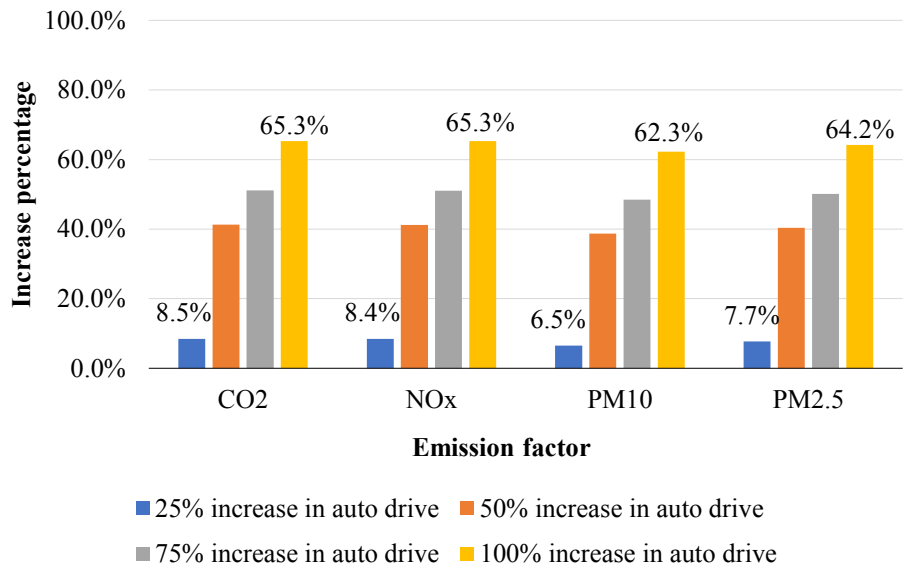


Figure 8.5 Emissions increase (percentage) in scenario 2 from base case

8.6 Conclusions

Commuters to large universities should be considered as a sizeable sub-population that needs special consideration in regional travel demand models. This study contributes by examining and modeling daily activity patterns (DAP) of sub-populations at the largest university in the Maritime Provinces of Canada. The data were drawn from the Environmentally Aware Commuter Travel Diary (EnACT) Survey undertaken in Dalhousie University, Canada. Activity sequencing, timing, and activity engagement of undergraduate students, graduate students, staff, and faculty members were investigated through a series of disaggregated statistical analyses and transition matrixes. The DAP patterns of these groups were modeled using the CART classifier algorithm. The model was trained with 75% of the dataset and the remaining 25% was used for testing the model performance.

Results show that university travel behavior deviates from the typical time-of-day distributions of travel. The empirical results from comparison of daily activity patterns among four different university populations reveal significant differences in activity travel behavior between student groups and workers (staff and faculty). For instance, undergraduate students were found to spend more time for shopping activity compared to other university groups. Workers often have an episode of physical activity (e.g. sport activity) before the start of work activity. Staff members are found to spend more time at school in their daily activity patterns in comparison with other university groups, usually starting work activity earlier than other university groups. Faculty members are found to participate in fewer leisure activities, whereas undergraduate students participate in higher numbers of recreation activities.

Consistent with previous studies, results reveal that the sequencing of home-school-home activities (H-S-H) and home-work-home activities (H-W-H) are the most frequent DAPs. Also, the sequence home-school/work-shopping-home activities is identified as the second most frequent DAP for all university community groups. However, each of the four sub-groups shows distinct differences when the probabilities of other sequences are considered.

This study provides valuable insights for policy discussions and transportation planning for university settings. The findings of commuting time and travel frequency patterns of university populations may help university and municipal authorities to develop practical policies to improve the traffic conditions on and near campuses, and to plan transit corridors to provide accessibility to university campuses. The pattern of student activity sequencing and travel timing is particularly important in this regard, since students

typically live close to campus and tend to make multiple trips to campus per day. This can have a substantial localized impact on the transit system, which is not well captured elsewhere. Furthermore, as our results revealed that the walking mode is the primary mode for student commuting to the university, the findings of this study may assist university administrators to improve sustainability through provision of a more pedestrian and bicycle friendly environment on and near campus. Details information of transit usage related to different university segments, including ridership frequency and travel time can be extracted from the DAP. Such these information can be used to improve transit level of service specially during morning and afternoon peak hours around university campuses.

To build on this study, we are proposing several avenues of research. Firstly, it is possible to investigate the association between land-use, built environment, and mode choice across the different university groups. Secondly, a population synthesis technique will be utilized as an alternative method to match the sample distribution to the overall university population, and to expand the sample size. Thirdly, the rich EnACT survey data that has been used for modeling in this study includes spatial information on residence locations and activity locations. Therefore, one potential extension would be development of a location choice model, to investigate the influence of spatial-temporal factors on locational choices for activities, by different university groups.

Finally, the approach and techniques employed in this study are transferable to study of DAPs and trip patterns at other large universities in North America, to develop sub-models for travel demand modeling. The approach may be applied more broadly, too, to other major sub-populations or special trip generators, such as those for large hospitals or large port areas.

Chapter 9 Conclusion

9.1 Summary

The research work presented in this dissertation focused on the development of the Scheduler for Activities, Locations, and Travel (SALT) disaggregated travel demand microsimulation model. The SALT modeling framework follows the concept of activity-based travel demand modeling techniques and theories. The work particularly concentrated on the development of the state-of-art machine learning micro-behavioral modules for modeling various aspects of activity-travel decisions, including activity selections and scheduling behaviors of population cohorts within the SALT modeling framework.

The SALT model system is designed based on the multi-layer hybrid machine learning techniques and is comprised of a series of behaviorally realistic advanced econometric, pattern recognition, and inference modules. Machine learning techniques provide the opportunity for more precise modeling and learning of preferences and behaviors in complex circumstances. These techniques incorporate an inter-related series of models with the aim to automatically learn to distinguish complicated patterns and make a creative decision based on the trained data.

The first stage of this research developed a novel pattern recognition model that identifies population clusters with homogeneous time-use activity patterns within the SALT model. Time-use activity patterns in each identified cluster can be modeled using a series of behaviorally realistic advanced econometric and machine learning micro-behavioral modules. The next stage of the research developed an agent-based ensemble model, which

infers type and frequency of activities in the schedule and their sequential arrangement, for modeling agenda formation within the SALT model. The next stage of research entailed the development of new agent-based inference model, which predicts temporal information associated with the traveler's daily activity schedule. Furthermore, a rule-based heuristic algorithm is developed to schedule the activities of individuals with varying characteristics and behavior for each cluster based on their priority importance and empirical guide information gained from the representative activity pattern in each cluster of the SALT model. The last stage of research entailed the development of a population synthesizer procedure to implement the SALT model for the entire region. The model is tested using a unique GPS-validated time-diary data set drawn from the large Halifax Space-Time Activity Research (STAR) household survey. In addition, this study modeled the daily time-use activity patterns and estimated emission factors for university commuters, considered as a special trip generator in regional travel demand models.

The advanced machine learning based micro-behavioral models utilized in the SALT model system are novel, time-efficient, and of practical use. A unique feature of the developed model that makes it different from other existing techniques is its degree of efficiency both in computational time and in minimizing exogenous errors. Furthermore, the proposed pattern recognition model enriches the traditional models, since it uses socio-demographic variables to classify the population and provide clusters based on identified time-use mobility patterns. Another advantage of the new proposed model is that, unlike previous approaches, the algorithm can recognize groups of people who typically tend to avoid travel in peak traffic periods. Furthermore, the inference model predicts both frequent and infrequent activities in the traveler's agenda. The implementation of the

scheduling model shows that the proposed model can accurately assemble the traveler's schedule with an average 82% accuracy in the 24-hour period.

In summary, the new micro-simulation modeling framework proposed in this study offers a straightforward and easy-to-implement tool for transport modelers to model time-use activity patterns for different population cohorts in the region. Moreover, the proposed modeling framework can be used to advance transportation demand management for different cohorts of the urban population as well as to analyze environmental mitigation and transport policy scenarios. The SALT modeling framework is being established for a medium-sized Canadian city and its travel-to-work area, but it can be modified and adapted to the modeling of urban transportation demands for major urban centers in North America.

9.2 Conclusions of Research Findings

The findings of this study provide deeper insights for modeling the travel behavior of population cohorts in transportation planning and management of urban transportation systems. Conclusions drawn from the outcomes of this study are summarized in the following.

9.2.1 Population Clusters with Homogeneous Time-Use Activity Patterns

- The Fuzzy C-Means (FCM) algorithm recognized pattern complexity of activity sequences in the dataset that resulted in the identification of distinct clusters for out-of-home workers, non-workers and non-students, students, and individuals who mostly spend their time at home.

- Each cluster contained homogeneous daily activity patterns and generated crucial information of activities, such as type, sequential arrangement, start time, and duration probability distributions.
- Both in-home and out-of-home time-use activity patterns of individuals were modeled. Each individual's daily pattern of activity was transformed into 288 three dimensional five-minute intervals. The first dimension contains temporal information on activities, the second dimension contains socio-demographic characteristics associated to activities, and the third dimension comprises spatial information related with activities.
- Using the subtractive clustering algorithm instead of randomized cluster number and cluster centroids can increase the accuracy of cluster identification.
- The FCM algorithm, unlike previous methods, can recognize and derive clusters of people who typically tend to avoid travel in peak traffic periods.
- The FCM clustering algorithm resulted in a superior convergence of the local minima of the squared error principle, compared to other potential clustering algorithms.

9.2.2 Representative Set of Activity Patterns

- Using the progressive alignment method instead of the sequence alignment method can increase the accuracy of identification of representative activity patterns.
- The progressive alignment method improved the model accuracy through iterative profile-alignment of tree portions to maximize the sum of pairs score.

- The Classification and Regression Tree (CART) classifier algorithm characterizes the cluster memberships through inter-dependencies among their socio-demographic attributes.
- Using the CART classifier algorithm, compared to other decision tree algorithms, can increase the performance of the decision tree by adding an additional cross-validation step in the model structure.
- Cluster memberships can be identified with better accuracy by developing a CART prediction tree based on the socio-demographic features of individuals.

9.2.3 Activity Engagement Patterns of Population Groups

- The Random Forest (RF) model predicted activity sequences and agenda of the entire population with 70% and 78% accuracy, respectively. Furthermore, the inference model predicts both frequent and infrequent activities in the traveler's agenda.
- Comparison between the observed and replicated patterns through activity episode transitions matrixes showed a mean absolute error of only 7.3%, and revealed that the RF models could successfully replicate episodes and position in agenda.
- The RF model replicated activity sequences of more than 70% of the population, with the mean distance equal to 0.47. Additionally, both in-home and out-of-home activity patterns of population were modeled.
- Using the RF model, compared to other alternative machine learning algorithms such as support vector machine and back-propagation neural network, can decrease the generalization error. Furthermore, the RF model is less likely to overfit.

9.2.4 Activity Timing and Building The 24-Hour Activity Schedule

- The activity start time is predicted with 60.1% accuracy for eight bins with 180 minutes duration, and 36.3% accuracy for 48 bins with 30 minutes duration.
- The activity duration is predicted with 98.6% accuracy for four bins with 360 minutes duration, and 67.3% accuracy for 24 bins with 60 minutes duration.
- Using the RF model, compared to alternative algorithms such as AdaBoost and CHAID, can increase the prediction accuracy of the activity temporal attributes with a smaller duration interval.
- The scheduling model was able to assemble the traveler's schedule with an average 82% accuracy in the 24-hour period.

9.2.5 Baseline Synthetic Population for the Region

- A baseline synthetic population was generated for Halifax, Nova Scotia, and for the university community groups.
- The population synthesizer algorithm performed well in terms of the computational time and distribution of seed data presented in the final synthesized population list.
- Synthetic populations were generated both at the regional and dissemination area (DA) levels with reasonable computational time and accuracy.
- The population synthesizer accurately represented the seed data (observed population) in the 100% synthetic population within the acceptable range of error for the designated study area

9.2.6 Travel Behavior of University Commuters as a Special Trip Generator in Regional Travel Demand Models

- Results showed that there are noteworthy differences in activity travel behavior between different university population segments, and that these in general deviate from the typical time-of-day distributions of travel.
- Shopping activity was found to be the most frequent out-of-home activity after school/work activity for all four university segments.
- The most used transport mode for commuting to and from university campuses was found to be the walking mode.
- The results of emission analysis revealed that the staff group emitted the highest pollutants compared to other university segments. Furthermore, on average, respondents who live farther than 10 km from university campuses emitted around 17 times more than those who live equal or less than 5 km from their campus/workplace.

9.3 Model Implementation

The prototype version of the SALT system enables component-based modeling and modular design for travel demand modelers. It offers the opportunity to fragment the model into innovative modules, and then model, simulate, and validate each module independently. Each of these can be developed and saved independently in the overall SALT modeling framework (Daisy 2018b). The SALT system outcome, at each time-stem of the simulation, can be stored at the individual-agent or object-level. This feature of the

SALT system allows multiple transport modelers to work in parallel on different components.

Before the implementation of the SALT system in real transportation planning policy analysis, a prototype model need to go through several phases. Model monitoring and validation should be integrated into all phases of the model cycle. Further disaggregate and aggregate spatio-temporal model verification and validation are essential for key system elements, including activity engagement and trip making. While the SALT system can be integrated with dynamic traffic assignment and network topology, and applied to infer regional activity patterns of individuals, additional assessment of the population synthesizer module, related to aspects such as the homogeneity issue and network topology, should be taken into account. Lastly, further analysis should be performed to assess the capability of the SALT system to forecast future year circumstances. This can be done by employing the backcasting technique to model for a period for which accurate empirical data at both start and finish years are available.

9.4 Recommendations for Future Work

The following recommendations provide suggestions for further research that would complement the work presented herein and gain a better understanding of the effect of machine learning techniques on activity-based travel demand forecasting models.

- The large Halifax STAR household survey data that has been used for constructing the SALT model in this study includes business hours survey data. Accordingly, future extension of the research would seek to incorporate operations hours in the modeling process.

- Newer activity-based models include interaction between household members, as these have a significant impact on others' travel. Therefore, another potential extension would be to incorporate joint activities and the interactions between activity schedules of household members into the SALT model.
- A more robust effort should be taken to determine how well the predicted changes in travel over time in the scheduling engine of the SALT model compare to actual changes in travel over time.
- For some of the population clusters that include more complex activity sequences, an alternative approach for tuning hyperparameters in the random forest, such as Bayesian optimization, might provide better outcomes.
- A straightforward mixed-integer linear programming approach can be incorporated into the decision rule-based algorithm to improve the conflict resolution between inserted activities.
- As explained, several new machine learning techniques, not previously explored in travel behavior analysis, are employed in the development of the different modules in the SALT model system. Therefore, one avenue to extend this research is to compare the corresponding reproducibility and computational time between alternative techniques and proposed approaches in this study for the same task.
- The advanced machine learning based techniques developed in this dissertation can also be adapted for modeling other components of activity-based travel demand models, such as work and residential location choice models, and transport mode. Future work could examine the application of such techniques in modeling the above-mentioned elements of activity-based travel demand models.

- The concept of self-driving and/or autonomous vehicles and ride-sharing services such as Uber and Lyft can be added into the SALT model. Given their potentially rapid adoption, the impact of these new technologies and ride-sharing services on daily activity travel patterns should be considered in further extensions of the current study.

References

Akar, G., C. Flynn and M. Namgung. (2012). "Travel choices and links to transportation demand management: Case study at Ohio State University". Transportation Research Record: Journal of the Transportation Research Board 2319, pp. 77-85.

Allahviranloo, M. (2016). "Pattern recognition and personal travel behavior". Presented at the 95th Annual Meeting of the Transportation Research Board (TRB), Washington, D.C., USA.

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

Allahviranloo, M., R. Regue, and W. Recker. (2016). "Modeling the activity profiles of a population". Transportmetrica B: Transport Dynamics, pp. 1-24.

Anas, A. (1994). "METROSIM: A unified economic model of transportation and land-use". Alex Anas and Associates, Williamsville, NY.

Anderson, P., B. Farooq, D. Efthymiou, and M. Bierlaire. (2014). "Associations generation in synthetic population for transportation applications graph-theoretic solution". Transportation Research Record: Journal of the Transportation Research Board 2429, pp. 38-50.

Arentze, T. A. and H. J. P. Timmermans. (2009). "A need-based model of multi-day, multi-person activity generation". Transportation Research Part B: Methodological. 43(2), pp. 251-265.

Arentze, T. and H. Timmermans. (2000). "ALBATROSS: A learning based transportation oriented simulation system". European Institute of Retailing and Services Studies (EIRASS), Technische Universiteit Eindhoven, Netherlands.

Arentze, T. and H. Timmermans. (2007). "Parametric action decision trees: Incorporating continuous attribute variables into rule-based models of discrete choice". Transportation Research Part B: Methodological. 41(7), pp. 772-783.

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

Auld, J. and A. Mohammadian. (2009). "Framework for the development of the agent-based dynamic activity planning and travel scheduling (ADAPTS) model". Transportation Letters. 1(3), pp. 245-255.

Auld, J., A. Mohammadian, and S. T. Doherty. (2009). "Modeling activity conflict resolution strategies using scheduling process data". Transportation Research Part A: Policy and Practice. 43(4), pp. 386-400.

Auld, J., M. Hope, H. Ley, V. Sokolov, B. Xu, and K. Zhang. (2016). "POLARIS: Agent-based modeling framework development and implementation for integrated travel demand and network and operations simulations". Transportation Research Part C: Emerging Technologies. 64, pp. 101-116.

Axhausen, K. W. (1996). "The design of environmentally aware travel diaries". Transportation Planning and Technology. 19, pp. 275-290.

Ballas, D., G. Clarke, D. Dorling, and D. Rossiter. (2007). "Using SimBritain to model the geographical impact of national government policies". Geographical Analysis. 39(1), pp. 44-77.

Balmer, M., K. Meister, M. Rieser, K. W. Axhausen. (2008). "Agent-based simulation of travel demand: Structure and computational performance of MATSim". ETH, Eidgenossische Technische Hochschule Zurich, IVT Institut fur Verkehrsplanung und Transportsysteme.

Balsas, C. J. L. (2003). "Sustainable transportation planning on college campuses". Transport Policy 10(1): 35-49.

Bao, Q., B. Kochan, Y. Shen, T. Bellemans, D. Janssens, and G. Wets. (2016). "Activity-Based travel demand modeling framework FEATHERS". Transportation Research Record: Journal of the Transportation Research Board 2564, pp. 89-99.

Bar-Gera, H., K. C. Konduri, B. Sana, X. Ye, and R.M. Pendyala. (2009)."Estimating survey weights with multiple constraints using entropy optimization methods". Presented at the 88th Annual Meeting of the Transportation Research Board. Washington, DC.

Barthelemy, J., T. Suesse, M. Namazi-Rad and M. J. Barthelemy. (2015). "MIPFP: Multidimensional iterative proportional fitting and alternative models". https://cran.r-project.org/web/packages/mipfp/index.html.

Bates, J. (2007). "History of demand modeling". Handbook of Transport Modeling. 1, pp. 11-34.

Beckman, R. J., K. A. Baggerly, and M. D. McKay. (1996). "Creating synthetic baseline populations". Transportation Research Part A: Policy and Practice. 30(6), pp. 415-429.

Ben-Akiva, M. and J. L. Bowman. (1998a). "Integration of an activity-based model system and a residential location model". Urban Studies. 35(7), pp. 1131-1153.

Ben-Akiva, M. E. and J. L. Bowman. (1998b). "Activity based travel demand model systems". Equilibrium and Advanced Transportation Modeling, Springer US: Boston, MA. p. 27-46.

Bhat, C. R. and F. S. Koppelman. (1999). "Activity-based modeling of travel demand". Handbook of Transportation Science, Springer US: Boston, MA. p. 35-61.

Bhat, C. R., K. G. Goulias, R. M. Pendyala, R. Paleti, R. Sidharthan, L. Schmitt, and H. H. Hu. (2013). "A household-level activity pattern generation model with an application for Southern California". Transportation. 40(5), pp. 1063-1086.

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

Biau, G. and E. Scornet. (2016). "A random forest guided tour". TEST. 25(2), pp. 197-227.

Bishop, C. (2007). "Pattern recognition and machine learning". Springer, New York.

Black, J., C. Mason, and K. Stanley. (1999). "Travel demand management: Policy context and an application by the University of New South Wales (UNSW) as a large trip generator". Transport Engineering in Australia. 5(2), pp. 1-11.

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

Boyce, D. E. and H. C. W. L. Williams. (2015). "Forecasting urban travel: Past, present and future". Edward Elgar Pub. Limited.

Breiman, L. (2001). "Random forests". Machine Learning. 45(1), pp. 5-32.

Castiglione, J., M. Bradley and J. Gliebe. (2015). "Activity-based travel demand models: A primer". TRB's second Strategic Highway Research Program (SHRP 2) Report S2-C46-RR-1.

Chen, X. (2012). "Statistical and activity-based modeling of university student travel behavior". Transportation Planning and Technology 35(5): 591-610.

Chenna, R., H. Sugawara, T. Koike, R. Lopez, T. J. Gibson, D. G. Higgins, and J. D. Thompson. (2003). "Multiple sequence alignment with the clustal series of programs". Nucleic Acids Research. 31(13), pp. 3497-3500.

Christakos, G. (2000). "Modern spatiotemporal geostatistics". Oxford University Press, New York.

Clarke, M. (1986). "Activity modeling-a research tool or a practical planning technique". Behavioral Research for Transport Policy, pp. 3-15.

Daisy, N. S. (2018a). "Modeling activity-travel behavior for activity-based travel demand modeling". Doctoral Research in Transport Modeling. Presented at the 97th Annual Meeting of Transportation Research Board (TRB), Washington, D.C., USA.

Daisy, N. S. (2018b). "Microsimulation of activity participation, tour complexity, and mode choice within an activity-based travel demand model system". PhD Dissertation. Department of Civil and Resource Engineering, Dalhousie University.

Daisy, N. S., L. Liu, and H. Millward. (2017a). "Analyzing tours: Application of a traveler grouping based cluster analysis". Presented at the 53rd Canadian Transportation Research Forum (CTRF). Ottawa, Canada.

Daisy, N. S., L. Liu, and H. Millward. (2017b). "Optimizing daily travel sequences and time-use patterns of individuals". Presented at the 53rd Canadian Transportation Research Forum (CTRF). Ottawa, Canada.

Daisy, N. S., H. Millward, M. H. Hafezi, and L. Liu. (2017). "Trip-chaining and tour complexity: contrasts between worker and non-worker groups in Halifax, Nova Scotia". Presented at the 28th Atlantic Division of the Canadian Association of Geographers (ACAG/ACGA) Conference: Saint Mary's University, Halifax, Nova Scotia.

Daisy, N. S., M. H. Hafezi, L. Liu, and H. Millward. (2018a). "Understanding and modeling the activity-travel behavior of university commuters at a large Canadian university". Journal of Urban Planning and Development. 144(2).

Daisy, N. S., M. H. Hafezi, L. Liu, and H. Millward. (2018b). "Housing location and commuting mode choices of university students and employees: An application of bivariate Probit models". Peer reviewed ASCE proceedings of the International Conference on Transportation and Development (ICTD). Pittsburgh, Pennsylvania, USA.

De Palma, A., R. Lindsey, E. Quinet, and R. Vickerman. (2011). "A handbook of transport economics". Edward Elgar.

Deville, J. C., C. E. Sarndal, and O. Sautory. (1993). "Generalized raking procedures in survey sampling". Journal of the American Statistical Association. 88(423), pp. 1013-1020.

Dong, X., M.E. Ben-Akiva, J.L. Bowman, and J.L. Walker. (2006). "Moving from trip-based to activity-based measures of accessibility". Transportation Research Part A: Policy and Practice. 40(2), pp. 163-180.

Edwards, K.L. and G.P. Clarke. (2009). "The design and validation of a spatial microsimulation model of obesogenic environments for children in Leeds, UK: SimObesity". Social science and medicine. 69(7), pp. 1127-1134.

Ellegard, K. (1999). "A time-geographical approach to the study of everyday life of individuals-A challenge of complexity". GeoJournal. 48(3), pp. 167-175.

Ellegard, K. and B. Vilhelmson. (2004). "Home as a pocket of local order: Everyday activities and the friction of distance". Geografiska Annaler: Series B, Human Geography. 86(4), pp. 281-296.

Ellegard, K. and U. Svedin. (2012). "Torsten Hagerstrand's time-geography as the cradle of the activity approach in transport geography". Journal of Transport Geography. 23, pp. 17-25.

ENR. (2016). Environment and natural resources - Local weather forecasts. Retrieved on July 4, 2016 from https://weather.gc.ca/city/pages/ns-19_metric_e.html.

Eom, J. K., J. R. Stone, and S. K. Ghosh. (2009). "Daily activity patterns of university students". Journal of Urban Planning and Development. 135(4), pp. 141-149.

Erisoglu, M., N. Calis, and S. Sakallioglu. (2011). "A new algorithm for initial cluster centers in k-means algorithm". Pattern Recognition Letters. 32(14), pp. 1701-1705.

Estevao, V. M. and C. E. Sarndal. (2006). "Survey estimates by calibration on complex auxiliary information". International Statistical Review 74(2), pp. 127-147.

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

Ettema, D., A. Borgers, and H. Timmermans. (1996). "SMASH (simulation model of activity scheduling heuristics): some simulations". Transportation Research Record: Journal of the Transportation Research Board 1551, pp. 88-94.

Fosgerau, M. (2002). "PETRA-An activity-based approach to travel demand analysis". National Transport Models: Recent Developments and Prospects, Springer Berlin Heidelberg: Berlin, Heidelberg. pp. 134-145.

Garling, T., J. K. Brannas, R. Garvill, G. Golledge, S. Gopal, E. Holm and E. Lindberg. (1989). "Household activity scheduling. Transport policy, management and technology towards 2001". Selected proceedings of the fifth world conference on transport research.

Garling, T., M. P. Kwan, and R. G. Golledge. (1994). "Computational-process modeling of household activity scheduling". Transportation Research Part B: Methodological. 28(5), pp. 355-364.

Goldner, W. (1971). "The Lowry model heritage". Journal of the American Institute of Planners. 37(2), pp. 100-110.

Goran, J. (2001). "Activity based travel demand modeling-a literature study". Technical report, Danmarks Transport-Forskning.

Goulias, K. G. (1999). "Longitudinal analysis of activity and travel pattern dynamics using generalized mixed Markov latent class models". Transportation Research Part B: Methodological. 33(8), pp. 535-558.

Guo, J. Y. and C. R. Bhat. (2007). "Population synthesis for microsimulating travel behavior". Transportation Research Record: Journal of the Transportation Research Board 2014, pp. 92-101.

Hafezi, M. H. (2018). "Modeling representative time-use behavior for activity-based travel demand modeling". Doctoral Research in Transport Modeling. Presented at the 97th Annual Meeting of Transportation Research Board (TRB), 2018, Washington, D.C., USA.

Hafezi, M. H., H. Millward, and L. Liu. (2018a). "Activity-based travel demand modeling: Progress and possibilities". Peer reviewed ASCE proceedings of the International Conference on Transportation and Development (ICTD). Pittsburgh, Pennsylvania, USA.

Hafezi, M. H., H. Millward, and L. Liu. (2018b). "Inferring activity selection and scheduling behavior of population cohorts for travel demand modeling". Presented at the 53rd Canadian Transportation Research Forum (CTRF). Ottawa, Canada.

Hafezi, M. H., H. Millward, N. S. Daisy, and L. Liu. (2017)."How people schedule their daily activities? Evidence from a mid-sized Canadian city". Presented at the 28th Atlantic Division of the Canadian Association of Geographers (ACAG/ACGA) Conference. Saint Mary's University, Halifax, Nova Scotia.

Hafezi, M. H., L. Liu, and H. Millward. (2017a). "Modeling activity scheduling behavior of individuals for travel demand models". Presented at the 52nd Canadian Transportation Research Forum (CTRF). Winnipeg, Manitoba.

Hafezi, M. H., L. Liu, and H. Millward. (2017b). "Identification of representative patterns of time use activity through fuzzy c-means clustering". Transportation Research Record: Journal of the Transportation Research Board 2668, pp. 38-50.

Hafezi, M. H., L. Liu, and H. Millward. (2017c). "A time-use activity-pattern recognition model for activity-based travel demand modeling". Transportation, pp. 1-26.

Hafezi, M. H., L. Liu, and H. Millward. (2018a). "Learning daily activity sequences of population groups using random forest theory". Transportation Research Record: Journal of the Transportation Research Board.

Hafezi, M. H., L. Liu, and H. Millward. (2018b). "Modeling activity scheduling behavior of travelers for activity-based travel demand model". Presented at the 97th Annual Meeting of the Transportation Research Board (TRB), Washington, D.C.

Hafezi, M. H., N. S. Daisy, H. Millward, and L. Liu. (2018a). "Commuting to campus: Findings from the Dalhousie EnACT travel survey". Department of Civil and Resource Engineering, Dalhousie University. Halifax, Canada.

Hafezi, M. H., N. S. Daisy, H. Millward, and L. Liu. (2018b)."Emissions analysis for synthetic baseline population of a large Canadian university". Presented at the 97th Annual Meeting of the Transportation Research Board (TRB), Washington, D.C.

Hafezi, M. H., N. S. Daisy, L. Liu, and H. Millward. (2017)."Daily time-use activity patterns at a large Canadian university". Presented at the 96th Annual Meeting of the Transportation Research Board (TRB), Washington, D.C.

Hafezi, M. H., N. S. Daisy, L. Liu, and H. Millward. (2018). "Daily activity and travel sequences of students, faculty, and staff at a large Canadian university". Transportation Planning and Technology.

Hagerstrand, T. (1970). "What about people in regional science?". Papers in Regional Science. 24(1), pp. 7-24.

Hand, D. J., H. Mannila, and P. Smyth. (2001). "Principles of data mining". Cambridge: MIT Press.

Hayes-Roth, B. and F. Hayes-Roth. (1979). "A cognitive model of planning". Cognitive Science. 3(4), pp. 275-310.

Hensher, D. and T. Ton. (2002). "TRESIS: A transportation, land use and environmental strategy impact simulator for urban areas." Transportation 29(4): 439-457.

Hermes, K. and M. Poulsen. (2012). "A review of current methods to generate synthetic spatial microdata using reweighting and future directions". Computers, Environment and Urban Systems. 36(4), pp. 281-290.

Heung, V. C. S. and J. S. L. Leong. (2006). "Travel demand and behavior of university students in Hong Kong." Asia Pacific Journal of Tourism Research 11(1): 81-96.

Hildebrand, E. D. (2003). "Dimensions in elderly travel behavior: A simplified activity-based model using lifestyle clusters". Transportation. 30(3), pp. 285-306.

Ho, C. and C. Mulley. (2013). "Tour-based mode choice of joint household travel patterns on weekend and weekday". Transportation. 40(4), pp. 789-811.

HRM. (2016). "Halifax Transit moving forward together plan". Halifax Regional Municipality. URL: https://www.halifax.ca/sites/default/files/documents/transportation/halifax-transit/MFTP_PlanOnly.pdf.

Huang, Z. and P. Williamson. (2001). "A comparison of synthetic reconstruction and combinatorial optimization approaches to the creation of small-area microdata". Population Microdata Unit, Department of Geography, University of Liverpool, Liverpool L69 3BX. URL http://pcwww.liv.ac.uk/~william/microdata.

Hunt, J. D., D. S. Kriger, and E.J. Miller. (2005). "Current operational urban land-use transport modeling frameworks: A review". Transport Reviews. 25(3), pp. 329-376.

Jang, Y., Y. C. Chiu, and H. Zheng. (2013). "Modeling within-day activity rescheduling decisions under time-varying network conditions". Advances in Dynamic Network Modeling in Complex Transportation Systems, Springer New York: New York, NY. p. 225-244.

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

Jiang, S., J. Ferreira, and M. C. Gonzalez. (2017). "Activity-based human mobility patterns inferred from mobile phone data: A case study of Singapore". IEEE Transactions on Big Data. 3(2), pp. 208-219.

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

John Lu, Z. (2010). "The elements of statistical learning: Data mining, inference, and prediction". Journal of the Royal Statistical Society. 173(3), pp. 693-694.

Jones, P. M., M. C. Dix, M. I. Clarke, and I. G. Heggie. (1983). "Understanding travel behavior". Journal of Forecasting. 4, pp. 315-316.

Jordan, R., M. Birkin, and A. Evans. (2014). "An agent-based model of residential mobility: Assessing the impacts of urban regeneration policy in the EASEL district". Computers, Environment and Urban Systems. 48(0), pp. 49-63.

Jovicic, G. (2001). "Activity based travel demand modeling-A literature study". Danmarks TransportForskning. p. 64.

Kamruzzaman, M., J. Hine, B. Gunay and N. Blair. (2011). "Using GIS to visualize and evaluate student travel behavior." Journal of Transport Geography 19(1): 13-32.

Khattak, A., X. Wang, S. Son and P. Agnello. (2011). "University student travel in Virginia: Is it different from the general population." Transportation Research Record: Journal of the Transportation Research Board 2255, pp. 137-145.

Kitamura, R. (1988). "An evaluation of activity-based travel analysis". Transportation. 15(1), pp. 9-34.

Kitamura, R., C. Chen, and R. Pendyala. (1997). "Generation of synthetic daily activity-travel patterns". Transportation Research Record: Journal of the Transportation Research Board 1607, pp. 154-163.

Kitamura, R., C. Chen, R. M. Pendyala, and R. Narayanan. (2000). "Micro-simulation of daily activity-travel patterns for travel demand forecasting". Transportation. 27(1), pp. 25-51.

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

Koupal, J., H. Michaels, M. Cumberworth, C. Bailey, and D. Brzezinski. (2002)."EPA's plan for MOVES: A comprehensive mobile source emissions model". Proceedings of the 12th CRC On-Road Vehicle Emissions Workshop, San Diego, CA.

Krizek, K.J. (2003). "Neighborhood services, trip purpose, and tour-based travel". Transportation. 30(4), pp. 387-410.

Kubat, M. (2015). "An introduction to machine learning". Springer International Publishing.

Kwan, M. P. (1997). "GISICAS: An activity-based travel decision support system using a GIS-interfaced computational-process model". Activity-based approaches to travel analysis, pp. 263-282.

Kwan, M. P. and R. G. Golledge. (1997). "Computational process modeling of disaggregate travel behavior". Recent Developments in Spatial Analysis: Spatial Statistics, Behavioral Modeling, and Computational Intelligence, Springer Berlin Heidelberg: Berlin, Heidelberg. pp. 171-185.

Kyriakidis, P. C. (2004). "A geostatistical framework for area-to-point spatial interpolation." Geographical Analysis 36(3): 259-289.

Lee, D. H. and Y. Fu. (2011). "Cross-entropy optimization model for population synthesis in activity-based microsimulation models." Transportation Research Record: Journal of the Transportation Research Board 2255, pp. 20-27.

Lemaitre, G. and J. Dufour. (1987). "An integrated method for weighting persons and families". Survey Methodology. 13(2), pp. 199-207.

Leszczyc, P. T. L. P. and H. Timmermans. (2002). "Unconditional and conditional competing risk models of activity duration and activity sequencing decisions: An empirical comparison". Journal of Geographical Systems. 4(2), pp. 157-170.

Li, S. and D. H. Lee. (2017). "Learning daily activity patterns with probabilistic grammars". Transportation. 44(1), pp. 49-68.

Li, Z. C., W. H. K. Lam, and S. C. Wong. (2014). "Bottleneck model revisited: An activity-based perspective". Transportation Research Part B: Methodological. 68, pp. 262-287.

Liao, F., T. Arentze, and H. Timmermans. (2013). "Incorporating space-time constraints and activity-travel time profiles in a multi-state supernetwork approach to individual activity-travel scheduling". Transportation Research Part B: Methodological. 55, pp. 41-58.

Liao, L., D. J. Patterson, D. Fox, and H. Kautz. (2007). "Learning and inferring transportation routines". Artificial Intelligence. 171(5), pp. 311-331.

Lim, P. P. and D. Gargett. (2013)."Population Synthesis for Travel Demand Forecasting". Presented at the 36th Australasian Transport Research Forum (ATRF), Brisbane, Queensland, Australia.

Limanond, T., T. Butsingkorn and C. Chermkhunthod. (2011). "Travel behavior of university students who live on campus: A case study of a rural university in Asia". Transport Policy 18(1): 163-171.

Liu, F., D. Janssens, J. Cui, G. Wets, and M. Cools. (2015). "Characterizing activity sequences using profile Hidden Markov Models". Expert Systems with Applications. 42(13), pp. 5705-5722.

Liu, L., M. H. Hafezi, N. Daisy. (2016). "Results of the 2016 Dalhousie Environmentally Aware Travel Diary (EnACT) Survey". Department of Civil and Resource Engineering, Dalhousie University. Halifax, Canada.

Lockwood, A., S. Srinivasan, and C. Bhat. (2005). "Exploratory analysis of weekend activity patterns in the San Francisco Bay Area, California". Transportation Research Record: Journal of the Transportation Research Board 1926, pp. 70-78.

Long, Y. and Z. Shen. (2013). "Disaggregating heterogeneous agent attributes and location". Computers, Environment and Urban Systems. 42, pp. 14-25.

Lovelace, R., M. Birkin, D. Ballas and E. van Leeuwen. (2015). "Evaluating the performance of iterative proportional fitting for spatial microsimulation: New tests for an established technique". Journal of Artificial Societies and Social Simulation 18(2): 21.

Lu, J. (2010). "The elements of statistical learning: data mining, inference, and prediction". Journal of the Royal Statistical Society 173(3): 693-694.

Ma, L. (2011). "Generating disaggregate population characteristics for input to travel-demand models". PhD Dissertation. University of Florida.

Malouf, R. (2002). "A comparison of algorithms for maximum entropy parameter estimation". Proceedings of the 6th conference on Natural language learning-Volume 20. Association for Computational Linguistics.

Marcotte, P. and S. Nguyen. (2013). "Equilibrium and advanced transportation modeling". Springer US.

Mathez, A., K. Manaugh, V. Chakour, A. El-Geneidy, and M. Hatzopoulou. (2013). "How can we alter our carbon footprint? Estimating GHG emissions based on travel survey information". Transportation. 40(1), pp. 131-149.

McFadden, D. (1980). "Econometric models for probabilistic choice among products". The Journal of Business. 53(3), pp. S13-S29.

McNally, M. G. (2007). "The four step model". Handbook of transport modeling. 1, pp. 35-41.

Miller, E. and M. Roorda. (2003). "Prototype model of household activity-travel scheduling". Transportation Research Record: Journal of the Transportation Research Board 1831, pp. 114-121.

Millward, H. and J. Spinney. (2011). "Time use, travel behavior, and the rural–urban continuum: results from the Halifax STAR project". Journal of Transport Geography. 19(1), pp. 51-58.

Mitloehner, F. (2016). "Quantification of the Emission Reduction Benefits of Mitigation Strategies for Dairy Silage". California Air Resources Board.

Muller, K. and K. W. Axhausen. (2010). "Population synthesis for microsimulation: State of the art". ETH Zurich, Institut fur Verkehrsplanung, Transporttechnik, Strassen-und Eisenbahnbau (IVT).

Muller, K. and K. W. Axhausen. (2011). "Hierarchical IPF: Generating a synthetic population for Switzerland". Eidgenossische Technische Hochschule Zurich, IVT.

Nagle, N. N., B. P. Buttenfield, S. Leyk and S. Spielman. (2013). "Dasymetric modeling and uncertainty". Annals of the Association of American Geographers 104(1): 80-95.

Nakamura, H., Y. Hayashi, and K. Miyamoto. (1983). "A land use-transport model for metropolitan areas". Papers in Regional Science. 51(1), pp. 43-63.

Needleman, S.B. and C.D. Wunsch. (1970). "A general method applicable to the search for similarities in the amino acid sequence of two proteins". Journal of Molecular Biology. 48(3), pp. 443-453.

Ngo, L. T. and B. H. Pham. (2012). "A type-2 fuzzy subtractive clustering algorithm". Mechanical Engineering and Technology, Springer Berlin Heidelberg: Berlin, Heidelberg. p. 395-402.

NRN. (2016). GeoBase - National Road Network (NRN) - NS, Nova Scotia. Retrieved on May 5, 2016 from http://geogratis.gc.ca/api/en/nrcan-rncan/ess-sst/28fd9d45-5680-4dc7-b016-9a56bbab88eb.html.

Oberkampf, W. L., S. M. DeLand, B. M. Rutherford, K. V. Diegert, and K. F. Alvin. (2002). "Error and uncertainty in modeling and simulation". Reliability Engineering and System Safety. 75(3), pp. 333-357.

Oketch, T. and M. Carrick. (2005)."Calibration and validation of a micro-simulation model in network analysis". Presented at the 84th Annual Meeting of the Transportation Research Board (TRB), Washington, D.C.

Oppenheim, N. (1995). "Urban travel demand modeling: From Individual choices to general equilibrium". Wiley.

Ortuzar, J. D. D. and L. G. Willumsen. (2011). "Modeling transport". Wiley.

Outwater, M. L. and B. Charlton. (2006). "The San Francisco Model in practice". Innovations in Travel Demand Modeling, pp. 24.

Pena, J. M., J. A. Lozano, and P. Larranaga. (1999). "An empirical comparison of four initialization methods for the K-Means algorithm". Pattern Recognition Letters. 20(10), pp. 1027-1040.

Popovich, N. (2014). "Results of the 2013-14 campus travel survey". Institute of Transportation Studies, University of California, Davis.

Pritchard, D. R. and E. J. Miller. (2009)."Advances in agent population synthesis and application in an integrated land use and transportation model". Presented at the 88th Annual Meeting of the Transportation Research Board (TRB), Washington, D.C.

Rasouli, S. and H. Timmermans. (2012). "Uncertainty in travel demand forecasting models: literature review and research agenda". Transportation Letters. 4(1), pp. 55-73.

Rasouli, S. and H. Timmermans. (2014). "Activity-based models of travel demand: promises, progress and prospects". International Journal of Urban Sciences. 18(1), pp. 31-60.

Recker, W. W. (2001). "A bridge between travel demand modeling and activity-based travel analysis". Transportation Research Part B: Methodological. 35(5), pp. 481-506.

Recker, W. W., M. G. McNally, and G. S. Root. (1986a). "A model of complex travel behavior: Part I-Theoretical development". Transportation Research Part A: General. 20(4), pp. 307-318.

Recker, W. W., M. G. McNally, and G. S. Root. (1986b). "A model of complex travel behavior: Part II-An operational model". Transportation Research Part A: General. 20(4), pp. 319-330.

Rissel, C., C. Mulley and D. Ding. (2013). "Travel mode and physical activity at Sydney university". International Journal of Environmental Research and Public Health 10(8): 3563-3577.

Rodriguez, D. A. and J. Joo. (2004). "The relationship between non-motorized mode choice and the local physical environment". Transportation Research Part D: Transport and Environment 9(2): 151-173.

Roorda, M. J., E. J. Miller and K. M. N. Habib. (2008). "Validation of TASHA: A 24-h activity scheduling microsimulation model". Transportation Research Part A: Policy and Practice 42(2): 360-375.

Ruther, M., G. Maclaurin, S. Leyk, B. Buttenfield, and N. Nagle. (2013). "Validation of spatially allocated small area estimates for 1880 Census demography". Demographic Research. 29(22), pp. 579-616.

Salvini, P. and E. J. Miller. (2005). "ILUTE: An operational prototype of a comprehensive microsimulation model of urban systems". Networks and Spatial Economics 5(2): 217-234.

Schroeder, J. P. (2007). "Target-density weighting interpolation and uncertainty evaluation for temporal analysis of census data". Geographical Analysis. 39(3), pp. 311-335.

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

Shafique, M. A. and E. Hato. (2015). "Use of acceleration data for transportation mode prediction". Transportation. 42(1), pp. 163-188.

Shan, R., M. Zhong, and C. Lu. (2013)."Comparison between traditional four-step and activity-based travel demand modeling - A case study of Tampa, Florida". Presented at the 2nd International Conference on Transportation Information and Safety (ICTIS).

Shannon, T., B. Giles-Corti, T. Pikora, M. Bulsara, T. Shilton and F. Bull. (2006). "Active commuting in a university setting: Assessing commuting habits and potential for modal change". Transport Policy 13(3): 240-253.

Shieh, H. L. (2014). "Robust validity index for a modified subtractive clustering algorithm". Applied Soft Computing. 22, pp. 47-59.

Spinney, J. E. and H. Millward. (2011). "Weather impacts on leisure activities in Halifax, Nova Scotia". International Journal of Biometeorology. 55(2), pp. 133-145.

Statistics Canada. (2010). "General social survey-Overview of the time use of Canadians, Statistics Canada". URL http://www.statcan.gc.ca/pub/89-647-x/89-647-x2011001-eng.htm.

Statistics Canada. (2011). "2006 census public use microdata file (PUMF), hierarchical file: Documentation and user guide". Statistics Canada. URL http://equinox.uwo.ca/docfiles/2006_Census/pumf/hier/pumf%20user%20guide.pdf.

Statistics Canada. (2015). "Canadian vehicle survey: Annual (53-223-X)". URL http://www5.statcan.gc.ca/olc-cel/olc.action?objId=53-223-X&objType=2&lang=en&limit=1

Stopher, P. R., D. T. Hartgen, and Y. Li. (1996). "SMART: simulation model for activities, resources and travel". Transportation. 23(3), pp. 293-312.

Suthaharan, S. (2015). "Machine learning models and algorithms for big data classification: Thinking with examples for effective learning". Springer US.

Tamminga, G., P. Knoppers and J. W. C. van Lint. (2014). "Open traffic: A toolbox for traffic research". Procedia Computer Science 32: 788-795.

Tan, P. N., M. Steinbach, and V. Kumar. (2005). "Classification: Basic concepts, decision trees, and model evaluation". Introduction to Data Mining. Addison-Wesley Companion Book Site. p. 145-205.

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

TURP. (2008). "TURP (Time Use Research Program)". Halifax regional space time activity research (STAR) survey: A GPS-assisted household time-use survey, survey methods. Halifax: Saint Mary's University.

Ubillos, B. J. and F. A. Sainz. (2004). "The influence of quality and price on the demand for urban transport: The case of university students". Transportation Research Part A: Policy and Practice 38(8): 607-614.

VDOT. (2009). "Implementing Activity-based models in Virginia". VTM Research Paper 09-01.

Villanueva, K., B. Giles-Corti and G. McCormack. (2008). "Achieving 10,000 steps: A comparison of public transport users and drivers in a University setting". Preventive Medicine 47(3): 338-341.

Voas, D. and P. Williamson. (2000). "An evaluation of the combinatorial optimization approach to the creation of synthetic microdata". International Journal of Population Geography 6(5): 349-366.

Voas, D. and P. Williamson. (2001). "Evaluating goodness-of-fit measures for synthetic microdata". Geographical and Environmental Modeling 5(2): 177-200.

Volosin, S. E., S. Paul, R. M. Pendyala, V. Livshits, and P. Maneva. (2014)."Activity-travel characteristics of a large university population". Presented at the 93th Annual Meeting of the Transportation Research Board (TRB), Washington, D.C.

Vovsha, P., E. Petersen, and R. Donnelly. (2002). "Microsimulation in travel demand modeling: Lessons learned from the New York best practice model". Transportation Research Record: Journal of the Transportation Research Board 1805, pp. 68-77.

Vovsha, P., E. Petersen, and R. Donnelly. (2004). "Model for allocation of maintenance activities to household members". Transportation Research Record: Journal of the Transportation Research Board 1894, pp. 170-179.

Wang, X., A. Khattak, and S. Son. (2012). "What can be learned from analyzing university student travel demand?". Transportation Research Record: Journal of the Transportation Research Board 2322, pp. 129-137.

Weiner, E. (1999). "Urban transportation planning in the United States: An historical overview". Praeger.

Wen, C. H. and F. S. Koppelman. (2000). "A conceptual and methodological framework for the generation of activity-travel patterns". Transportation. 27(1), pp. 5-23.

Widen, J., A. Molin, and K. Ellegard. (2012). "Models of domestic occupancy, activities and energy use based on time-use data: deterministic and stochastic approaches with application to various building-related simulations". Journal of Building Performance Simulation. 5(1), pp. 27-44.

Wu, B. M., M. H. Birkin, and P. H. Rees. (2011). "A dynamic MSM with agent elements for spatial demographic forecasting". Social Science Computer Review. 29(1), pp. 145-160.

Ye, X., K. Konduri, R. M. Pendyala, B. Sana, and P. Waddell. (2009)."A methodology to match distributions of both household and person attributes in the generation of synthetic populations". Presented at the 88th Annual Meeting of the Transportation Research Board (TRB), Washington, D.C.

You, J., J. Wang, and J. Guo. (2017). "Real-time crash prediction on freeways using data mining and emerging techniques". Journal of Modern Transportation. 25(2), pp. 116-123.

Appendix A Copyright Permission

December 18, 2017

Transportation

I am preparing my Ph.D. thesis for submission to the Faculty of Graduate Studies at Dalhousie University, Halifax, Nova Scotia, Canada. I am seeking your permission to include a manuscript version of the following paper(s) as a chapter in the thesis:

[Mohammad Hesam Hafezi, Lei Liu and Hugh Millward, "A time-use activity-pattern recognition model for activity-based travel demand modeling", Transportation, 2017: 1-26, DOI:10.1007/s11116-017-9840-9.]

Canadian graduate theses are reproduced by the Library and Archives of Canada (formerly National Library of Canada) through a non-exclusive, world-wide license to reproduce, loan, distribute, or sell theses. I am also seeking your permission for the material described above to be reproduced and distributed by the LAC(NLC). Further details about the LAC(NLC) thesis program are available on the LAC(NLC) website (www.nlc-bnc.ca).

Full publication details and a copy of this permission letter will be included in the thesis.

Yours sincerely,

Mohammad Hesam Hafezi

SPRINGER NATURE LICENSE TERMS AND CONDITIONS

Dec 20, 2017

This Agreement between Dalhousie University -- Mohammad Hesam Hafezi ("You") and Springer Nature ("Springer Nature") consists of your license details and the terms and conditions provided by Springer Nature and Copyright Clearance Center.

License Number	4253151334052
License date	Dec 20, 2017
Licensed Content Publisher	Springer Nature
Licensed Content Publication	Transportation
Licensed Content Title	A time-use activity-pattern recognition model for activity-based travel demand modeling
Licensed Content Author	Mohammad Hesam Hafezi, Lei Liu, Hugh Millward
Licensed Content Date	Jan 1, 2017
Type of Use	Thesis/Dissertation
Requestor type	academic/university or research institute
Format	print and electronic
Portion	full article/chapter
Will you be translating?	no
Circulation/distribution	>50,000
Author of this Springer Nature content	yes
Title	A time-use activity-pattern recognition model for activity-based travel demand modeling
Instructor name	Lei Liu
Institution name	Dalhousie University
Expected presentation date	Mar 2018
Portions	Full article/chapter
Requestor Location	Dalhousie University 6299 South Street Halifax, NS B3H 4R2 Canada Attn: Dalhousie University
Billing Type	Invoice
Billing Address	Dalhousie University 6299 South Street Halifax, NS B3H 4R2 Canada Attn: Dalhousie University
Total	0.00 USD
Terms and Conditions	

Springer Nature Terms and Conditions for RightsLink Permissions

Springer Customer Service Centre GmbH (the Licensor) hereby grants you a non-exclusive, world-wide licence to reproduce the material and for the purpose and

requirements specified in the attached copy of your order form, and for no other use, subject to the conditions below:

1. The Licensor warrants that it has, to the best of its knowledge, the rights to license reuse of this material. However, you should ensure that the material you are requesting is original to the Licensor and does not carry the copyright of another entity (as credited in the published version).

   If the credit line on any part of the material you have requested indicates that it was reprinted or adapted with permission from another source, then you should also seek permission from that source to reuse the material.

2. Where **print only** permission has been granted for a fee, separate permission must be obtained for any additional electronic re-use.

3. Permission granted **free of charge** for material in print is also usually granted for any electronic version of that work, provided that the material is incidental to your work as a whole and that the electronic version is essentially equivalent to, or substitutes for, the print version.

4. A licence for 'post on a website' is valid for 12 months from the licence date. This licence does not cover use of full text articles on websites.

5. Where **'reuse in a dissertation/thesis'** has been selected the following terms apply: Print rights for up to 100 copies, electronic rights for use only on a personal website or institutional repository as defined by the Sherpa guideline (www.sherpa.ac.uk/romeo/).

6. Permission granted for books and journals is granted for the lifetime of the first edition and does not apply to second and subsequent editions (except where the first edition permission was granted free of charge or for signatories to the STM Permissions Guidelines http://www.stm-assoc.org/copyright-legal-affairs/permissions/permissions-guidelines/), and does not apply for editions in other languages unless additional translation rights have been granted separately in the licence.

7. Rights for additional components such as custom editions and derivatives require additional permission and may be subject to an additional fee. Please apply to Journalpermissions@springernature.com/bookpermissions@springernature.com for these rights.

8. The Licensor's permission must be acknowledged next to the licensed material in print. In electronic form, this acknowledgement must be visible at the same time as the figures/tables/illustrations or abstract, and must be hyperlinked to the journal/book's homepage. Our required acknowledgement format is in the Appendix below.

9. Use of the material for incidental promotional use, minor editing privileges (this does not include cropping, adapting, omitting material or any other changes that affect the meaning, intention or moral rights of the author) and copies for the disabled are permitted under this licence.

10. Minor adaptations of single figures (changes of format, colour and style) do not require the Licensor's approval. However, the adaptation should be credited as shown in Appendix below.

Appendix — Acknowledgements:

For Journal Content:
Reprinted by permission from [**the Licensor**]: [**Journal Publisher** (e.g. Nature/Springer/Palgrave)] [**JOURNAL NAME**] [**REFERENCE CITATION** (Article name, Author(s) Name), [**COPYRIGHT**] (year of publication)

For **Advance Online Publication papers:**
Reprinted by permission from [**the Licensor**]: [**Journal Publisher** (e.g. Nature/Springer/Palgrave)] [**JOURNAL NAME**] [**REFERENCE CITATION** (Article name, Author(s) Name), [**COPYRIGHT**] (year of publication), advance online publication, day month year (doi: 10.1038/sj.[JOURNAL ACRONYM].)

For Adaptations/Translations:
Adapted/Translated by permission from [**the Licensor**]: [**Journal Publisher** (e.g. Nature/Springer/Palgrave)] [**JOURNAL NAME**] [**REFERENCE CITATION** (Article name, Author(s) Name), [**COPYRIGHT**] (year of publication)

Note: For any republication from the British Journal of Cancer, the following credit line style applies:

Reprinted/adapted/translated by permission from [**the Licensor**]: on behalf of Cancer Research UK: : [**Journal Publisher** (e.g. Nature/Springer/Palgrave)] [**JOURNAL NAME**] [**REFERENCE CITATION** (Article name, Author(s) Name), [**COPYRIGHT**] (year of publication)

For **Advance Online Publication** papers:
Reprinted by permission from The [**the Licensor**]: on behalf of Cancer Research UK: [**Journal Publisher** (e.g. Nature/Springer/Palgrave)] [**JOURNAL NAME**] [**REFERENCE CITATION** (Article name, Author(s) Name), [**COPYRIGHT**] (year of publication), advance online publication, day month year (doi: 10.1038/sj. [JOURNAL ACRONYM])

For Book content:
Reprinted/adapted by permission from [**the Licensor**]: [**Book Publisher** (e.g. Palgrave Macmillan, Springer etc) [**Book Title**] by [**Book author**(s)] [**COPYRIGHT**] (year of publication)

Other Conditions:

Version 1.0