Title	Modeling occupant behavior for residential energy demand simulation: enhancement of diversity by incorporating spatial variation
Author(s)	Li, Yuanmeng
Citation	大阪大学, 2022, 博士論文
Version Type	VoR
URL	https://doi.org/10.18910/89625
rights	
Note	

Doctoral Dissertation

Modeling occupant behavior for residential energy demand simulation: enhancement of diversity by incorporating spatial variation

LI YUANMENG

June 2022

Graduate School of Engineering,
Osaka University

Abstract

Occupant behavior (OB) models, that simulate the daily activity of residents, have been developed to be integrated into building simulation tools to estimate residential building energy demand. OB models contribute to improving accuracy as they can capture the impact of the OBs on energy demand. However, previous OB studies mainly focused on the simulation algorithms and paid less attention to the design of the overall model, especially the pre-simulation process — data processing, variable selection, and parameter preparation. In addition, the diversity among the occupants particularly the spatial variation was greatly underestimated. Therefore the existing OB models hinder generating realistic behavior profiles thereby leading to less reliable energy predictions for future building design and planning.

Based on the aforementioned background, this thesis aims to provide a systematic investigation in three steps. First, various machine learning based OB models are evaluated and compared to illustrate the importance of the pre-simulation process for the OB model development. Second, the existence of spatial variation and historical change in OBs were confirmed. The significance of these factors on model performance is further evaluated. Finally, new OB models incorporating spatial variation are proposed to enhance the diversity exhibited over given heterogeneous regions. The thesis is divided into 6 chapters.

In Chapter 1, the thesis centers on introducing the energy use and energy demand modeling in the residential sector, the role of OB in influencing the energy demand, the development of OB modeling, and diversity in OBs. Then, three critical reviews are presented to reflect the current research status. One is conducted to summarize the model engine, the modeling methods to prepare the parameter, and variables used in the model in previous studies. The remaining two summarize previous studies on modeling with spatial variation for both the engineering field and other fields in terms of the aspects — empirically representing the spatial variation and simulating the research object with the consideration of the spatial variation. Then, the research gap was found based on the review. In addition, the overall framework of research targeting the assessment of OB model performance and development of the model that can involve spatial variation are outlined.

In Chapter 2, the pre-simulation process of OB modeling is analyzed and its importance is evaluated. In this chapter, two crucial questions representing the vital components of the pre-simulation process that have been paid less attention to in previous studies were solved: 1) which variables should be considered and 2) what is the most appropriate parameter preparation method. Using four machine learning based parameter preparation methods combined with three cases including different variable consideration conditions based on the single-year American time use survey (ATUS) data, the significance of the design of the pre-simulation process of OB modeling for residential energy demand simulations had been highlighted.

In Chapter 3, the existence of the spatial variation and historical change are confirmed through a case study of watching television activity for the women population based on the multi-year ATUS data. In this chapter, the spatial variation is checked for the time interval with the largest discrepancies in the probability of undertaking watching television among the years from 2009 to 2019. The historical change is detected in a five-year period (2009–2014–2019). In addition, the performance of the conventional logistic regression model to reproduce the spatial variation is checked by comparing the distribution in space with the observations.

In Chapter 4, new OB models that can incorporate spatial variation are established. In this chapter, the spatial variation for four activities — sleeping, cooking and washing up, watching television, and commuting for 6 groups of the subpopulation of women is confirmed and modeled. Three new OB models are developed based on different representations of spatial variation and assessed in terms of indicators at the national and state levels. Further potential applications including the broader range such as energy demand simulation considering the spatial variation in OB are discussed.

Chapter 5 discusses the main outcomes of the thesis, as well as its limitations and further work.

Chapter 6 summarizes the study findings, conclusions, and contributions.

Overall, this work has contributed to broadening the knowledge of the pre-simulation and successfully incorporating spatial variation to enhance the model diversity thus greatly improving the understanding of the systems and identifying areas to support sustainable decision-making depending on the time-use of people in different regions. These findings can be extended to develop more realistic energy demand models in future work.

List of Publications

1. Li, Y., Yamaguchi, Y., & Shimoda, Y. (2022). Impact of the pre-simulation process of occupant behavior modeling for residential energy demand simulations. Journal of Building Performance Simulation, 15(3), 287–306.
2. Li, Y., Yamaguchi, Y., Chen, C. F., & Shimoda, Y. Spatial variation and historical change in occupant behavior: statistical analysis and application on household activities and time scheduling. Proceedings of Building Simulation 2021: 17th Conference of IBPSA.
3. Li, Y., Yamaguchi, Y., Torriti, J., & Shimoda, Y. Modeling of occupant behavior considering spatial variation: geostatistical analysis and application based on American Time Use Survey data. Energy and Buildings, under review.

Acknowledgment

This work has been made possible by the tireless support and effort of my academic advisors. I thank Professor Yamaguchi for being there at every step throughout the study period. His great contribution to the ideas that gave birth to the outputs in this dissertation cannot be overstated. I thank Professor Shimoda for their support with the technical guidance and their insightful comments during the seminars. As well as much thanks to Professor Chien-fei Chen, Professor Jacopo Torriti, and Professor Kondo.

I also thank my colleagues in the Shimoda laboratory; especially Usama Perwez who sacrificed some of his time to help me out with the difficult areas, especially sharing useful papers and discussing the new ideas that related to my work. I drew a lot of motivation from each of the members, especially Takai Shiho, Okubo Hiromi, Furutani Sumiko, and Okumura Miyuki, and thank them for their assistance whenever I needed it. Special mention of Xiangyun Shi whom I consulted from time to time although based in a different laboratory.

Special mention of my parents who have been my support base since the start of the foreign trip and the start of my academic journey. Their undying support both moral and financial has seen me exploit my potential to the best of my ability. Together with my lovely friends Yixuan Wang, Yue Li, Hannan Fu, Yilin Xie, and Xiaoye Tian who are all been great pillars for me.

Table of Contents

Figures

Tables

Equation

Abbreviation and Acronyms

ANN	artificial neural network
ATUS	American time use survey
GWR	Geographically Weighted Regression
MLR	multinomial log-linear regression
Moran’s *I* test	Global Moran Index test
MSD	mean standard deviation
OB	occupant behavior
ORs	odds ratios
RCs	regression coefficients
RMSE	root mean squared error
RMSE_GA	averaged root mean squared error of subgroups
SAR	spatial autoregressive
SVM	support vector machine
TAE	total absolute error
TUD	time use data

This page is intentionally left blank

1 Introduction

This chapter composed a concise introduction to occupant behavior (OB) modeling and the overview of this thesis. First, how energy is consumed in the residential sector is introduced. Second, the role of OB in building energy performance is discussed. Third, the development of OB modeling underpinning the energy demand simulation is presented. Forth, the diversity between occupants and their behaviors resulting in major uncertainty in predicting building energy performance is highlighted. Fifth, three critical reviews are presented to reflect the current research status. Then, the research gap was found based on the review. Finally, the aim and objectives, contributions, and the outline of the thesis are explained respectively.

1.1 Energy use in the residential sector

The residential sector accounts for a relatively high proportion (16–50%) of the national energy consumption and it is the major sector in terms of electricity consumption (Martinaitis et al., 2015; Wilke et al., 2013). Since the great impact on the energy supply side, the residential sector draws more and more attention to help reduce the cost and the energy demand throughout the day as well as to better support the design of the control algorithms for the supply-side systems. Therefore, the residential sector has great energy-saving potential. At the same time, residential building energy systems are tightly related to national or regional energy and environmental policies (Yu et al., 2011). Hence, the energy demand models which are the foundation of making related strategies and plans for the entire industry progress have been developed.

In recent decades, two types of modeling approaches — top-down and bottom-up have been applied to simulate the building energy demand (Kavgic et al., 2010). The energy demand models applied with the top-down approach are established at an aggregated level, typically by fitting historical time series data of national energy consumptions or greenhouse gas emissions. These models aim to illustrate the inter-relationships between the energy sector and the macroeconomics primarily based on considering the relationship between energy use and market economic factors such as fuel prices and technological progress. Models can be generally categorized into econometric and technological two groups by different variables to represent the economy. However, these top-down models are incapable of explaining factors such as the building physical factor and OBs that can affect energy demand.

On the contrary, the bottom-up modeling approach is widely applied to simulate aggregated residential energy consumption by characterizing individual appliances and loads within a building (McKenna et al., 2018). Therefore, various modules such as the external temperature model, thermal demand model, and solar photovoltaic model are combined to estimate the overall energy demand. Nowadays, the OB model is gradually integrated into the bottom-up based energy demand models as they can encapsulate the full range and timing of OB (i.e., occupants' presence, activities, and dependent behaviors) on the buildings' energy balance. Moreover, bottom-up based energy demand models are able to predict the future changes in the physical composition of buildings or the ownership of appliances as well as the changes in the population's demographic/behavioral characteristics (Wilke et al., 2013). Hence, this thesis is focusing on researching the OB model underpinning the bottom-up based energy demand model. Bottom-up based energy demand models can be categorized into three groups — engineering models, statistical models, and hybrid models. Most research efforts for the engineering models have focused on residential buildings using archetypes (i.e., representative buildings or prototype buildings) (Lim & Zhai, 2017). Each archetype is defined by specific features in terms of four main areas: form, envelope, system, and operation (Corgnati et al., 2013). These engineering models simulate the energy demand for the archetypes instead of the whole building stock. The total energy demand is then aggregated for all predicted energy demands of each archetype with proper weighting factors such as the floor area. Regarding the statistical models, most of them are based on regression techniques. Such models are capable of taking demographics and OBs that have a significant influence on energy consumption into account. Regarding the hybrid models, they combine modeling components where both building physics and statistical approaches were applied and they can solve more practical problems (Kavgic et al., 2010). Since these three groups of the bottom-up models are established at a disaggregated level, sufficient databases of empirical data that can support the description of each model component to characterize each individual load are required (Shorrock & Dunster, 1997).

1.2 Role of occupant behavior

Simulation studies (Mastrucci et al., 2017; Wilke et al., 2013; Zhao et al., 2014) have confirmed that OB is an important determinant of building energy consumption in the bottom-up based statistical or hybrid type of energy demand models and a leading source of uncertainty in predicting building energy use, as energy-consuming appliances are generally operated to satisfy people's daily needs in correspondence to the activities that the occupants perform. For instance, the oven for cooking meals, the washing machine for washing the clothes, lights for lighting, and the air-conditioner for adjusting the temperature during the summer and winter days. Occupants also adjust the settings of the indoor environment to pursue comforts, such as operating window openings and shading devices (Mosteiro-Romero et al., 2017; Ruan et al., 2017) to improve the indoor air quality or keep the indoor temperature within a comfortable range. Furthermore, OB is a vital factor in the assessment of technologies employed in building design and retrofit (Yan et al., 2015). Many case studies have demonstrated that OB influences the adaptability and implementation of building technologies for better assessing the building energy performance as well as accurately simulating the energy demand (Belessiotis & Mathioulakis, 2002; Fabi et al., 2013).

1.3 Occupant behavior modeling

Since the significance of the OB is clear, numerous models (called OB models) for capturing the occupancy, activities, and actions of building occupants have been developed for understanding, modeling, and analyzing OBs and their impacts on building energy demand. To capture the dynamic changes in the building energy demand and diversity among the households, OB models are gradually employed as one module in energy demand models. Generally, for residential buildings, most of the divisions of the building stock are based on building physics of form, envelope, and system characteristics, only a few have based on the relevance of OBs, which can define the archetype in the operation area (Buttitta et al., 2017). Heinrich et al., 2022 built archetypes that are related to specific housing contexts and energy consumption levels based on the seven clusters of OBs in the residential sector.

OB, as mentioned in Chapter 1.2, can be modeled in terms of occupancy, activities, and actions. The occupancy model simulates the presence and absence status of the occupants in the targeted building. The activity model takes into account the various daily activities of occupants over the complete time range to provide a better time-dependent activity profile. Chapter 1.5.1 reviewed OB models including all these three types. However, this thesis mainly focuses on the activity model. In the following chapters, unless explicitly stated, all OB models refer to the activity model. Moreover, the OB models considered and developed in Chapters 2–4 also indicate the activity model. The action model models specific actions such as the window opening for simulating certain loads or evaluating indoor air quality.

The majority of OB models were established based on the time use data (TUD) as TUD is an important data source that collects invaluable information — sociodemographic information, housing information, and the daily activity schedule for recorded households. Particularly, TUD is widely used as it was collected at the national level for many countries (e.g., America, United Kingdom, German, Australia, China, and Japan). Existing OB models use either deterministic or stochastic modeling techniques (Happle et al., 2018). This thesis focuses on the stochastic models because deterministic models only capture the average behavior of energy demand, whereas stochastic models enable the production of stochastic behavior in building energy demand. Stochastic models employ empirical statistical data such as the TUD to model the probability that the occurrence or undertaking of activity thus reflecting the OBs more realistically (Jeong et al., 2021). Therefore, the stochastic OB models can better assist to simulate the actual building energy demand in terms of diversity and the variability among the simulated occupants (Yamaguchi et al., 2019).

To develop an OB model or energy demand model with consideration of OBs, three processes should be carefully designed. Figure 1-1 shows the whole general procedure with three processes for simulating building energy demand using OBs as inputs. The OBs input for the energy demand model is located in the post-simulation process. According to Figure 1-1, OBs comprise data regarding occupancy states, activities, and/or actions referred to as “model objectives”. These data are stochastically generated by a model engine during the simulation of OBs. The model engine has several model parameters based on which OBs are generated. Model parameters can be used to differentiate OBs among simulated occupants according to the conditions given in the simulation process. The model parameters were prepared based on a model developed during the pre-simulation process. In the pre-simulation process, first, the input data were prepared based on raw data, for example, TUD through data preprocessing. Then, a certain parameter preparation method was applied based on the input data to develop the model to prepare the model parameters for the simulated occupants. For example, many studies have quantified model parameters based on sample distribution. Statistical and machine learning methods can be developed for preparation. The parameter preparation method may consider several variables, such as demographic conditions so that the influence of the considered variables can be reflected in the model parameters and the resultant OBs in the simulation process. In this thesis, we refer to the process of combining data processing, variable selection, and parameter preparation as the pre-simulation process.

Each process matters for the simulation results as all processes are closely linked. For this thesis, three vital selections as shown in Figure 1-1: selection of considered variables, selection of the parameter preparation method, and the selection of the engine which require thorough considerations to allow the better development of the framework together with better design appropriate combinations of these three processes are highlighted. The selection of the engine is located in the simulation process and most studies had paid great attention to improving model methods for the engine. As for the first two selections located in the pre-simulation process, few studies had considered although greater knowledge of the pre-simulation process is needed as revealed in Chapter 1.5.1.

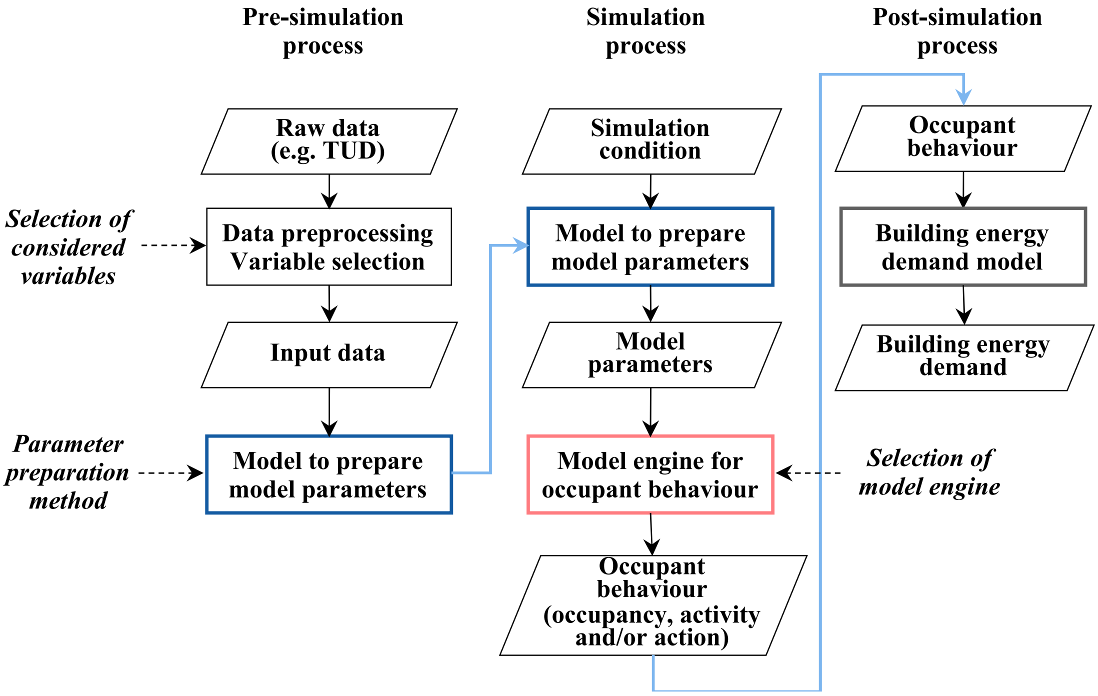

Pre-simulation process
Simulation process
Post-simulation process
Raw data (e.g. TUD)
Selection of considered variables
Data preprocessing Variable selection
Input data
Parameter preparation method
Model to prepare model parameters
Simulation condition
Model to prepare model parameters
Model parameters
Model engine for occupant behaviour
Selection of model engine
Occupant behaviour (occupancy, activity and/or action)
Occupant behaviour
Building energy demand model
Building energy demand

Figure 1-1. Procedure for simulating energy demand considering occupant behavior.

1.4 Diversity

Modeling OBs is critically important and it has become increasingly important, as the modeling of energy demands with a high spatiotemporal resolution has attracted attention. To this end, researchers in academia and industry have developed various OB models. However, there exists complicated decision-making of the occupants to conduct their daily behaviors. Therefore, they will unlikely exhibit the reaction as what the researchers assumed or set in the model scenarios (O'Brien & Gunay, 2015). Also, most of the previous studies seem to model the typical or representative occupants for the building energy demand simulation. As revealed by researchers, the performance gap exists between the energy demand simulation and reality (Happle et al., 2018; Martinaitis et al., 2015; Yan et al., 2017) and it comes from the following points: 1) the use of oversimplified assumptions such as a fixed occupancy schedule (Delzendeh et al., 2017), 2) inappropriate consideration of the interactions between appliances and building systems (Diao et al., 2017), and 3) ignorance of the diversity resulting from different sociodemographic conditions and/or other influencing conditions for OBs (Happle et al., 2018). Among all, the underestimation of the diversity among the occupants especially the general occupants is one of the major sources resulting in the performance gap.

Diversity itself has been defined in different ways in different contexts by different fields, even in the engineering field (O’Brien et al., 2017a). Herein, the simple way to understand the diversity is the variability response from the behaviors that occupants conducted. In order to model the diversity in building OBs thereby modeling demand loads more accurately, especially the dynamic changes in the energy demand, various factors have been considered to represent the diversity. The variable representing the household composition which is related to one of the most important factors — the demographic factor has a significant influence on energy consumption as it varies significantly among households (Jones et al., 2015). Likewise, the type of the housing unit, climate, and day of the week these variables also play important role in influencing the residential energy demand. We have summarized commonly used factors in the literature in Chapter 1.5.1.

Among all factors, the geographic factor (also called the spatial factor) to represent the spatial variation has not been fully investigated (Li et al., 2019). Spatial variation essentially refers to the rules or tendencies of objects of the research exhibited in a given space. It can be represented and considered in the modeling in different ways. Many researchers have proven that spatial variation plays an important role in simulating energy demand. Druckman & Jackson, 2008 demonstrated that household energy use and the associated carbon dioxide emissions vary significantly with household socioeconomic conditions and locations. Rural/urban environments are another important factor in devising policies for a low-carbon society. Halleck Vega et al., 2022 pointed out that although the spatial perspective has received limited attention in the literature, it is a significant factor in energy-related policy considerations. They observed that the spatial factor is important, and ignoring it can lead to inaccurate conclusions. Furthermore, spatial variation also exists in time use. Several studies showed differences in the time use of occupants among countries, which revealed spatial variation existed in the time spent on OBs (Al-Mumin et al., 2003; Jeong et al., 2021; Torriti, 2012). Esteban Ortiz-Ospina & Roser, 2020 found that OBs conducted by people are spatially varied in European countries, which cannot be effectively explained by economic or demographic differences. Such spatial variation in OBs may further occur within a country or even within a region. Studying how people spend their time over space provides an important perspective for understanding living conditions, economic opportunities, and general well-being. However, a consistent approach to empirically represent spatial variation in OB and to consider it in OB modeling is currently lacking, but useful spatial analysis and modeling methods have been developed in other fields as shown in review Chapters 1.5.2 and 1.5.3.

1.5 Critical literature review

The literature review is divided into three parts. Chapter 1.5.1 summarizes the reviewed studies that developed an OB model or an energy demand simulation with consideration of OBs. We mainly focus on three important sub-process as mentioned in Chapter 1.3 for energy demand modeling with consideration of OBs — the model engine, the methods for preparing the parameters for the model engine, and variables. Chapter 1.5.2 summarizes the reviewed studies related to OB and energy modeling with spatial variation. Chapter 1.5.3 summarizes the reviewed studies relevant to spatial variation in other fields.

1.5.1 Studies related to OB and energy demand modeling

This chapter is a summary of the studies in terms of the three selections of the model development. First, the model engine of two types is reviewed. Secondly, the parameter preparation methods considered in the previous studies are summarized. Finally, the type of factors used in the existing models is categorized.

The selection for the engine is a core process for the OB models as the engine determines the OB model outputs which are also the inputs for the energy demand model. Osman & Ouf, 2021 summarized the model engine in modeling occupants' presence and behaviors. Most of the reviewed studies used discrete-time or discrete-event approaches, which are the main approaches, as summarised in the second column of Table 1-1. A discrete-time approach considers a fixed time interval, and the changes in the model objectives are examined at each time step. In this approach, the time-inhomogeneous Markov chain model is widely used (Aerts et al., 2014; Diao et al., 2017; Richardson et al., 2008; Widén et al., 2009). Time-inhomogeneous Markov chain model consists of the current state space and the probabilities associated with the transitions from each of the states into the others (Ramírez-Mendiola et al., 2019). The stochastic processes that can be adequately described by a time-inhomogeneous Markov chain model are said to satisfy the Markov property. This property for the time-inhomogeneous Markov can also be said as memoryless: $\mathrm{P}(x_{t+1} = i_{n+1} | x_t = i_n, x_{t-1} = i_{t-1}, \dots, x_0 = i_0) = \mathrm{P}(x_{t+1} = i_{n+1} | x_t = i_n) = p$, which indicated that state x at time $t+1$ only related to that state at time t. Advanced methods have also been applied. Liisberg et al., 2016 used hidden Markov models to create methods for the indirect observation and characterization of OB. Kleinebrahm et al., 2021 applied neural networks which combined state-of-the-art long short-term memory (LSTM) and attention-based autoregressive models with imputation models to generate weekly activity profiles capable of capturing long-term dependencies in mobility and activity patterns. A discrete-event approach reproduces an OB as an ordered event sequence. Each event has a specific start time and duration. Wilke et al., 2013 presented an approach to model residential activities based on time-dependent probabilities for the start of activities and the corresponding distributions of activity durations.

Table 1-1. Model engines and parameter preparation methods of previous studies.

Literature	Model engine	Parameter preparation method	Segmentation
Diao et al., 2017	Discrete-time: Markov chain	Clustering, neural network	Yes
Liisberg et al., 2016	Discrete-time: Markov chain	Sample distribution	No
Ramírez-mendiola et al., 2019	Discrete-time: Markov chain	Sample distribution	Yes
Richardson et al., 2008	Discrete-time: Markov chain	Sample distribution	Yes
Aerts et al., 2014	Discrete-time: Markov chain	Sample distribution	Yes
Widén et al., 2009	Discrete-time: Markov chain	Sample distribution	Yes
Jones et al., 2017	Discrete-time	Multivariate logistic regression	No
Okada et al., 2020	Discrete-event	Logistic regression	Yes
Yamaguchi & Shimoda, 2017	Discrete-event	Sample distribution	Yes
Tanimoto et al., 2008b, 2008a	Discrete-event	Sample distribution	Yes
Fischer et al., 2015	Discrete-event	Sample distribution	Yes
Wilke et al., 2013	Discrete-event	Logistic regression and sample distribution	No
Deng & Chen, 2019	Discrete-event	Neural network	Yes
Kleinebrahm et al., 2021	Discrete-time	Neural networks	Yes

Existing OB models have used the various parameter preparation methods listed in the third column of Table 1-1. The methods were divided into three groups. The first group used a sample distribution or fitted distribution (Aerts et al., 2014; Fischer et al., 2015; Liisberg et al., 2016; Ramírez-mendiola et al., 2019; Richardson et al., 2008; Tanimoto et al., 2008b, 2008a; Widén et al., 2009; Yohei Yamaguchi & Shimoda, 2017). For example, Richardson et al., 2008 used the transition probability derived from the TUD by dividing the occurrence of transitions in the occupancy state by the number of samples to model occupancy. In the sample-based method, the same modeling parameters are applied to simulated individuals; thus, diversity is ignored. The second group used regression to quantify the modeling parameters. Logistic regression is the most frequently used method to consider variations owing to various factors (Jones et al., 2017; Okada et al., 2020; Wilke et al., 2013). In recent decades, multinomial log-linear regression models have become useful to model OBs. It is a generalization of binomial logistic regression which can deal with the classification of multiple labels of dependent variable Y which is the probability of activity. The mathematical expression of multinomial log-linear regression models is:

$$\ln\left(\frac{\mathrm{P}(Y_i = y_k)}{\mathrm{P}(Y_i = y_1)}\right) = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_m x_{mi} = Z_K, \qquad K = 2, \ldots, N \qquad 1\text{-}1$$

$$\mathrm{P}(Y_i = y_1) = \frac{1}{1 + \sum_{j=2}^{N} e^{Z_j}} \tag{1-2}$$

$$\mathrm{P}(Y_i = y_k) = \frac{e^{Z_k}}{1 + \sum_{j=2}^{N} e^{Z_j}} \tag{1-3}$$

where the i means the ith observation and N means the total number of activities, x represents the attribution of occupants. $Y_i = y_1$ is selected as the base case and the choice of the base case does not change the calculations of probability, it only affects the coefficients and ways to explain the odds ratio. An example of the use of multinomial log-linear regression is Wilke et al., 2013, who proposed the use of it to model the activity-starting probability.

The third group used advanced data-driven methods (e.g., statistical tests and random forests). Nowadays, the use of the artificial neural network to develop a model based on big data become more popular. Same as regression models the artificial neural network (ANN) also contains the input, calculation functions (i.e., one or more layers), and output. The ANN has two types: feedforward and feedback network architectures. One of the distinct characteristics of the ANN is it learns from experience and examples and then can adapt to changing situations (Rafiq et al., 2001). Kleinebrahm et al., 2021 applied advanced ANNs to simulate OBs including synthetic weekly mobility schedules. Deng & Chen, 2019 applied an ANN method to model the OB occurrence. As presented in the fourth column of Table 1-1, many of the existing models grouped the input data before applying the parameter preparation method. Many studies have used basic demographic conditions for segmentation, such as age and gender (Okada et al., 2020) and the distinction between weekdays and weekends (Ramírez-Mendiola et al., 2019; Richardson et al., 2008). Diao et al., 2017 and Aerts et al., 2014 grouped TUD samples based on the characteristics of time allocation observed in TUD using a clustering method. Segmentations can improve the reproduction of diversity in OBs, even when a sample-based parameter preparation method is used.

For the segmentation and development of statistical OB models, previous studies considered several variables to address their influence on OBs and to enhance the diversity among the simulated occupants (Okada et al., 2020), although many of the models suppressed occupant diversity (O’Brien et al., 2017b). Haldi et al., 2017 and Tahmasebi & Mahdavi, 2018 revealed that the consideration of diversity enhances the diversity in energy demand among households and improves the reproducibility of building energy demand models, including extreme values.

We categorized the variables used in the existing OB models into the eight categories listed in Table 1-2. This categorization was originally used by Stazi et al., 2017 and modified by the authors based on the reviewed studies listed in Table 1-3. We refer to the categories as "influencing factors". Table 1-3 includes studies that were not listed in Table 1-1 because they did not provide an OB model but provided relevant evidence indicating the significant influence of a factor on OB.

According to Table 1-3, variables related to demographic and time factors are most commonly considered regardless of the model objectives. The consideration of variables representing the demographic factor enables consideration of the inter-person/household diversity of OBs, whereas time factor variables enhance the reproducibility of temporal variations in OBs which are important for reproducing the time-dependent characteristics of building energy demand. Occupancy factor variables contribute to the dependency on the designated location (e.g., performed at home). Notably, psychological and environmental factors are only considered in the modeling of occupant actions; for example, window opening.

Table 1-2. Summary of influencing factors categories and corresponding representative variables considered in OB models.

Influencing factor	Representative variables
Demographic	Individual attributes: age, gender, employment status Household attributes: household size, household composition Housing condition: housing Attributes of other household members: employment status, age
Time	Time of day, day of week, distinction between weekday and weekend
Activity	Previous activity, accompanying people
Geographic	Metropolitan status, region, nation
Appliance	Ownership, appliance control
Environmental	Local weather, climate zone, humidity
Occupancy	Presence, arrival, awake status
Psychological	Motives, goals, setting preferences

Table 1-3. Summary of influencing factors considered in previous studies.

Literature	**Influencing factor**							
	Demo-graphic	**Time**	**Activity**	**Geo-graphic**	**App-liance**	**Environ-mental**	**Occu-pancy**	**Psycho-logical**
Diao et al., 2017	√	√		√				
Liisberg et al., 2016	√	√						
Ramírez-mendiola et al., 2019	√							
Richardson et al., 2008	√	√					√	
Aerts et al., 2014		√	√					
Widén et al., 2009	√	√						
Jones et al., 2017		√				√		
Okada et al., 2020	√	√		√				
Yamaguchi & Shimoda, 2017	√	√	√					
Tanimoto et al., 2008b, 2008a	√	√						
Fischer et al., 2015	√	√	√					
Wilke et al., 2013	√	√		√	√		√	
Deng & Chen, 2019	√					√		
Chiou et al., 2011	√	√						
Anderson, 2016	√	√						
Buttitta et al., 2017	√	√						
Torriti, 2017		√						
De Lauretis et al., 2017	√	√		√	√			
Toftum, 2010	√							√

1.5.2 Studies related to OB and energy modeling with spatial variation

Spatial variation essentially refers to the rules or tendencies of objectives of the research exhibited in a given space. Spatial variation can be represented and considered in the modeling in different ways. Table 1-4 summarizes reviewed studies in terms of the research sector, objective, spatial variation, modeling scale, and modeling method.

There is a significant development in OB-related modeling that addresses space use. These space use studies considered spatial choice or individual preference based on geo-referenced data to determine space use (Chiou et al., 2011; Ibrahim et al., 2020). Tabak, 2009 developed a model called the User Simulation of Space Utilization that simulates space utilization in an office building by calculating the distances between the locations of different activities based on measured data. In addition to spatial utilization, the mobility and occupancy patterns of people can also be estimated based on dynamic spatial choices or preferences (Dziedzic et al., 2020; Feng et al., 2015; Kleinebrahm et al., 2021; Mohammadi & Taylor, 2017; Nassar & Elnahas, 2007; C. Wang et al., 2011). As shown in Table 1-4, the majority of the previous studies considered space use to integrate spatial variation into their works. Although these studies conducted analyses or developed models with spatial variation, the variation in OBs over space has not been discussed.

According to the fourth column of Table 1-4, only limited studies have used some independent variables related to the spatial factor to consider spatial variation during the modeling process to enhance the diversity of the model (Li et al., 2022). These spatial variables are usually used as the general variables in data-driven methods such as regression analysis and neural networks. Halleck Vega et al., 2022 assessed various factors, including seven spatial factor variables (e.g., urban–rural gradient, city center, and village center), to develop a suitable policy for increasing the uptake of carbon emission reduction measures. They also highlighted the importance of using the spatial factor for designing energy policy frameworks. Marín-Restrepo et al., 2020 identified OB patterns in office environments through data analysis and the Chi-squared test based on spatial (e.g., spatial layout and occupant orientation relative to control elements) and human factor variables. Wilke et al., 2013 considered an independent variable that indicated whether an occupant lives in an urban/suburban area to simulate the starting probability of activities through a multinomial logit model. Okada et al., 2020 applied the same method by considering city size as an independent variable to simulate the probability of undertaking activities. Rafiee et al., 2019 revealed through regression analyses that spatial context (e.g. building density and urban form) is a significant determinant of household heat consumption. Abbasabadi et al., 2019 presented an urban energy use model that captures both urban building operational energy and transportation energy consumption by localizing the energy performance data and considering various urban socioeconomic factors and spatial contexts (e.g., urban density and accessibility).

Moreover, the scale of the modeling with spatial variation in previous studies is almost limited to the building or room levels as shown in the fifth column of Table 1-4. However, modeling at the larger scale such as the neighbor scale or urban scale can improve the understanding of urban energy use by informing decision-making regarding urban morphological and spatial patterns that can affect the city structure and subsequent building operational and transportation energy end-uses (Abbasabadi et al., 2019). Further, developing a model that can be applied to multi-scale is the final goal for all related researchers.

Based on the above-mentioned, less focus has been paid to spatial variation in the OB modeling at a larger scale. Spatial variation has been insufficiently represented based on the actual data in previous studies. Although some studies used the spatial factor, there is still a lack of modeling methods to better reproduce spatial variation in OBs.

Table 1-4. The reviewed studies with spatial variation in OB modeling and energy demand modeling with consideration of OBs.

Literature	Sector	Objective	Spatial variation	Scale	Method
Chiou et al., 2011	Residential	Energy use	Space use	Whole & sub-house	Bootstrap sampling
Tabak, 2009	Office	Building performance simulation	Movement	Building	USSU system
Mohammadi & Taylor, 2017	Residential	Electricity demand	Mobility patterns	City	Autoregressive models
Dziedzic et al., 2020	Residential & office	Occupant	Movement	Room	Segregation technique and scanning method
Nassar & Elnahas, 2007	–	Occupant	Movement	Open space	Random access measure
Kleinebrahm et al., 2021	Residential	Occupant	Mobility behavior	–	Neural networks
C. Wang et al., 2011	Office	Occupant	Movement	Building	Markov chain method
Feng et al., 2015	Office	Occupant	Occupancy	Room & building	Software module
Hoes et al., 2009	Office	Building performance assessment.	Space use	Room	USSU and occupancy control models
Deng & Chen, 2019	Office	HVAC	Occupancy	Room	Behavioral artificial neural network model
Kim & Cha, 2019	University	Occupant	Spatial choice	Building	Empirical validation: Brier scores and t-test

Table 1-4. Continued.

Literature	**Sector**	**Objective**	**Spatial concern**	**Scale**	**Method**
Ibrahim et al., 2020	Residential	Thermal	Space use	Room	Multi-nominal techniques and regression
An et al., 2017	Residential	Cooling	Occupancy	District	Stochastic method and DeST
Ueda & Mita, 2015	Residential	Lighting	Spatial choice	Room	Regression and homeostatic function
Shahzad et al., 2019	Office	Thermal	Spatial preference	Room	Visual thermal landscaping and Pearson correlation
Nguyen et al., 2020	Residential	Architectural parametric design	Social-spatial processes	Building	Prototype agent-based model
Carlucci et al., 2016	Residential	Energy performance	Occupancy	Neighborhood	Optimization algorithm and statistical test
Marín-Restrepo et al., 2020	Office	Occupant	Spatial factor	Room	Chi-square test, logistic regression
Okada et al., 2020	Residential	Occupant	spatial factor	Community	Logistic regression and segmentation
Halleck Vega et al., 2022	Residential	Energy transition efficiency	Spatial factor	Regional and neighborhood	Empirical analysis
Rafiee et al., 2019	Residential	Heating	Spatial fator	Housing unit	Regression
Abbasabadi et al., 2019	–	Energy use	Spatial factor	urban	Advance machine learning methods

1.5.3 Studies related to spatial analysis and modeling in other fields

Disciplines associated with the fields of epidemiology, environmental meteorology, and econometrics have applied sound spatial analysis methods to solve subject-specific problems. Epidemiological studies that analyze how health objectives are related to risk factors that vary geographically or predict the spatial spread range of infectious diseases are becoming increasingly popular (Wang, 2012; Zhu et al., 2016). The prediction of the diffusion of air pollutants and the prediction of precipitation or other meteorological phenomena for the unmeasured areas have always been the research hotspots in the environmental meteorology field (Degré et al., 2015; Monestiez et al., 2001; Xie et al., 2017). A lot of econometric research focused on investigating how the fluctuations or changes of the social-economic items of interest such as the wage or house price (Chasco et al., 2007; Murakami et al., 2017) varied by spatial area. This chapter summarized such methods used to either empirically represent the spatial variation or simulate the research objective with the consideration of the spatial variation. Figure 1-2 shows the summary of the methods.

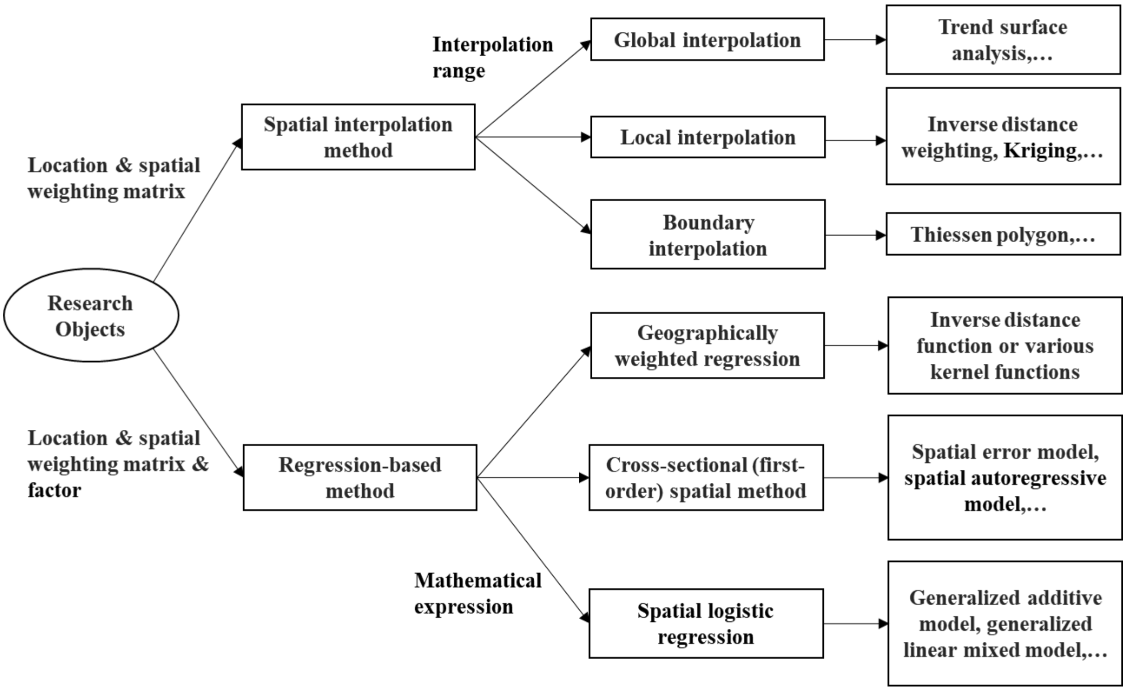

Research Objects
Location & spatial weighting matrix
Spatial interpolation method
Interpolation range
Global interpolation
Trend surface analysis,…
Local interpolation
Inverse distance weighting, Kriging,…
Boundary interpolation
Thiessen polygon,…
Location & spatial weighting matrix & factor
Regression-based method
Geographically weighted regression
Inverse distance function or various kernel functions
Cross-sectional (first-order) spatial method
Spatial error model, spatial autoregressive model,…
Mathematical expression
Spatial logistic regression
Generalized additive model, generalized linear mixed model,…

Figure 1-2. Summary of the methods for spatial analysis and modeling.

Based on the mechanism and data input, the methods used in these studies can be classified as spatial interpolation and regression-based methods. Spatial interpolation methods simulate the spatial autocorrelation of surrounding observations to represent the spatial trend of the objectives or to generate spatial predictions for unmeasured areas. Based on the interpolation range, models can be further divided into global (Nath, 2014), local (Lu & Wong, 2008; Oliver & Webster, 2007), and boundary (Faisal & Gaffar, 2012) spatial interpolation models. The global interpolation model uses all observations to conduct the feather fitting for the whole study area. The typical method is the trend surface analysis. The local interpolation model uses the limited observations within a defined neighboring area to build the mathematical function that can reflect the changes in this neighboring space. The typical method is the inverse distance weighting interpolation method and the kriging interpolation method. The boundary interpolation model assumes that objectives within the boundary are the same (i.e. uniform and homogeneous), changes only occur on the boundary of the region. The typical method is the Thiessen polygon method. Olaf Berke, 1999 applied the trend surface analysis and universal kriging to simulate acid-precipitation in Lower Saxony. Olaf Berke, 2001 also developed the modified median polish kriging method to generate more robust spatial predictions for Wolfcamp-Aquifer. Varouchakis, 2021 applied median polish kriging and sequential Gaussian simulation to explore the spatial distribution of source rock data in terms of total organic carbon weight concentration.

In regression-based methods, they incorporate additional factors, such as sociodemographic factor variables, into the modeling process. According to the mathematical expression, these regression-based models can be divided into Geographically Weighted Regression (GWR), cross-sectional (first-order) spatial model, and logistic spatial model three categories. The mathematical expression of GWR is similar to the conventional regression model, however, the calculation of the regression coefficients is different which involves the information of the locations (Chasco et al., 2007; Mcmillen, 1996; McMillen & McDonald, 1997):

$$y_{s_i} = \beta_{s_i,0} + \sum\nolimits_{k=1}^{m} \beta_{s_i,k} * x_{s_i,k} + \varepsilon \qquad \text{1-4}$$

where $\beta_{s_i,k}$ indicates the coefficient for each variable $x_k, (x_1\ , x_2\ , \dots x_{m,})$ and location $s_i, (s_1, s_2, \dots s_{N,})$. Therefore the coefficient β is not a $m \times 1$ dimensional vector but a $m \times N$ dimensional matrix. To estimate the β the weights should be assigned to each objective by different distances to the location s_i:

$$\tilde{\beta} = \left(X' W_{s_i} X\right)^{-1} X' W_{s_i} y \qquad \text{1-5}$$

where $\tilde{\beta} = \left(\tilde{\beta}_{s_i,0}, \tilde{\beta}_{s_i,1}, \dots, \tilde{\beta}_{s_i,k}\right)$ is the vector of estimated coefficient for location s_i and $\mathrm{W}_{s_i} = \begin{bmatrix} w_{s_i,s_1} & \cdots & 0 & 0 \\ 0 & w_{s_i,s_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & w_{s_i,s_N} \end{bmatrix}$ is the weighting matrix which can be defined by the inverse distance function or various kernel functions (Bivand et al., 2021). Chasco et al., 2007 analyzed the spatially varying impacts of some conventional variables, such as unemployment rate and average housing price, on the per capita household income in Spanish provinces based on geographically weighted regression. In the cross-sectional spatial model, the spatial lags can exist in any parameters of the model (Arraiz et al., 2010; Bivand et al., 2021; Kelejian & Prucha, 1998). The mathematical expression of the cross-sectional spatial model is defined as follows:

$$\Upsilon = \pi^T \Upsilon + \beta^T \mathrm{x} + \alpha^T \mathrm{Wx} + \lambda^T \mathrm{Wy} + \mu \qquad \text{1-6}$$

$$\mu = \rho^T \mathrm{M}\mu + \varepsilon \qquad \text{1-7}$$

where y is the dependent variable on objectives, Υ and x are (endogenous and exogenous) independent variables, and μ indicates the disturbance. W and M are defined spatial weighting matrices. α, λ, and ρ are scalar spatial autoregressive parameters. The variables Wx, Wy, and $\mathrm{M}\mu$ are referred to as spatial lags. Various models can be generated if we set different restrictions to the coefficients in Equations 1-6 and 1-7. Such as the spatial error model (let $\pi = \alpha = \lambda = 0$), spatial autoregressive model ($\pi = \alpha = \rho = 0$), and spatial Durbin model ($\pi = \rho = 0$). To solve these cross-sectional spatial models, some assumptions should be pre-defined to simplify the solution process and ensure the uniqueness of the solution. The spatial logistic regression model has a similar mathematical expression to the conventional logistic regression model. The only difference is that it considered the smooth function $g(s_i; \theta)$ parameterized by θ over location s. According to different $g(s_i; \theta)$, the spatial logistic model can derive different models as well (e.g., generalized additive model and generalized linear mixed model) (Paciorek, 2007). Xie et al., 2000 employed spatial logistic regression to obtain the development patterns in regions and to assess the prognostic capacity of the model based on several factors such as population density and availability of usable sites. Paciorek, 2007 compared several models for fitting spatial logistic regression models and suggested that the spectral basis model is the best to provide a good compromise between the quality of fit and computational speed for the estimation of the spatial surface.

1.6 Research gap

As highlighted in Chapter 1.3 and revealed in the review Chapter 1.5.1, there is ample evidence of the significance of the OB models for simulating energy demand. Many of the bottom-up based energy demand models consider the OB model modules accordingly. However, the development framework of the OB is not well documented in previous literature. In particular, the pre-simulation in terms of the selection of the variable used in the model and the selection of the parameter preparation method lack discussion. Considered variables are mainly from the sociodemographic and time factors. Other factors, while potentially important, were ignored. In addition, the majority of the parameter preparation methods are sample distribution or fitted distribution based. These choices of the OB model design result in poor performance in terms of diversity.

Further, while diversity has been recognized as an essential cause resulting in the performance gap, the deeper investigation and the practical application in the OB model have not been implemented yet. In particular, spatial variation in OBs has attracted more and more attention, but little relevant progress has been made. Most of the research only concerns the space use or the mobility of the occupants, the variation in OB over space is not within the scope of their study. Further, proper methods to deal with the spatial variation have been sparsely considered and applied in the energy field although other fields have already developed some robust spatial analysis methods. These insufficient considerations of diversity in OBs can lower the impact of the occupants on building energy performance and create less accurate energy consumption estimations.

1.7 Aims and objectives

Based on the aforementioned research gap, this thesis aims to improve the OB model by analyzing and evaluating the modeling process, especially the underrated pre-simulation process thus providing reference guides for designing the OB model framework. In particular, this thesis intends to develop new OB models that can incorporate spatial variation into the modeling process thereby better enhancing the diversity among simulated occupants in a given space.

Three objectives are studied related to the research aims:

1) Address research questions on which variables should be considered and what is the most appropriate parameter preparation method to improve the pre-simulation process of OB modeling.

2) Address research questions on whether spatial variation and historical change exist in OB, whether variables can represent spatial and temporal variations, and whether the conventional modeling method can reproduce the spatial variation.

3) Address research questions on when spatial variation exists in OB, how can spatial variation in OB be represented quantitatively, and how can spatial methods reproduce spatial variations in OB to develop a new method for OB modeling with consideration of spatial variation.

1.8 Contributions

The overall research studies would be beneficial in numerous ways to the people who are committed to the residential energy sector both in academia and industry.

First of all, one of the contributions of the work connected to this thises is the comprehensive summary of the review literature. As shown in Chapter 1.5.1, the summary of the use of the modeling method, variables, and engine are given. It contributes to existing knowledge on the development of OB models underpinning energy demand simulation. It also contributes to further understanding of the impact of the significant factors on building energy consumption. More importantly, it reveals the significance of pre-simulation, as the existing studies had seldom paid attention to the design of the framework of an OB model or energy demand model with the consideration of the OBs. In addition, based on the review in Chapter 1.5.1, a further investigation and a deeper discussion of the pre-simulation process were conducted. These contribute to explaining the reason for the selection of the engine, model method, and parameter as well as their combination which are rarely explained in the literature thereby providing useful references and model development direction for other researchers.

Furthermore, the most important contribution is the development of the methodology — OB modeling incorporating spatial variation. Previous studies have pointed out the importance of diversity. However, most of the studies just stop at the discussion aspect of how to enhance diversity. According to Chapter 1.5.2, diversity especially for the spatial variation was insufficiently represented based on the actual data in previous studies. Although some studies used spatial factor variables, there is still a lack of modeling methods to better reproduce spatial variation in OBs. To address the research gap, this thesis developed new OB models incorporating spatial variation to better reproduce the spatial variation in OBs. The new OB model broadens the knowledge of diversity and highlights its significance for the engineering field. Moreover, the outcomes of the model will beneficial to engineering and environment professionals to simulate energy demand and design policy advocacy. Further, the present methodology to consider the diversity in this thesis can be investigated not only for modeling OBs but also for modeling and analyzing the adoption of appliances or other objectives of the research.

In short, the contributions of this thesis are highlighting the importance of the pre-simulation process of OB modeling as well as successfully developing new OB models that can consider spatial variation. For the above-mentioned reasons, this thesis provides great contributions toward more advanced OB modeling underpinning the energy demand simulation and moves forward the state-of-the-art in the field.

1.9 Thesis outline

The work consists of seven chapters. Figure 1-3 presents the overall flow of the thesis. As shown by the figure, Chapters 2–4 revolve around the research objectives to fill the research gap mentioned in Chapter 1.6 respectively.

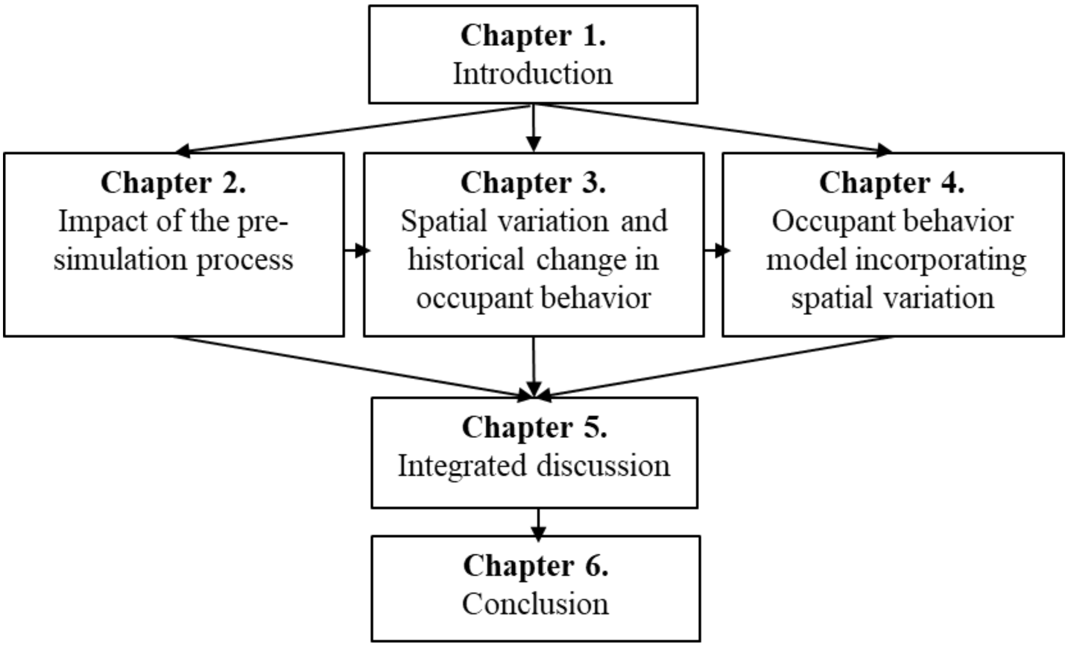

Chapter 1.
Introduction
Chapter 2.
Impact of the pre-simulation process
Chapter 3.
Spatial variation and historical change in occupant behavior
Chapter 4.
Occupant behavior model incorporating spatial variation
Chapter 5.
Integrated discussion
Chapter 6.
Conclusion

Figure 1-3. Thesis structure.

Chapter 1 centers on introducing the energy use and energy demand modeling in the residential sector, the role of OB in energy demand modeling, the development of OB modeling, and diversity in OBs. Then a critical literature review to summarize the methods of OB modeling or energy demand modeling with consideration of OBs, and methods of spatial analysis and modeling in both the engineering field and other fields is given. Based on the review, the research gap was found. Finally, the overall objectives and the framework of research targeting the assessment of OB model performance and development of the model that can involve spatial variation are outlined.

Chapter 2 presents a case study to highlight the importance of the design of the pre-simulation process based on single-year American time use survey (ATUS) data. It covers the two vital sub-processes in the pre-simulation thus evaluating their corresponding impact on OB model outcomes.

Chapter 3 and Chapter 4 present analyses related to spatial variation in OBs based on multiple-year ATUS data. In Chapter 3, the spatial analysis method learned from the geostatistics field was applied to assess the performance of the OB model that was selected based on the results from Chapter 2. The existence of the spatial variation and the historical change is confirmed by the spatial analysis method.

Chapter 4 proposes the research method to develop the new OB models incorporating spatial variation. This chapter includes the theoretical concepts of the spatial analysis and modeling methods reviewed in previous studies, as well as the implementation of the new model (i.e., the data collection, the tool, the evaluation standard).

Chapter 5 presents a combined in-depth discussion of the preceding chapters, giving an overview of the complex nature of the OB modeling systems that were analyzed. The chapter highlights the significance of the design of the OB modeling and put forwards the pre-conditions and appropriate ways to involve spatial variation in the OB modeling process. The limitation and future work are also discussed.

Chapter 6 presents the achievement of research objectives, research conclusions, and research contributions.

2 Impact of the pre-simulation process of occupant behavior modeling for residential energy demand simulations

2.1 Purpose

OB models play an important role in building energy demand modeling. As OBs can control the operation status of energy-consuming appliances, reflect the occupancy status of the rooms or buildings which helps to evaluate the regular energy system, and adjust the indoor environment to meet the needs of the occupants. Useful simulation algorithms for OB modeling have been developed in previous studies as summarized in Chapter 1.5.1. However, previous studies have generally focused on model engines. Less attention has been paid to the pre-simulation process, even though it has a significant influence as analyzed in this chapter. Although existing OB models have used various pre-simulation processes, the reason for choosing a pre-simulation process is not well documented, and alternative methods are rarely compared to improve model performance. To obtain better OB models, the following questions which were mentioned in Chapter 1.7 should be addressed by model developers and users: (1) which variables should be considered, and (2) what is the most appropriate parameter preparation method. None of the previous studies have addressed these questions.

The study in this chapter aimed to provide a reference for addressing the aforementioned two research questions and to improve the pre-simulation process of OB modeling. To this end, this study evaluated how model performance is affected by changes in the pre-simulation processes. Through cross-comparison, this study provided a better understanding of the influences of the selection of variables and parameter preparation methods on OB model performance. The study also provides recommendations for developing improved OB models for different application contexts. Chapter 2.2 introduces the methods and materials used in this study. The results are presented in Chapter 2.3, Chapter 2.4 discusses the findings, and Chapter 2.5 concludes this study.

2.2 Data material and methodology

This study considered the development of a discrete-event model that stochastically generated an activity sequence as the model objective. The model used two modeling parameters: 1) the starting probability of activities and 2) the statistical distributions of activity durations. The activity sequence was stochastically generated by a model engine that repeats two processes: 1) selection of an activity that starts at the first vacant time slot by random selection based on the activity-starting probability, and 2) selection of the duration of the selected activity based on the statistical distribution of the activity duration. An example of this model can be found in the studies of Wilke et al., 2013 and Okada et al., 2020. The time resolution of the OB model was dependent on the unit length of the activity-duration modeling. For example, it was 5 min when the activity duration was modeled using the cumulative probability distribution quantified with a 5-min interval. The studies confirmed that this modeling framework is capable of producing realistic temporal sequences of activities (Wilke et al., 2013) and differentiating them by considering various influencing factors in the modeling of the starting probability and statistical distribution of activity durations (Okada et al., 2020). Although the original TUD can be used as an input for a building energy demand model, the use of the OB model is beneficial when applied to a large number of households; for example, in urban and building stock energy models.

This study only considered the activity-starting probability for the evaluation of the impact of the pre-simulation process and did not include the simulation process using the model engine. The activity-starting probability, which is used in the first process of the activity sequence generation in the discrete-event OB model, represents the composition of the probabilities for selecting each target activity. This study quantified the activity-starting probability within each of the individual 24 h intervals of the day based on a parameter preparation method. The evaluation procedure is illustrated in Figure 2-1.

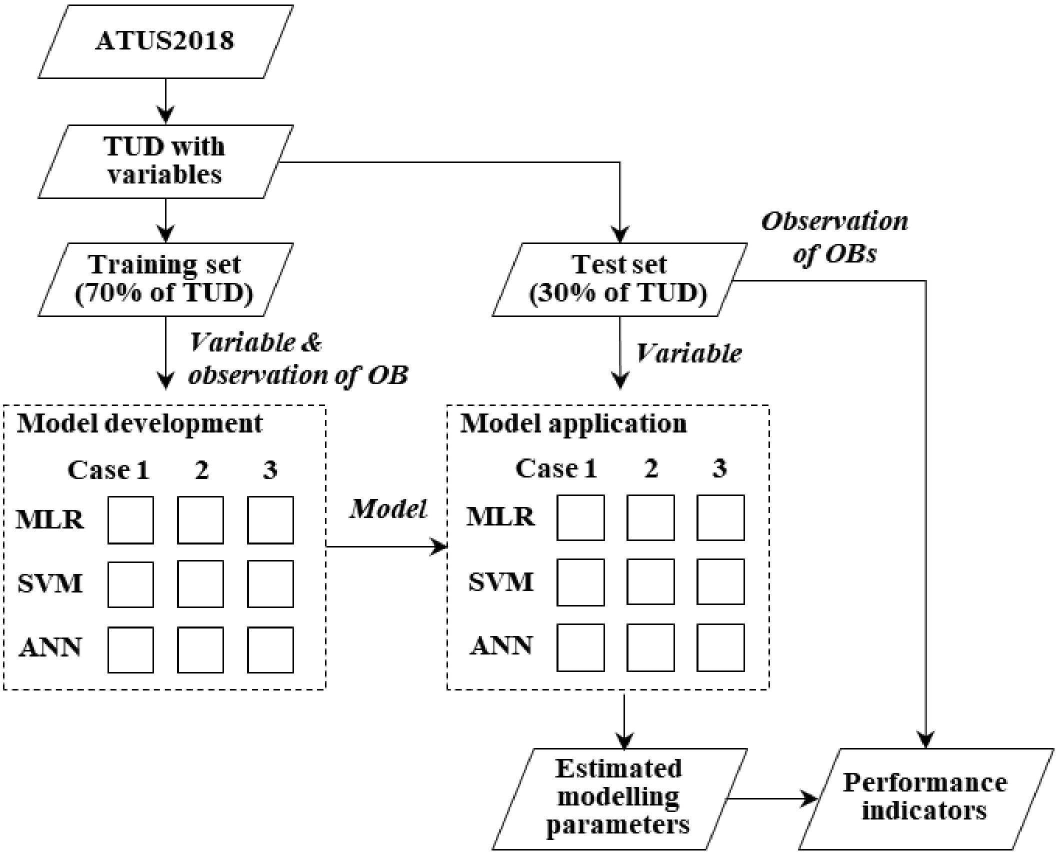


Figure 2-1. Analysis procedure.

This study used the TUD collected from the ATUS in the year 2018. The ATUS is sponsored by the Bureau of Labor Statistics and conducted by the U.S. Census Bureau. The TUD was first randomly divided into training and test sets. The training set comprised 70% of all TUD used to develop models that estimated the activity-starting probability for simulated occupants. We developed nine models based on the training set, combining three evaluation cases and three parameter preparation methods. The evaluation cases, Cases 1–3, were designed to have different combinations of variables considered in the parameter preparation method to evaluate the impact of variable selection on the model performance (explained in Chapter 2.2.2). To evaluate the impact of the selection of the parameter preparation method, we considered three methods (Chapter 2.2.3): 1) a multinomial log-linear regression (MLR), 2) support vector machines (SVMs), and 3) a feedforward artificial neural network (ANN). The remaining 30% of the TUD was used as the test set to evaluate the performance of the developed models for validation. In the validation, the nine developed models were applied to the test set to quantify several performance indicators, and the performance was cross-compared based on well-designed indicators (Chapter 2.2.4).

2.2.1 Data

The ATUS collected time use diaries for 24 h beginning at 4:00 on a survey day from 9,370 individuals. The diary contains activity codes representing the activity performed and the times at which the activity started and ended. Table 2-3 in Appendix A. ATUS data record shows an example of the ATUS data. The ATUS data use 18 major activity categories with hundreds of subcategories. The ATUS data contain the identification number used in the current population survey data that contain the demographic attributes of individuals. Using the identification number, demographic attributes are attached to the activity data.

For the modeling, we converted the activity code in the ATUS data into 25 activities listed in Table 2-1 such that each category had similar appliance usage, and the activity locations could be grouped as indoor or outdoor. It should note that the text in the brackets is the abbreviation of the corresponding activity. Activities 1–10 and 14–21 were indoor activities. Activities 11–13, 22, and 23 were outdoor activities. Activity 24 is an unspecified personal or private activity performed at an unspecified location. Activity 25 involves activities that are missing for various reasons (for example, survey participants refused to provide information, and an activity code could not be assigned). Based on these features, we classified the activities into activity clusters C1–C5 as listed in Table 2-1 to be referred to in the results chapter. C1 includes basic life activities such as sleeping, eating, drinking, and personal care activities that occur indoors. C2 contains indoor housework activities. C3 contains work and study activities. C4 and C5 contain other indoor and outdoor activities, respectively.

Table 2-1. Activities and codes used in this study.

Cluster	Code	Activity	Cluster	Code	Activity
C1	1	Eating and drinking (Eating&D)	C4	14	Computer
	2	Personal care		15	Telephone
	3	Sleeping		16	Television
C2	4	Laundry		17	Household and personal management and planning (Plan&M)
	5	Caring		18	Leisure and hobby (Leisure&H)
	6	Housework		19	Sports
	7	Food preparation and presentation (FoodP&P)		20	Religious, volunteer, and civic activities (Religious&VC)
	8	Kitchen and food clean-up (Kitchen&FC)		21	Shopping and using services (Shopping&S)
C3	9	Paid work or job (Work&J)	C5	22	Appliances for outside (Appliances_O)
	10	Studies, school work, and research (Studies&WR)		23	Other outside activities (Other act_O)
	11	Paid work or job outside (Work&J_O)	-	24	Personal activities (Personal act)
	12	Studies, school work, and research outside (Studies&WR_O)	-	25	Missing
	13	Commuting and school (Commute&S)			

As this study quantified the activity-starting probability using a 1 h interval, the activity records were classified into 24 groups based on the clock time as threshold values. For each group, the parameter preparation method was applied independently. However, we combined the time intervals from 0:00 to 5:59 to ensure that the events per variable (the number of activity records of each independent variable) was 10 or larger (Concato et al., 1995) for each activity. Figure 2-2 shows the number of observations in each time interval. As shown, the sample size at intervals from 1:00 to 4:59 was smaller than 1,000. When the parameter preparation methods were applied to this interval, five dummy variables representing each of the intervals from 0:00 to 4:59 were considered with 5:00–5:59 as the reference category. Appendix B evaluated the influence of the combined time intervals.

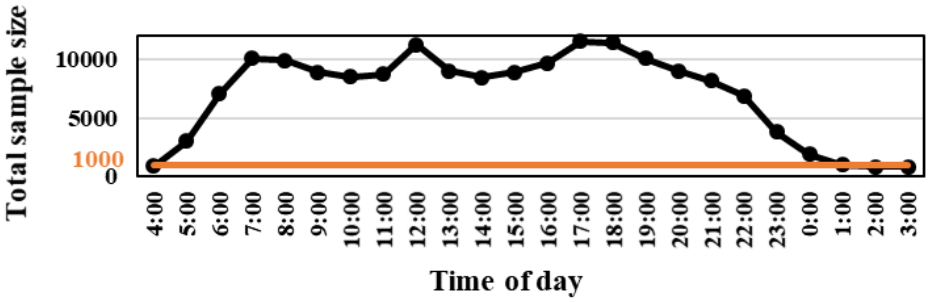


Figure 2-2. Number of observations at each time interval.

2.2.2 Case design

As shown in Figure 2-1, we designed three cases (i.e., Cases 1–3) characterized by the number and type of selected variables considering three levels of variables. The first level comprised the eight influencing factors explained in Chapter 1.5.1. The second level comprised variables included in the TUD, such as age. The third level consisted of independent variables input during the application of the parameter preparation methods. For example, we created three dummy independent variables representing young people 10–29 yrs, middle-aged 30–59 yrs, and seniors 60 yrs or older based on the age variable. Table 2-4 in Appendix C. Variables considered in the examined cases lists the variables and independent variables considered in this study.

Table 2-2 lists the type of variables, the number of variables, and the independent variables considered in each case. Case 1 contained the fewest variables and considered only the basic variables of the demographic and time factors. These six basic variables have been commonly used in existing OB models (Anderson, 2016; Diao et al., 2017; Fischer et al., 2015; Okada et al., 2020; Wilke et al., 2013).

Case 2 contained the variables considered in the models of Wilke et al., 2013 and Okada et al., 2020. Case 2 newly considered the ownership of the housing unit in the demographic factor and the metropolitan status in the geographical factor. The number of variables representing the demographic, time, and geographical factors increased to 14 and the number of independent variables increased to 26.

Case 3 assumed a situation in which as much available information as possible was considered in the modeling to include those rarely used in existing OB models and considered the variables in the remaining three influencing factors. The employment status of a spouse was from the demographic factor, the ownership of a telephone was represented as the appliance factor, and the other four variables, such as the type of person accompanying an occupant, were represented by the activity factor. The number of independent variables was 67.

Table 2-2. Number and type of variables considered in cases. The definition of the variables is listed in Table 2-4 in Appendix C.

Item		Case 1	Case 2	Case 3
Demographic	Individual	Age, gender, education, and employment	Variables considered in Case 1, occupation, health, and race	Variables considered in Case 2 and student
	Household	Num_people	Num_people, children, and income	Num_people, children, and income
	Housing		Housing	Housing
	Other members			Spouse employment
Time		Diary day	Diary day and holiday	Diary day and holiday
Activity				Time_care, num_people accompany, type_people accompany, and previous activity
Geographic			Metropolitan	Metropolitan and region
Appliance				Telephone
Number of variables / independent variables		6 / 8	14 / 26	22 / 67

2.2.3 Methods of parameter preparation

The activity-starting probability was modeled using three parameter preparation methods: 1) a MLR following (Wilke et al., 2013), 2) SVMs with a Gaussian radial basic kernel function (Jiawei Han, Micheline Kamber, 2014), and 3) a feedforward ANN with a backpropagation algorithm and one hidden layer. Generally, for an ANN, a single layer with an optimal number of neurons is sufficient for many practical problems (Goh, 1994; Rafiq et al., 2001). The number of neurons in the hidden layer was determined using $m = \log_2 n$, where m is the number of neurones in the hidden layer, and n is the number of neurones in the previous input layer, which should be between the number of input and output neurones (Sheela & Deepa, 2014).

Our intention was not to find the best method for modeling activity-starting probability but to evaluate how model performance changes with the selection of the method. Thus, we chose MLR because it has been widely used in OB modeling and is easy to develop. SVM and ANN were selected as potential alternatives capable of dealing with nonlinear relationships.

We applied the stepwise method to select variables for MLR, as in most previous studies. All of the variables were used to develop the models using the ANN because it required large-scale data and did not require feature extraction. For the SVM, we applied stepwise and LASSO regression

(Ranstam & Cook, 2018) in addition to the model using all variables to determine whether feature extraction was required for the TUD.

2.2.4 Model performance assessment

We assessed the model performance using five performance indicators related to three aspects. The first aspect was the reproducibility of the average activity-starting probability, which is crucial for obtaining a realistic average energy demand. To evaluate the reproducibility, we used the root mean squared error (RMSE) defined as

$$\text{RMSE} = \sqrt{\frac{\sum_{t=1}^{T}\sum_{m=1}^{M}\left(\varepsilon_{t,m}\right)^2}{T * M}} = \sqrt{\frac{\sum_{t=1}^{T}\sum_{m=1}^{M}\left(p_{obs_{t,m}} - p_{est_{t,m}}\right)^2}{T * M}} \qquad \text{2-1}$$

where RMSE quantifies the average error between observation and estimations for all activities and time intervals, t is the time interval (T = 24), m is the activity (M = 25), p is the activity-starting probability, and ε denotes the error.

The second aspect was diversity, which assesses how well the model represented the variation in OBs among the simulated occupants. We used two performance indicators. The first was the mean standard deviation (MSD) calculated as

$$\text{MSD} = \frac{\sum_{t=1}^{T}\sum_{m=1}^{M}\left|SD_{est_{t,m}}\right|}{T * M} \qquad \text{2-2}$$

where $SD_{est_{t,m}}$ is the standard deviation of the estimated probability among the sample for activity m at t. MSD measures the average amount of variation or dispersion of the estimates for all combinations of activities and time intervals.

The weakness of MSD is that it does not quantify the goodness-of-fit with the observations. To overcome this weakness, we considered a second indicator based on the Hosmer–Lemeshow test which is often used to evaluate the goodness-of-fit in logistic regression models. In the test, the samples were divided into several groups after sorting the samples according to the estimated probability from the lowest to the highest. Then, the statistical difference in the probability of each group was tested between the estimation and observation. However, this method is not effective when the occurrence of the model objectives is low (Paul et al., 2013) and it is inapplicable to activities with a low starting probability. Therefore, we designed an indicator that measured the RMSE between the averaged estimated probability and averaged probability of observations of subgroups as in the Hosmer–Lemeshow test, named RMSE_GA and calculated as

$$\text{RMSE\_GA} = \sqrt{\frac{\sum_{t=1}^{T}\sum_{m=1}^{M}\sum_{g=1}^{G}\left(Mean_{t,m,g}(P_{est}) - Mean_{t,m,g}(P_{obs})\right)^2}{T * M * G}} \quad 2\text{-}3$$

where g indicates the subgroups created based on the estimated probability of the test set. For activity m at t, we sorted the test set according to the estimated probability and then equally divided it into 10 subgroups (G = 10 as often used in the Hosmer–Lemeshow test) for all methods. This indicator quantified the difference in the distribution of observation and estimation, thereby assessing diversity.

The third aspect was the reproducibility of individuals' activities; that is, individual specificity. Individual specificity is important to accurately predict an individual's activities. The accuracy and F1 score were used to evaluate individual specificity, respectively defined as

$$\text{accuracy} = \frac{1}{T}\sum\nolimits_{t=1}^{T}\frac{N_{pred_t}}{N_{total_t}} \quad 2\text{-}4$$

$$\text{F1 score} = \frac{1}{T}\sum\nolimits_{t=1}^{T}\frac{2\ \text{Precision}_t * \text{Recall}_t}{\text{Precision}_t + \text{Recall}_t} \quad 2\text{-}5$$

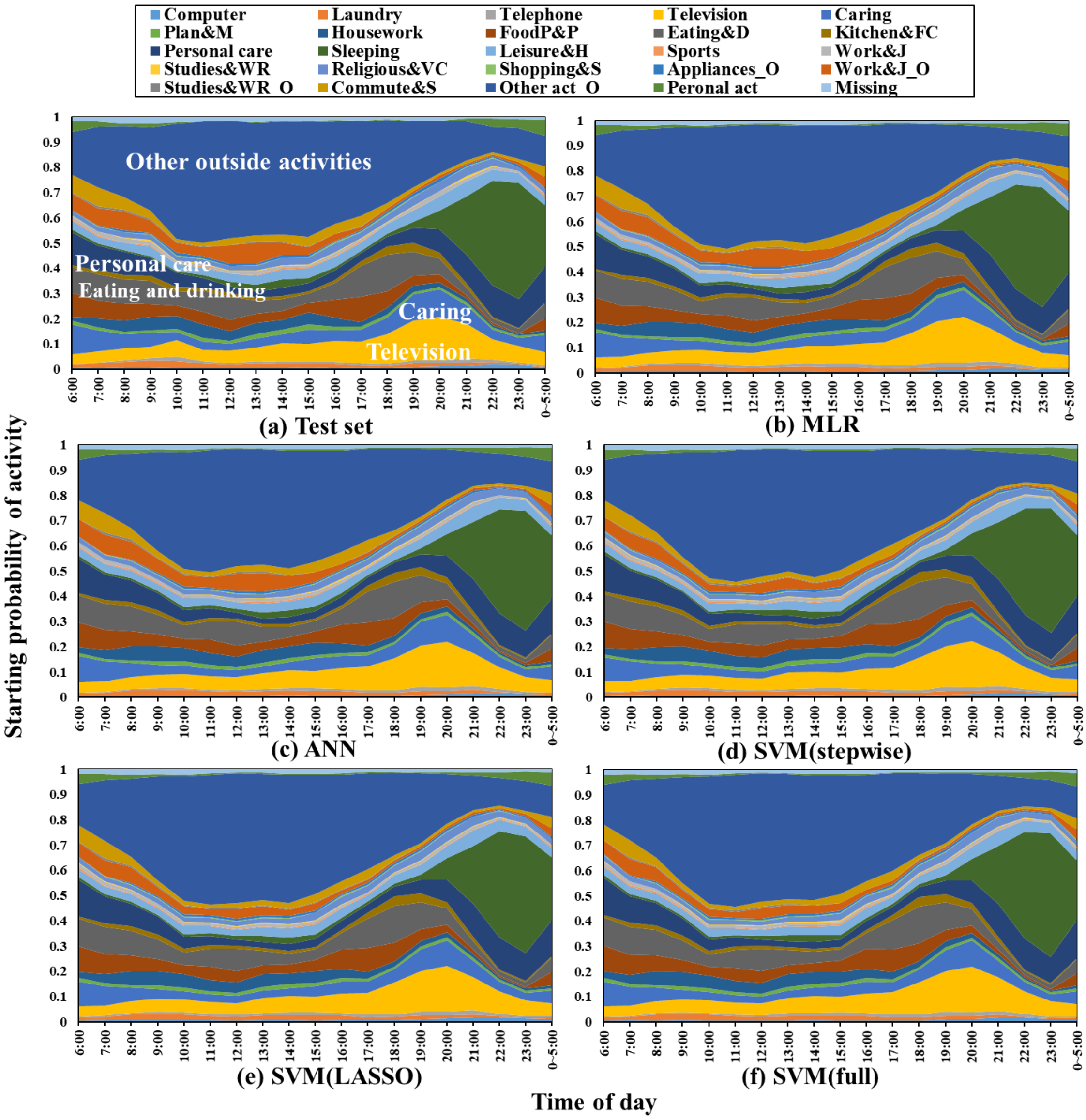
where N_{pred_t} indicates the number of correctly predicted cases at t. The accuracy measures the percentage of all correctly predicted cases. For prediction, the activity with the highest starting probability was selected. The precision and recall values of the F1 score were extracted from the confusion matrix. The F1 score measured incorrectly classified cases, which is an important metric when misprediction is costly.

2.3 Results

2.3.1 Average performance

Figure 2-3 (a) shows the sample distribution of the starting probability in the test set. Figure 2-3 (b)–(f) show the average of the activity-starting probability estimated by the three methods of Case 1. No evident differences were observed with the sample distribution. This result indicates that all of the models reproduced the average starting probability of the test set in Case 1. Similar results were obtained for Cases 2 and 3.

Figure 2-4 shows that the RMSE for all developed models was less than 1%. This scale of error is much smaller than that originating from the simulation processes of OB (in the middle of Figure 1-1) and building energy demand, which generally has an RMSE error greater than 10%. For example, the error of Yamaguchi & Shimoda, 2017 was 10%–20% for multiple activities. The error of Naspi et al., 2018 for the window-closing action was 15%. The RMSE for modeling the energy use of appliances in a low-energy house in the work of Candanedo et al., 2017 was greater than 65%. The most notable result is that the MLR and ANN had lower RMSE values than that of the SVMs. Case 3 had the smallest RMSE among all methods. The improvement in Case 2 from Case 1 was limited to the MLR and ANN. The effect of the SVM was significant when stepwise and LASSO regression were adopted for the selection of independent variables.



(a) Test set

(b) MLR

(c) ANN

(d) SVM(stepwise)

(e) SVM(LASSO)

(f) SVM(full)

Figure 2-3. Starting probability of activities at the time of day. Each colored area indicates the probability of an activity indicated by the graph legend (Table 2-1). Figure (a) shows the proportion of activities observed in the test set. Figures (b)–(f) show those estimated for the test set in Case 1.

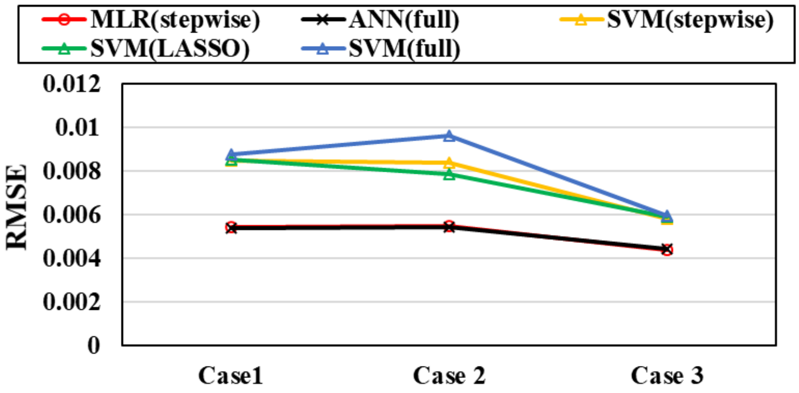


Figure 2-4. RMSE of developed models.

2.3.2 Diversity performance

As shown in Figure 2-5, the MSD for all developed models was smaller than 7%. The MSDs of MLR and ANN were higher than those of the SVMs. The MSD of Case 3 was higher than that of Cases 1 and 2, and a small improvement was observed from Case 1 to Case 2. This result indicates that the newly considered variables in Case 3 enhanced the diversity.

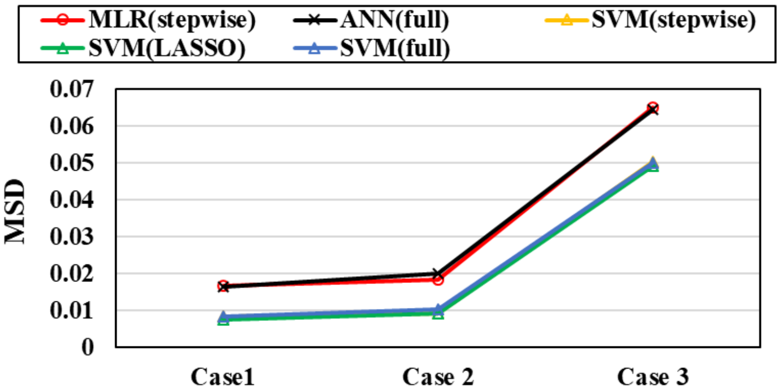


Figure 2-5. MSD of developed models.

Figure 2-6 shows the sorted probability of watching TV, kitchen, food clean-up, and sleeping at the representative hours of the day estimated by the MLR in Case 3 (red lines). The figure also shows the averaged probability in the 10 subgroups made by the MRL in Case 3 as explained in Chapter 2.2.4. The black line indicates the observations of the test set. All of the methods in Case

3 fit well with the averaged probabilities regardless of the activity. Compared to Case 3, the averaged probabilities of Cases 1 and 2 did not fit well with the subgroups in the test set. The most obvious example was sleeping, which ranged from 41% to 53% for all subgroups in Case 1 and ranged from 43% to 54% in Case 2.

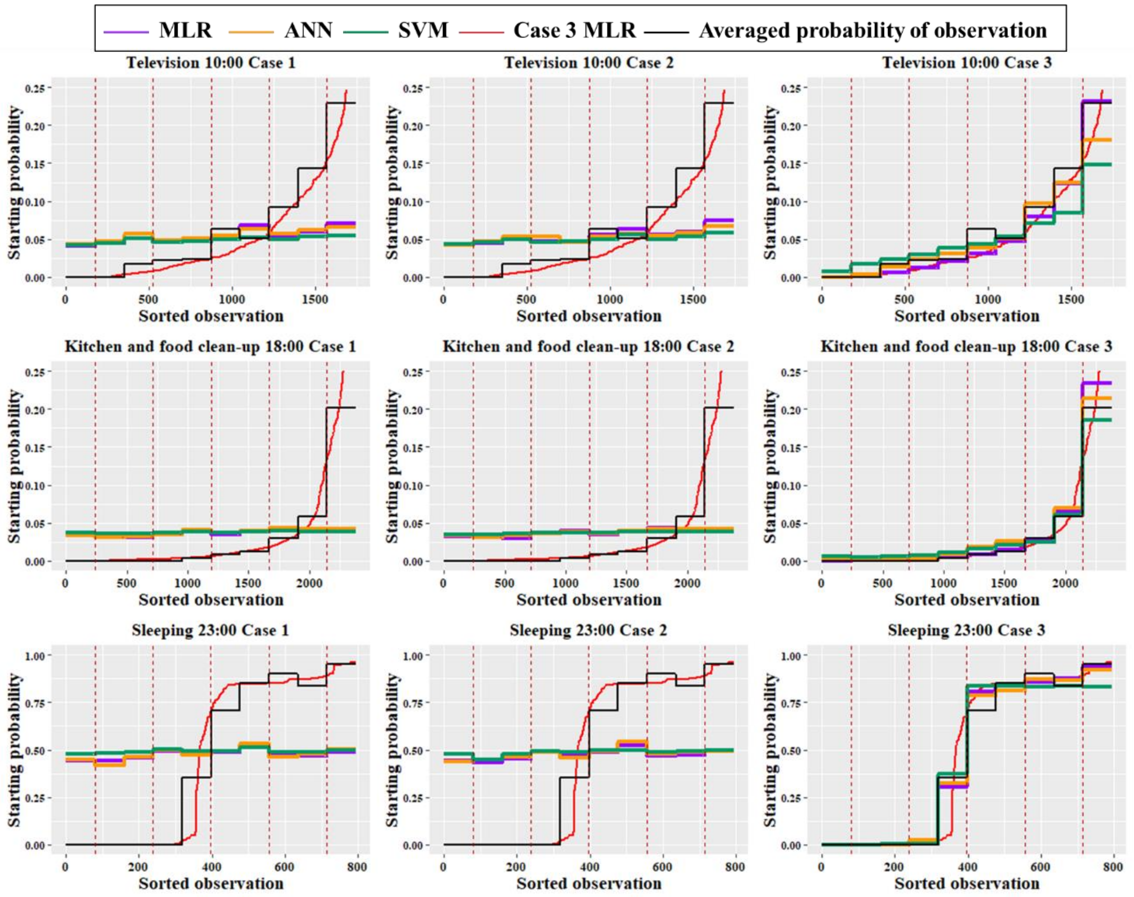


Figure 2-6. The activity-starting probability of the 10 subgroups created by sorting the samples based on the probability estimated by the Case 3 MLR-based model are indicated by the red line. The nine figures indicate the result of watching television, kitchen and food clean-up, and sleeping at the representative time intervals of 10:00, 18:00, and 23:00, respectively. The horizontal axis indicates the number of sorted samples. The black line indicates the average of the observation. The other colored lines indicate the results of the models.

However, this result did not indicate that Cases 1 and 2 did not fit the test set. Figure 2-7 shows the same results using the subgroups created based on the probability estimated by the MLR in Case 1. All the models fit well with the averaged probabilities of the subgroups in the test set shown in Figure 2-7. However, the ranges of the starting probability of all three representative activities were smaller than those of Case 3 in Figure 2-6, which is consistent with the MSD results. This result indicates that less diversity was produced in Cases 1 and 2 compared to that of Case 3.

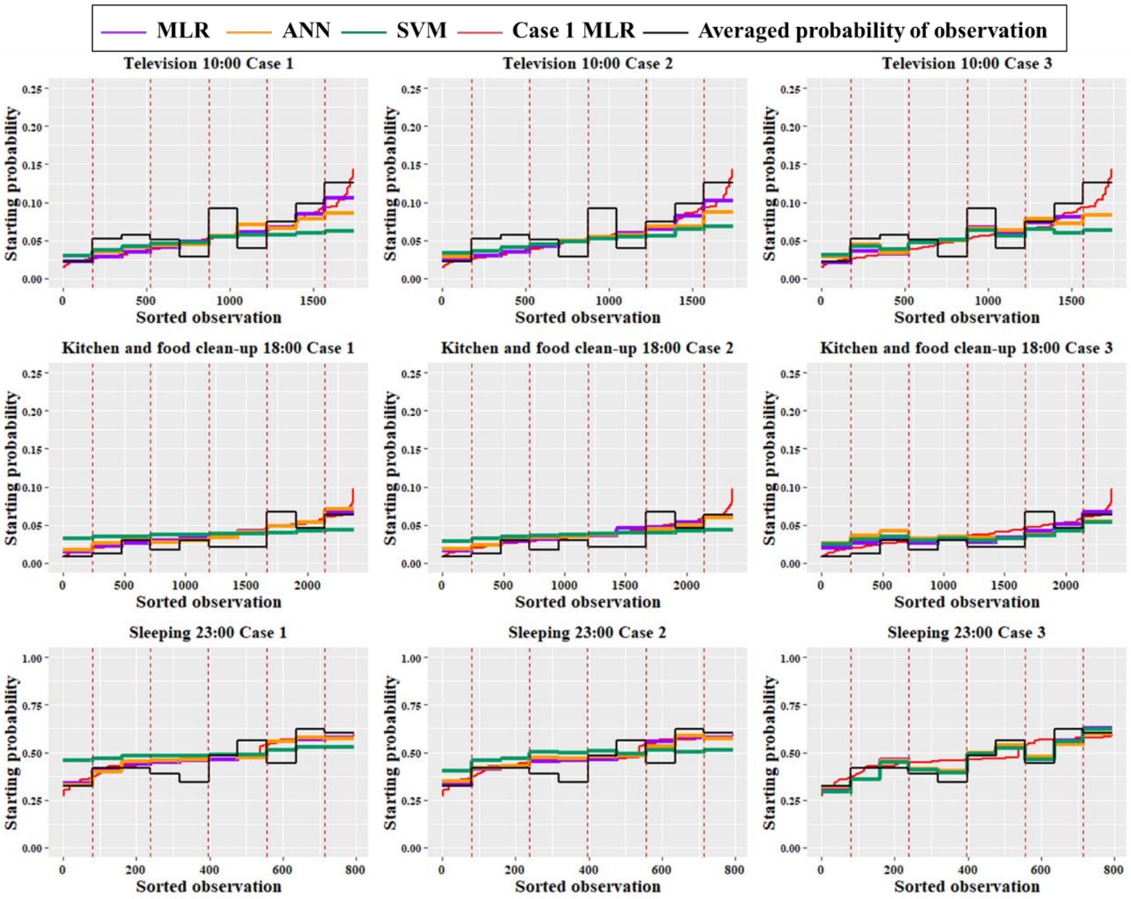


Figure 2-7. Activity-starting probability of the 10 subgroups created by sorting the samples based on the probability estimated by the Case 1 MLR-based model (indicated by the red line).

Figure 2-8 shows the base-10 logarithms of the estimated and observed probabilities of the test set for all combinations of activities, time intervals, and subgroups. The subgroups were divided based on Case 3 MLR. According to Figure 2-8 (a) and (b), the distribution was scattered owing to the error as shown in Figure 2-6. The figures present two R^2 values obtained with and without the transformation of the base-10 logarithm (R^2_{log} and R^2, respectively). The R^2 was 0.48 for the MLR and ANN in Case 1 without logarithmic transformation, and 0.36 and 0.35, respectively, with the transformation. Figure 2-8 (c) and (d) indicate that the Case 3 models fit the test set well, although there were discrepancies in the range with probabilities less than 0.01. These discrepancies are acceptable as the probabilities were very small because the figures are shown as a base-10 logarithm. R^2 was 0.99 for the MLR and ANN in Case 3 without logarithmic transformation, whereas the R^2_{log} values with the transformation were 0.48 and 0.73, respectively.

Therefore, the Case 3 models were more capable of enhancing the diversity range and reproducing the distribution of probability among the simulated occupants differentiated by the considered variables. This result was confirmed by the RMSE_GA.

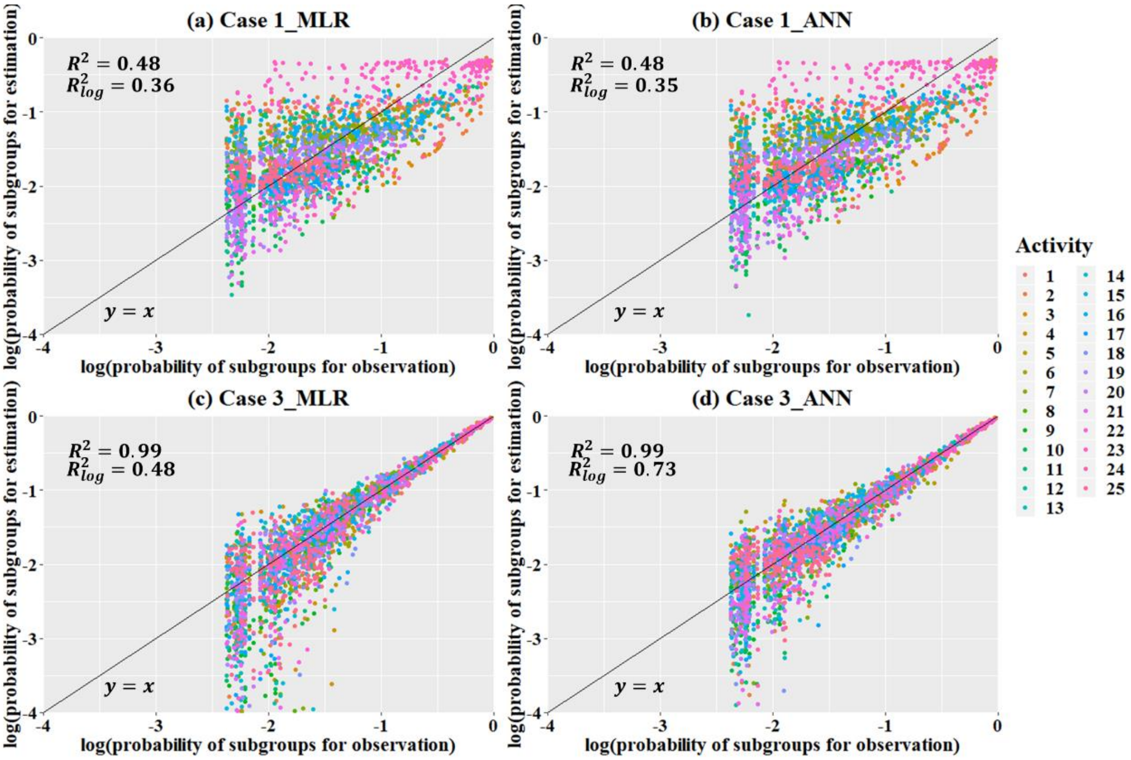


Figure 2-8. Averaged starting probability of activities estimated and observed in the subgroups created based on the estimation of the Case 3 MLR-based model. The horizontal axis shows the observed probabilities, and the vertical axis shows the estimated probabilities. The black line indicates the reference line $y = x$. Each dot indicates a combination of the activity, time interval, and subgroup. Logarithmic transformation was conducted in the range (−4, 0) × (−4, 0).

Figure 2-9 shows the RMSE_GA results, including all developed models with the subgroups created according to the estimated probabilities in Case 3. We found that all methods had the smallest RMSE_GA in Case 3, particularly the MLR. These results further demonstrate that Case 3 had higher diversity and better characterized the probability distribution in the test set. Cases 1 and 2 had similar RMSE_GA results, implying that the newly added variables in Case 2 did not enhance the diversity.

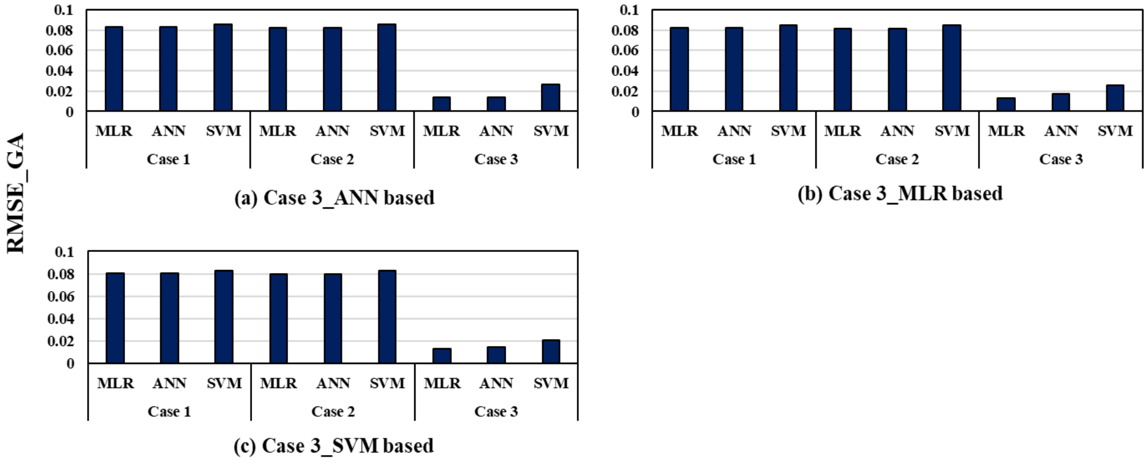


Figure 2-9. RMSE_GA for developed models when subgroups were divided based on Case 3.

2.3.3 Individual specificity performance

Figure 2-10 shows the results of the individual specificity performance. All of the parameter preparation methods had similar accuracies, and Case 3 had the highest accuracy (53% for MLR and SVM, 52% for ANN). Cases 1 and 2 exhibited similar accuracies. However, the accuracies of all developed models were less than 60%, which is smaller than that in other fields that conduct multi-classification (Silva-Palacios et al., 2017). The F1 score showed similar results in Cases 1 and 3. The SVMs had a higher F1 score than the ANN and MLR, particularly in Case 2. The F1 score deteriorated from Case 1 to Case 2 for the MLR, ANN, and stepwise SVM. Combining the two indicators, Case 3 had better individual specificity performance than the other cases.

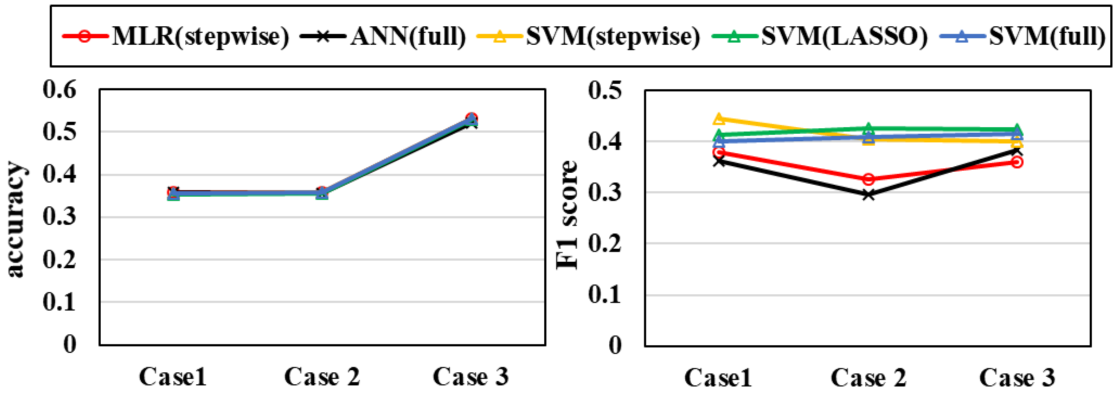


Figure 2-10. Accuracy and F1 score of developed models.

2.3.4 Significant variable

In this subchapter, the variables with a significant influence on the activity-starting probability are analyzed based on the MLR in Case 3, which provided the best performance for all three aspects of model performance. We applied the Wald test to calculate the P-values of the regression coefficients in the MLR model and evaluated the independent variables as to be significant when P-value < 0.05. A significant influence of a variable on target activity was recognized when the regression coefficient of one or more independent variables of the variable is significant.

Figure 2-11 shows the combination of activities and times of day at which a significance was observed in the six representative variables which were widely considered in previous studies — gender, diary day (weekends or weekdays), employment status, presence of children, number of people, and student status. The gender variable showed the significance of indoor activities in activity clusters C2 and C4 at most time intervals. The dairy day variable was significant for most of the activities during the daytime from 6:00 to 18:59. The employment status variable was significant in most of the time intervals, except those from 20:00 to 23:59 for most activities, particularly activities in C3. The presence of children, number of people, and student status variables were only significant for several activities in limited time intervals. Therefore, the significance of these commonly used variables differed depending on the activity and time of day.

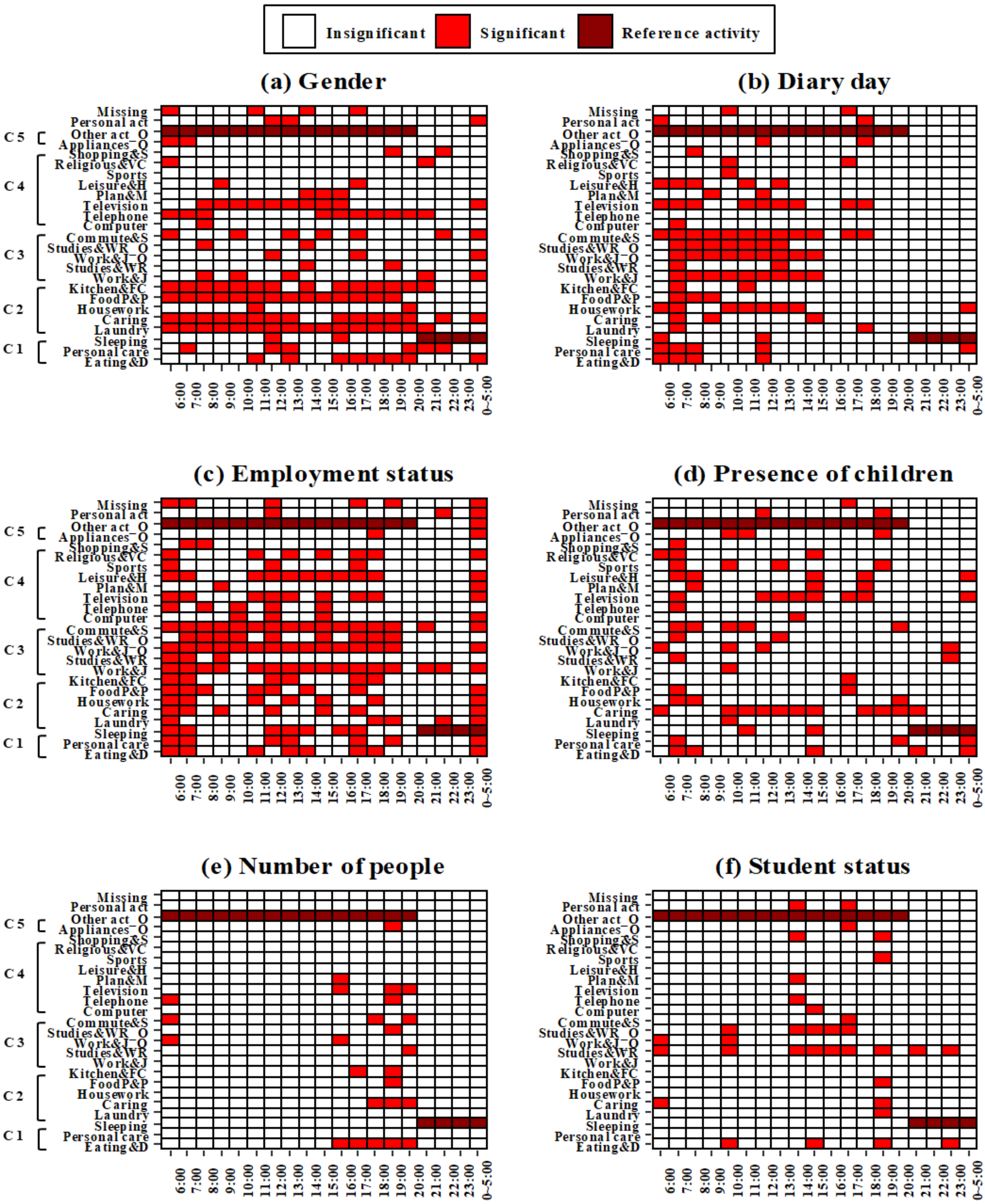


Figure 2-11. Significance of representative variables for a combination of activities (shown vertically) and time intervals (shown horizontally). Red cells indicate that the variable was significant in the combination, and dark red cells indicate an activity considered as a reference category in the MLR for which the significance of the variables could not be obtained.

Figure 2-12 shows the significance of all variables on the horizontal axis for the activities shown on the vertical axis. We identified the following significant relationships:

- The variables in the demographic factor for both individuals (variable ID 1, 2, and 4 on the horizontal axis) and the household (7), time (6), and activity factors (20–22) had a significant influence on the activities in all of the activity clusters C1–C5 because they had large circles. Contrary to our expectations, variables 9 and 17 of the geographic factor were insignificant.

- Activity clusters C1, C2, and C4 included activities conducted indoors. In addition to the variables mentioned above, the other variables in the demographic factor such as health status (12), race (13), and ownership of housing unit (14) showed significance for many activities. The time spent providing secondary care for children younger than 13 years (16) in the activity factor and the ownership of a telephone (18) in the appliance factor also had significance.

- Activity cluster C3 included activities conducted both indoors and outdoors. The work and education variables in the demographic factor such as education (3), occupation (10), health status (12), and student status (19) showed significance.

- Activity cluster C5 included activities conducted outdoors. Except for variables (20–22) in the activity factor, all other variables were significant at limited time intervals for these activities.

The horizontal axis in Figure 2-12 indicates the case in which the variables were included. We observed that most of the variables in Case 1 significantly influenced most activities. However, the circle sizes of the newly added variables in Case 2 were relatively small; therefore, Case 2 had a similar performance to that of Case 1. Case 3 included three highly significant variables in the activity factor: the number of people accompanying (20), type of person accompanying (21), and previous activity (22).

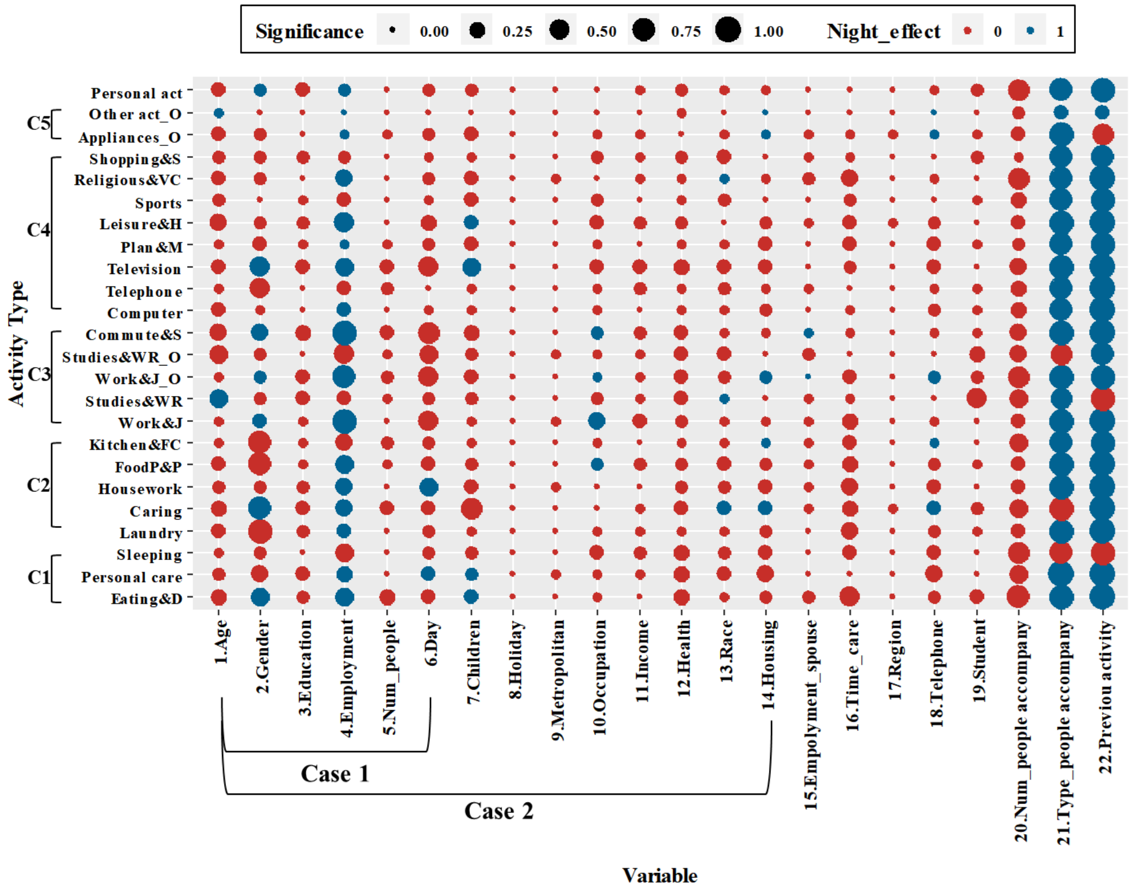
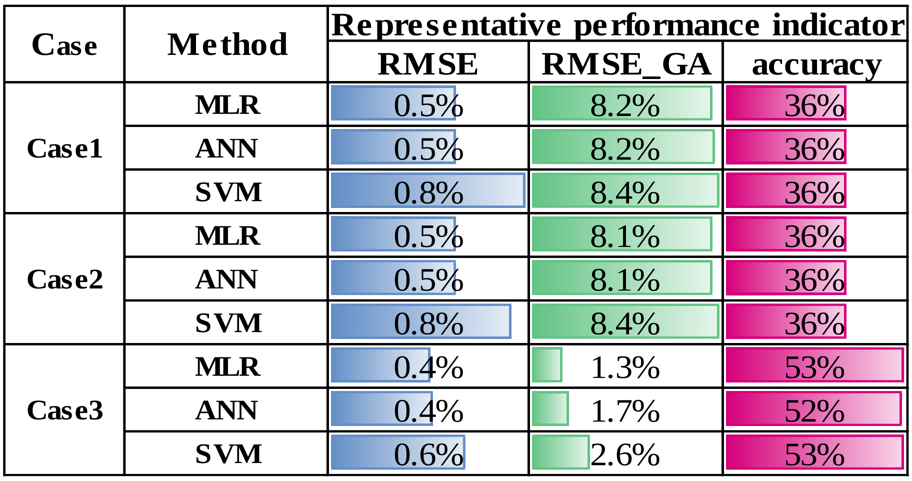


Figure 2-12. Significance of variables for each activity observed in the MLR of Case 3. The horizontal axis lists the variables, and the vertical axis lists the activities in which the clusters were labeled. The circle sizes ranging from 0.0 to 1.0 indicate the number of time intervals in which a variable significantly affected an activity (1 indicates all time intervals). The blue circles indicate that the variable had a significant effect during the period from 0:00 to 5:59.

2.4 Discussion

2.4.1 Summary of results

Figure 2-13 shows a summary of the three examined aspects of model performance for all developed models using representative indicators. Regarding the average performance, the MLR, ANN, and SVM methods had similar results in all three cases in terms of the RMSE. The maximum difference within each case was 0.3%, whereas that among the cases was 0.4%. Regarding the diversity performance, Case 3 had the lowest RMSE_GA, and all of the methods had similar performances in each case. The individual specificity performance represented by accuracy was similar among the methods and was low even in Case 3 with the highest accuracy, 53%.



Case	Method	Representative performance indicator		
		RMSE	RMSE_GA	accuracy
Case1	MLR	0.5%	8.2%	36%
	ANN	0.5%	8.2%	36%
	SVM	0.8%	8.4%	36%
Case2	MLR	0.5%	8.1%	36%
	ANN	0.5%	8.1%	36%
	SVM	0.8%	8.4%	36%
Case3	MLR	0.4%	1.3%	53%
	ANN	0.4%	1.7%	52%
	SVM	0.6%	2.6%	53%

Figure 2-13. Results of representative indicators of developed models.

Therefore, MLR, ANN, and SVM are all acceptable methods for OB modeling when the reproducibility of the average probability is important, regardless of the considered variables. Case 3 considering many variables exhibited the best diversity performance. These results indicate that consideration of important variables enhances the reproducibility of diversity in OB. However, simply increasing the considered variables may not guarantee improvement in performance as observed in Case 2.

Regarding the reproducibility of individual activities, all methods showed poor performance. This result implies that the examined methods cannot deliver a model with high accuracy using ATUS data. This is reasonable because 1) we only assessed the correctly predicted cases (same as the true positive cases in binary classification) since true negative cases cannot be directly obtained by the multi-classification method, and 2) the input variables were not sufficiently detailed to predict individual activities.

2.4.2 Determination of the most appropriate parameter preparation method

As mentioned in Chapter 2.4.1, all of the methods had similar performances and were sufficient for OB modeling in terms of the reproducibility of the average probability and diversity. However, from a more practical perspective, stepwise MLR was useful when only a small number of independent variables were considered as illustrated in Case 1. When considering a larger number of independent variables, ANN was useful because the feedforward ANN with a simple structure obtained similar results to the stepwise MLR with a much shorter run time because the ANN did not require a feature extraction process.

As discussed in Chapter 2.4.1, the examined methods cannot deliver a model with high accuracy using ATUS data. Therefore, other modeling approaches (e.g., Kleinebrahm et al., 2021) should be employed when individual specificity is considered.

In addition, at time intervals during the night that contained a small number of samples, all models showed noticeably larger errors compared to the other intervals because the TUD was unbalanced. Consequently, we recommend combining modeling time intervals or applying advanced techniques such as resampling or bootstrap to generate new reliable samples (Raudys & Jain, 1991) when the sample size is limited.

2.4.3 Determination of variable to be considered

Many previous studies have considered only the basic variables in the demographic and time factors. This approach enables the construction of OB models capable of reproducing the average probability as shown by the RMSE of the average performance. More variables should be included when diversity needs to be reproduced. However, the consideration of more variables does not guarantee an improvement in the model performance as indicated in Case 2. The model performance is improved only when highly significant variables are considered. Therefore, we recommend setting a reference group that includes the basic variables from previous studies to test whether the newly considered variables are worthy of being included in the modeling.

The results in Chapter 2.3.4 indicate that complex relationships exist among the activities, variables, and time of day. For example, the significance of basic variables in the demographic and time factors varied greatly with respect to the activity and time of day. The relationships should be reflected in the variable selection to better express diversity. Although variables in the activity and appliance factors (e.g., the type of person accompanying the occupant(s) and ownership of an appliance) were rarely considered in previous models. These factors have a significant impact on diversity.

2.4.4 Limitation

This study targeted only the activity-starting probability parameter for OB modeling and assumed that the other parameters would have similar results. We developed models with whole samples, although many existing OB models conducted segmentation as indicated in Table 1-1. Part of the bias or diversity among the subpopulations was ignored in this study. In addition, only OB-related performance indicators were used to evaluate the cross-comparisons. Indicators for measuring the influence of the OB model on the energy demand simulation were not included.

Another limitation was that we were unable to test the different cases combined with the different parameter preparation methods on big data because of data limitations. Advanced methods such as multilayer ANNs and other time-series methods may provide better performance (Calis et al., 2017; Kleinebrahm et al., 2021).

2.5 Conclusion

OB models have several modeling parameters (e.g., activity-starting probability) prepared to simulate the occupants in the pre-simulation process. The method used to quantify these modeling parameters for building occupants has a significant impact on the performance of OB models and subsequent energy demand models. However, the impact of the pre-simulation process of OB modeling has received less attention. A literature review of the existing OB models revealed that modeling parameters have been predominantly quantified based on a sample-based approach; i.e. using a sample distribution. Variables considered in the parameter preparation method were limited to basic demographic and time factors, and the selection of the methods and variables was not comprehensively designed. Therefore, this study elaborated on the pre-simulation process of OB modeling and evaluated how the design of the pre-simulation process influenced the average, diversity, and individual specificity performances, whereas previous studies mainly focused only on the average performance. Our analysis results showed that all the considered methods (MLR, ANN, and SVM) effectively reproduce the average activity-starting probability of a population with the basic variables of the demographic and time factors. An increase in the consideration of significant variables contributed to enhancing the reproducibility of diversity. Regarding the reproducibility of individuals' activities, the methods did not perform well, even with many variables. Furthermore, based on these findings, we offer the following practical recommendations for improving the pre-simulation process:

1. MLR with stepwise variable selection is the most practical method for cases in which the number of independent variables in the TUD is small. However, when the number of independent variables is large, the use of ANNs or other data-driven methods is more practical.
2. There is a complex relationship among variables, activities, and the time of day. Representing such relationships contributes to enhancing diversity in OB modeling. For activity modeling, in addition to basic variables in the demographic and time factors, variables in the activity (e.g., previous activity) and appliance (e.g., appliance ownership) factors are significant.
3. It is beneficial to use a reference model with a widely used parameter preparation method that considers basic variables to assess the pre-simulation process.

2.6 Appendix

2.6.1 Appendix A. ATUS data record

Table 2-3. Example of original ATUS records.

Case ID	Age	Start time	End time	Act code	Loc code
20180101180006	4	4:00:00	8:00:00	10101	-1
20180101180006	4	8:00:00	12:00:00	120303	1
20180101180006	4	12:00:00	12:10:00	181101	12
20180101180021	5	4:00:00	10:30:00	10101	-1
20180101180021	5	10:30:00	12:30:00	120303	1
20180101180054	7	10:00:00	10:01:00	70103	7
20180101180096	4	15:00:00	15:15:00	180301	14

Note: the age variable has eight categories; the numbers in bold in the Act code are the main activities of 18 categories. The Loc code has 26 categories for locations and -1 denotes missing values.

2.6.2 Appendix B. Evaluation of the effect of the combination of time intervals

Figure 2-14 shows the results of RMSE, MSD, and accuracy (defined in Chapter 2.2.4) of Case 1. The grey bars show the results of models with time intervals 0:00–5:59 combined using dummy variables indicating each time interval as explained in Chapter 2.2.1. The black bars indicate the models without the combination of time intervals where all of the time intervals were independently modeled. According to the figure, the latter models had a 20% larger RMSE compared to that of the former models. The latter model had a relatively larger MSD; however, the difference from the former model was within 0.8%. The two models had almost the same accuracies. Based on these results, we combined the time intervals in this study.

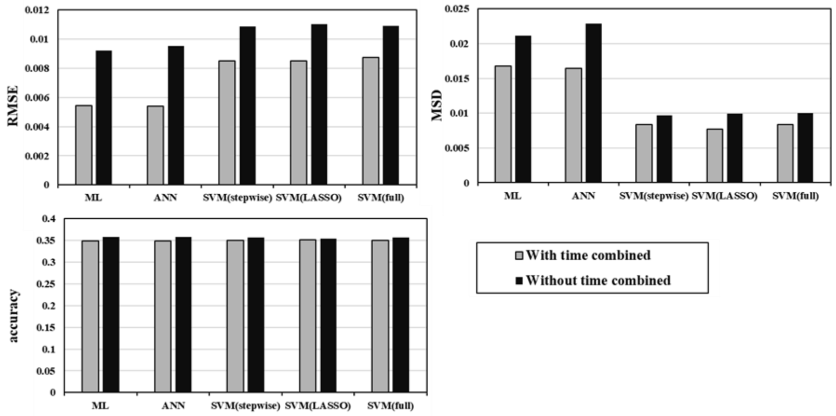


Figure 2-14. RMSE, MSD, and accuracy of methods with and without the combination of time intervals 0:00–5:59 in Case 1.

2.6.3 Appendix C. Variables considered in the examined cases

Table 2-4 lists variables and independent variables considered in Cases 1–3. The independent variables were created based on the variables listed in the second column.

Table 2-4. Variables and independent variables of different cases.

Case	Variable	Independent variables
Case 1	1. Age	Young (10–29 yrs), middle age (30–59 yrs), senior (60+ yrs)
	2. Gender	Male, Women *
	3. Education	Level of education above secondary education
	4. Employment status	Full-time worker *; part-time worker; no work (absent, unemployed, not in the labour force)
	5. Number of people	1–6
	6. Diary day	Weekends*, weekdays
Case 2	7. Presence of household children	No*, Yes
	8. Holiday	No*, Yes
	9. Metropolitan status	Metropolitan*, non-metropolitan, not identified
	10. Occupation	Management, professional, and related; service; sales and office; farming, fishing, and forestry; construction and maintenance; production, transportation, and material moving; no work*
	11. Household income	Lower income (household income below lowest 23%); middle income (50%)*; higher income
	12. Health status	With disability, without disability*
	13. Race	White*; black; Native and Indian; Asian; Hawaiian
	14. Ownership of housing unit	Owned or being bought by a household member*, occupied without payment of cash or rent for cash
Case 3	15. Spouse employment	No spouse or unmarried partner*; full-time spouse or unmarried partner; part-time spouse or unmarried partner; variable hours worked by spouse or unmarried partner; unemployed spouse or unmarried partner
	16. Time spent providing secondary care for children <13	0–1440 min
	17. Region	Northeast, midwest (formerly north central), south*, west
	18. Telephone	Own telephone in this house/apartment*, no telephone in this house/apartment
	19. Student status	Not a student*, student (1; full-time high school, part-time high school, full-time college or university, part-time college or university)
	20. Number of people accompanying	1–15
	21. Type of person accompanying	Unknown; alone*; family member or related person living within the household; unrelated person living within the household; family member or related person living outside the household; unrelated person living outside household; work-related person
	22. Previous activity	1–25 (reference category 23)

Note: * refers to the reference category.

3 Spatial variation and historical change in occupant behavior: statistical analysis and application on household activities and time scheduling

3.1 Purpose

Residents' occupancy and their activities at home have been recognized as two of the most important factors that determine residential energy demand, as they characterize the scale and temporal pattern of residential energy demand (Wilke et al., 2013; Zhao et al., 2014). OB models have been developed to capture residents' occupancy, activity, and action and reflect realistic patterns of buildings' energy demands. However, as mentioned in Chapter 1.5.2, the diversity in OB has not been fully investigated. For example, in some studies, the movement or mobility of people in space has been modeled to estimate the building energy consumption (Dziedzic et al., 2020; Mohammadi & Taylor, 2017), but spatial variation is not considered. Moreover, historical changes in OBs have not been taken into account (Deng & Chen, 2019; Hoes et al., 2009). The historical change represents long-term changes in people's lifestyles. The TUD has been used to observe the long-term changes in OB at an aggregate level. Some studies have considered the temporal variations by using measured time-series data to predict occupancy and energy demand (Calis et al., 2017; Piselli & Pisello, 2019; Yang et al., 2012). Temporal and spatial variations, however, were considered separately in these studies. Therefore, the spatial variation and historical change in OBs have not been effectively understood and assessed. Spatial variation and historical change are of high importance in OB modeling because the living location can predetermine the time required for some activities (e.g., the time required for commuting and shopping is different when the distance to travel is different among locations). Particularly, OB has generally been considered at the building level, but the nation-scale spatial analyses related to OB have seldom been conducted. More importantly, the specific modeling methods that are well suited to address spatial variation and historical change have not been established and assessed by conventional methods yet.

To address research gaps, this chapter presents a preliminary study to investigate the impacts of spatial variation and historical change of OB on residential energy demand. The purposes of this study were (1) to confirm the existence of spatial variation and historical change in household activities and time scheduling, (2) to find significant variables for representing spatial and temporal variations, and (3) to evaluate the performance of a logistic regression-based method for analyzing the spatial variation and historical change in household activities and time scheduling. The remainder of this chapter presents the methods, results, and discussion, followed by our conclusions about our approach to modeling the spatial variation and historical change in OB.

3.2 Data material and methodology

The data used in this study were still obtained from the ATUS, which includes multiple-year survey records. Although this data from 2003 to 2019 were available, we only used the data from 2009 to 2019 to ensure that the coding for each variable was consistent across each year since some variables were discontinued or newly introduced before 2009. The ATUS activity data subfile contained a summarised 24 h diary, starting from 4:00. The information in this subfile could be linked with other subfiles containing demographic attributes of the survey participants and the locations where the activities were performed. Although 17 main activities were defined in the multiple-year ATUS files, we only analyzed the activity of watching TV. As a result, 2,411,222 records of activities from 124,941 households were included in this study.

Figure 3-1 presents the analysis procedure. This study estimated the undertaking probability of watching TV activity and indicated the percentages of people watching TV at different times within 24 hours of the day. First, we analyzed the sample to confirm the existence of spatial variation, i.e., differences in the undertaking probability among living locations, and historical change, i.e., differences among the survey years. For this purpose, we quantified the average probability of watching TV for women with full-time jobs during the time interval from 21:00 to 21:59 in the U.S. in 2009, 2014, and 2019. Further, we aggregated the samples counted after multiplying a weight indicating the number of people represented by each sample that was given by the ATUS. We refer to this result as the ‘weighted subpopulation observation’. We chose to analyze watching TV activity and women with full-time jobs because watching TV is one of the main household activities in the ATUS (Xu & Chen, 2019) and the sample size of the women population was large and had various activity patterns.

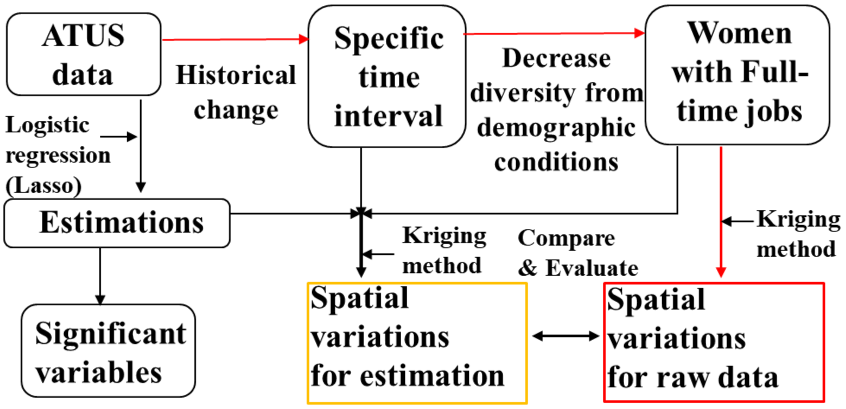

ATUS data
Historical change
Specific time interval
Decrease diversity from demographic conditions
Women with Full-time jobs
Logistic regression (Lasso)
Estimations
Kriging method
Compare & Evaluate
Kriging method
Significant variables
Spatial variations for estimation
Spatial variations for raw data

Figure 3-1. Research procedure.

Then, we conducted a logistic regression analysis to quantify the variation in the probability due to spatial and temporal variables for the entire population. For this analysis, data from the year 2009 to 2018 were used to develop the model (training model), and the data from 2019 were used to test this model (test model). Following the work of Wilke et al., 2013, we designed a case considering the socio-demographic conditions and the variables representing spatiotemporal variations. Year rank, population density, and spatial relationship were included to represent the spatial and temporal variations. Table 3-1 lists all the variables considered in this case.

Table 3-1. Predictor variables of the regression model for the whole population.

Variable	**Definition**	**Variable**	**Definition**
Disable	Respondent with disability.	Gender	Respondent is male.
Student	Respondent is a student.	Region	1: north-east; 2: mid-west; 3*: south; 4: west.
Carer	Respondent takes care of house or family.	Metropolitan status	1*: metropolitan; 2: non-metropolitan; 3: not clear.
Ill	Respondent is ill.	State code	1–56 (reference is CA: California).
Retire	Respondent is retired.	Day of week	Mon–Sun (reference group is Sunday).
Family income	1–14 levels (reference group is level 12: $100,000–$149,999)	Month	1–12 (reference group is January).
Work status	1: not in the labour force; 2*: full-time; 3: part-time; 4: with job, not at work; 5: unemployed.	Holiday	Dairy day is a holiday.
Housing type	1*: home, apartment, flat; 2: mobile home; 3: other types.	Year※	ATUS surveyed year.
Ownership of housing	1*: own; 2: rent; 3: other arrangements.	Year rank	1: 2009–2013; 2: 2014–2018; 3: 2019.
Education	1: Not completed secondary education/high school; 2*: high school; 3: college, no degree; 4: associate degree; 5: Bachelor’s degree; 6: Master’s degree; 7: professional school degree; 8: Doctorate degree.	Population density※	The number of people per unit of area (square mile).
Household size	1, 2*, 3, 4, 5, 6+.	Spatial relationship	Neighbor flag of a state. 1 means the state is the neighbor of the targeted state, 0 means the state is not the neighbor.
Age	1: 15–19; 2: 20–29; 3*: 30–39; 4: 40–49; 5: 50–59; 6: 60–69; 7: 70–79; 8: 80+.		
* indicates the reference group for each variable; ※ indicates that the variable is continuous.			

Subsequently, the developed regression models were evaluated based on the Hosmer–Lemeshow goodness-of-fit test and the following two indicators: total absolute error (TAE) and root mean squared error (RMSE). These indicators can be mathematically expressed as follows:

$$\text{TAE} = \left|\bar{p}_{Est_i} - \bar{p}_{Obs}\right| \quad 3\text{-}1$$

$$\text{RMSE} \;= \sqrt{\frac{\sum_{i=1}^{N}\left(p_{Est_i} - \bar{p}_{Obs}\right)^2}{N}} \quad 3\text{-}2$$

where N was the number of observations, and $\bar{p}_{Obs}$ and p_{Est} were the average undertaking probabilities observed in the sample and estimated by the regression model, respectively.

First, Lasso regression (Ranstam & Cook, 2018) was applied to select the variables for running the logistic regression for the whole population. Second, the significance of the spatiotemporal variables was identified based on the regression analysis. Then, we applied the ordinary kriging method to interpolate the spatial variation. The kriging method used the surrounding observations to predict the value of unmeasured locations. Its mathematical form was similar to a weighted regression. The prediction for unmeasured location (i_0, j_0), $\hat{Z}\left(s_{0_{(i_0,j_0)}}\right)$ was given by Equation 3-3:

$$\hat{Z}\left(s_{0_{(i_0,j_0)}}\right) = \sum_{k=1}^{N} \lambda_k Z\left(s_{k_{(i_k,j_k)}}\right) \quad 3\text{-}3$$

where $Z\left(s_{k_{(i_k,j_k)}}\right)$ was the observation value at the *k*th locations(i_k, j_k) and λ was unknown weight subject to $\sum_{k=1}^{N} \lambda_k = 1$. For the ordinary kriging the weight λ depended on 1) the distance between the locations of observations and the prediction and 2) the spatial relationship between the observations which surround the prediction. To obtain the weight λ, various empirical semivariograms were applied to fit the actual semivariogram so that they can reflect the spatial relationship between observations. In this study, we applied the widely considered empirical semivariogram-spherical model which was defined as:

$$\gamma(h) = \begin{cases} 0, h = 0 \\ c_0 + c_s\left\{\frac{3h}{2a} - \frac{1}{2}\left(\frac{h}{a}\right)^3\right\}, 0 < h \le a \\ c_0 + c_s, h \ge a \end{cases} \quad 3\text{-}4$$

where *h* was the lag which represents the distance to the observation. The parameters c_0 (nugget), c_s (partial sill) and a (major range) are non-negative constants that will be optimized to fit with the semivariogram.

Our kriging analysis was conducted only for the subpopulation of women with full-time jobs to minimize the influence of diversity of sociodemographic conditions. The estimated probabilities were then weighted to estimate the average probability for the entire subpopulation of the states. Then, we spatially interpolated the estimated probabilities for the states using the ordinary kriging method. The location of the states is represented by the internal points of the states. To assess the model's ability to replicate the spatial variation and historical change, the predicted probability was compared with the probability based on the weighted observations.

3.3 Results and discussion

3.3.1 Data Analysis

This subchapter first presents the results obtained based on the raw data. Figure 3-2 presents the yearly change in the undertaking probability of watching TV based on the weighted observations for the whole population and subpopulation, respectively. We observed a general decreasing trend during the 20:00 to 00:00 time period for both the whole population and the subpopulation of women with full-time jobs. The time interval from 21:00 to 21:59 exhibited the largest decrease (2.6% for the whole population and 2.4% for the subpopulation). Therefore, we picked this time interval to further visualize the spatial variation of the probability.

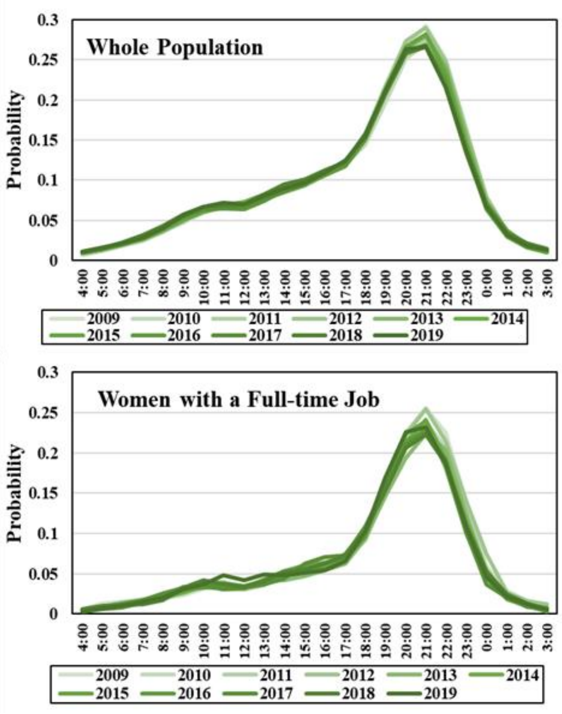
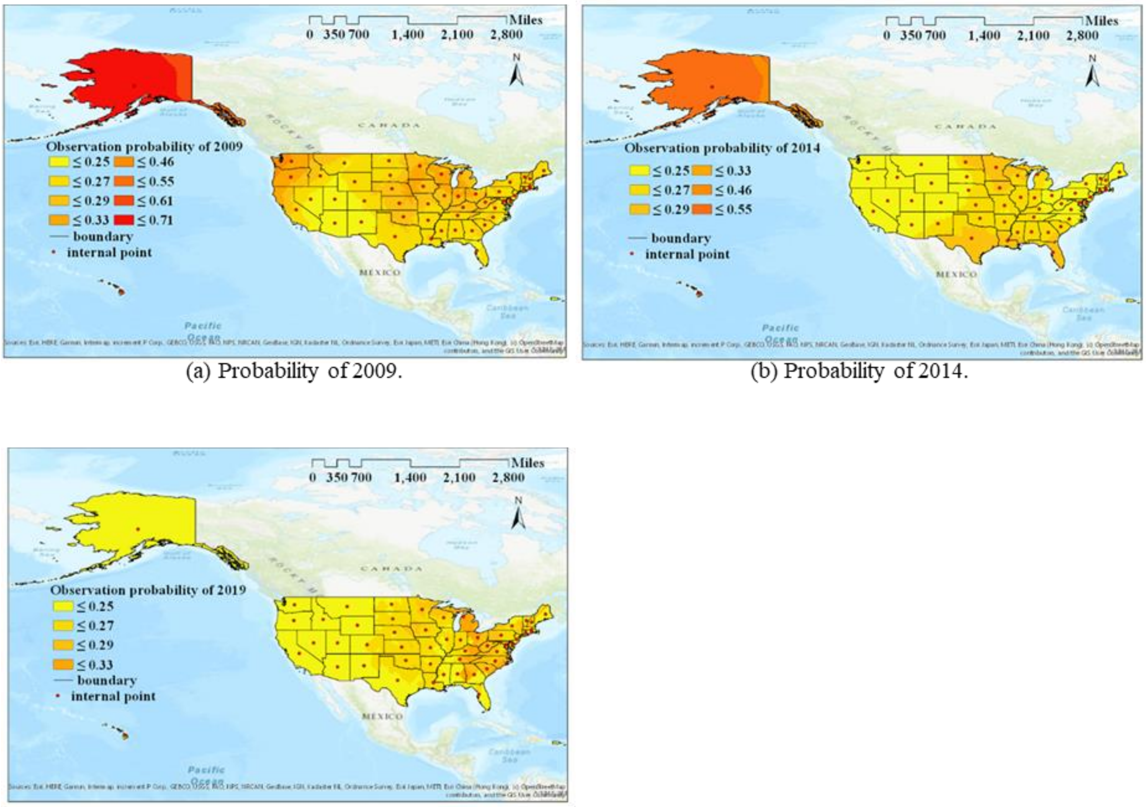


Figure 3-2. Weighted undertaking probability of watching TV per hour for each year.

Figure 3-3 presents the visualization results of the weighted subpopulation observations. The shapefile data of the U.S. in 2018 and the location data (latitude and longitude of the internal points representing each state) used to obtain the results were downloaded from the website of the U.S. Census Bureau. The color scale is consistent among the images in Figure 3-3 for comparison. Figure 3-3 depicts a spatial variation that changes over time. The general observed trend was that the region with a higher probability shifted from the north to the south and then to the east. In 2009, higher probability levels were located in most of the northwest and central parts (central but closer to the north) of the U.S. In 2014, the high probability levels shifted to the mid-south and mid-north areas. Finally, the higher probability levels relocated to be more concentrated in the southeast areas in 2019. We also observed a decreasing trend in the five-year periods as the probability of the areas with a high probability level decreased; this observation was consistent with the information presented in Figure 3-2.



(a) Probability of 2009.

(b) Probability of 2014.

(c) Probability of 2019.

Figure 3-3. Kriging based on weighted observations.

It is important to note that the results in Figure 3-3 did not fully represent the actual spatial variation. One of the significant limitations was that we used one internal point to represent the entire state. To capture the actual spatial variation, a higher granularity should have been incorporated. Additionally, the results for marginal areas may have contained an error originating from the characteristics of the ordinary kriging method. For example, Alaska, the state with the largest area, exhibited the highest probability. In reality, Alaska has the lowest population density, implying that the majority of the states must have been uninhabited or sparsely populated. Such areas in Alaska should have exhibited lower probability levels. Therefore, a single internal point, without nearby data points or with only a few nearby data points located in limited directions, might not have provided accurate spatial variations for the entire area.

3.3.2 Results of Logistic Regression

The initial model for the whole population during the time interval from 21:00 to 21:59 based on logistic regression, with Lasso regression to select the variables, did not pass the Hosmer–Lemeshow goodness-of-fit test. Therefore, to optimize the model, we conducted the Analysis of Variance (ANOVA) to reselect the significant variables. The improved model exhibited a relatively high prediction accuracy. It passed the Hosmer–Lemeshow goodness-of-fit test for both the training model and the test model, as the P-values of the improved model were 0.26 and 0.34 for the training and test model, respectively; both the P-values > 0.05, meaning that the model fits well with the observations. The TAE and RMSE for the test model of the improved model were 7.6%, and 1.5%, respectively. These two indicators verified that our improved model performed well in terms of probability errors.

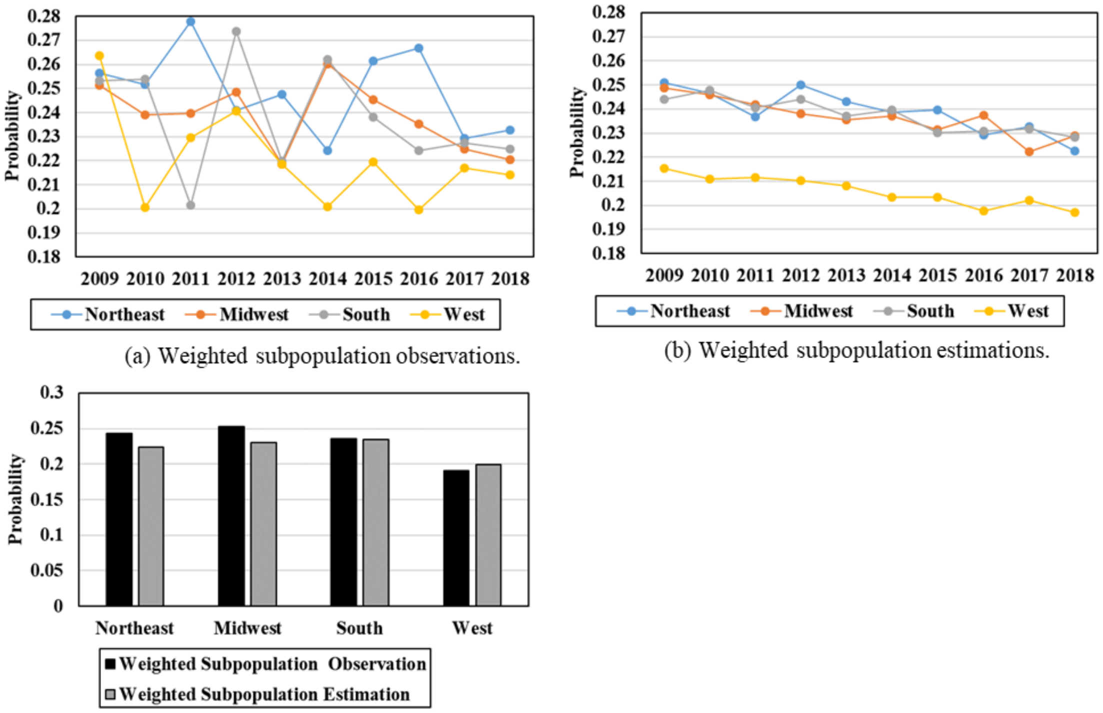
Table 3-2 lists the regression coefficients (RCs) and odds ratios (ORs) of the significant variables, determined by the improved model. Regarding the socio-demographic variables, the probabilities were lower for students and people who took care of family members in the household. On the contrary, the probabilities were higher for people who were not in the labor force or were unemployed. With respect to the temporal variables, almost all the temporal variables were significant. These temporal variables, such as year, month, and day of the week, had negative influences on the probability of watching TV, such that the probability within the reference groups was estimated to be the lowest in the respective categories. In terms of the spatial variables, a lower probability was estimated for people living in the western part of the U.S. than that for people living in other regions. Moreover, one special state code flag, Florida (FL), was found to be significant, indicating that people who lived in Florida were more likely to watch TV from 21:00 to 21:59 in 2019 than those in the other states.

Figure 3-4 presents the spatial variation and historical change for the subpopulation which was extracted from the improved model made based on the whole population. Figure 3-4 (a) and (b) depict the spatial variation and yearly change of observations and estimations for the four regions from 2009 to 2018. A large-scale fluctuation was seen in the observation probabilities, whereas a decreasing trend was observed in the estimation probabilities in the four regions. Figure 3-4 (c) depicts the predicted results for 2019. We observed that the absolute difference for each region was less than 2.3%. The spatial variation narrowed as the difference in the maximum and minimum probabilities decreased from 6.2% in the observation results to 3.6% in the estimation results. These results indicate that the logistic regression model with significant variables was capable of replicating spatial variation and historical change at the aggregate level. The scale of the prediction error was approximately 10% of the actual probability.

Table 3-2. Significant variables based on the improved model for the whole population.

Variable	Dummy Variable	RCs	ORs	Variable	Dummy Variable	RCs	ORs
Intercept	-	15.07***	2	Day of week	Monday	−0.09***	0.91
Male	-	0.22***	1.25		Tuesday	−0.11***	0.89
Student	-	−0.78***	0.74		Wednesday	−0.14***	0.87
Carer	-	−0.49***	0.75		Thursday	−0.14***	0.87
Family income	$10,000–$14,999	−0.05*	0.94		Friday	−0.12***	0.88
Work status	Not in labour	0.53***	1.43		Saturday	−0.08***	0.92
	Unemployed	0.11***	1.15	Month	Mar	−0.06**	0.94
Ownership of housing	Rent	−0.14***	0.95		Apr	−0.08***	0.92
	Other	−0.26***	0.82		May	−0.11***	0.9
Education	Less than high school	−0.13***	0.93		Jun	−0.17***	0.85
	Associate school	−0.12***	0.89		Jul	−0.19***	0.82
	College	−0.11***	0.9		Aug	−0.14***	0.87
	Bachelor’s degree	−0.18***	0.84		Sep	−0.13***	0.88
	Master’s degree	−0.3***	0.72		Oct	−0.1***	0.9
	Professional school	−0.27***	0.74		Nov	−0.09***	0.91
	Doctor	−0.46***	0.63		Dec	−0.13***	0.88
Number of people	1	0.2***	1.04	Year	-	−0.01***	0.99
	3	0.22***	1.09	State	FL	0.95**	0.94
	5	−0.1***	0.87	Region	West	−0.15***	0.8

*** < 0.001, ** < 0.01, and *< 0.05



(a) Weighted subpopulation observations.

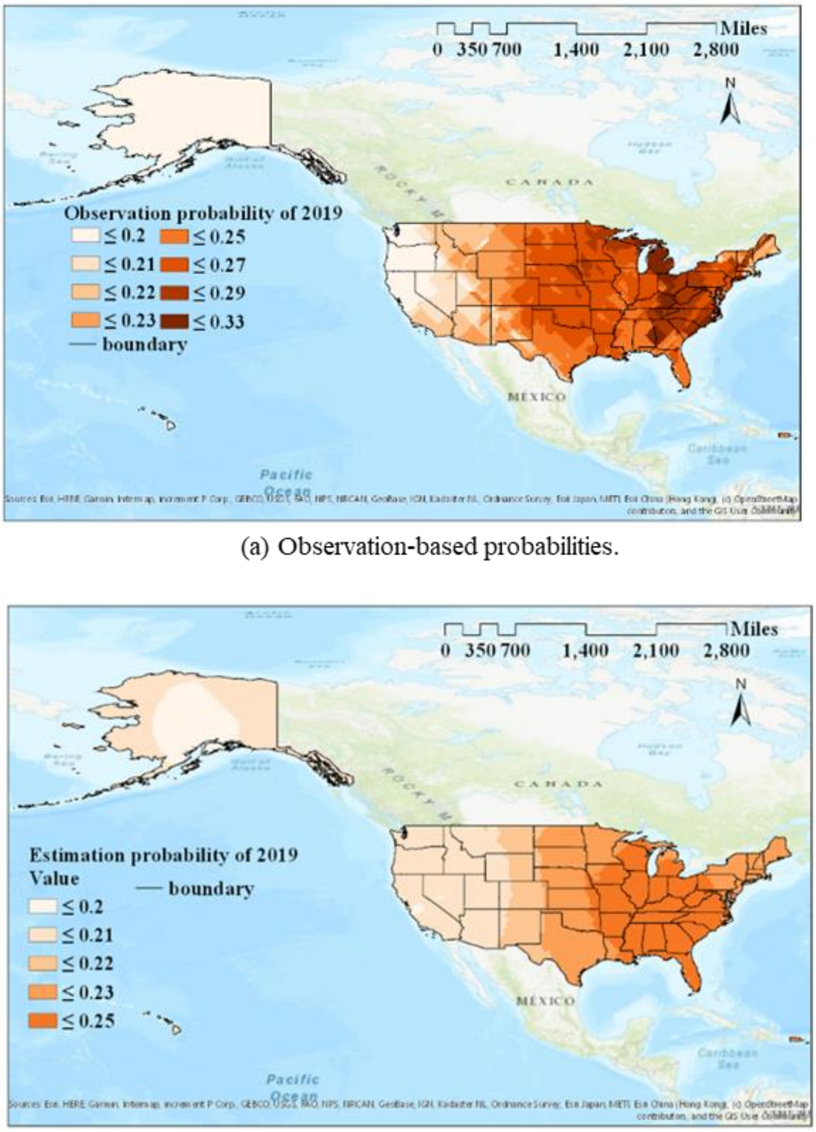
(b) Weighted subpopulation estimations.

(c) Predicted regional difference for 2019.

Figure 3-4. Spatial variation and historical change shown by the improved model.

3.3.3 Results of the kriging Method

To evaluate the performance of the logistic regression-based approach and further investigate the influence of spatial variations, the ordinary kriging method was applied to interpolate the probabilities. Figure 3-5 illustrates the comparison of the ordinary kriging results based on the weighted subpopulation observations and weighted subpopulation estimations. The classification of the undertaking probability levels was refined using Figure 3-3 for better comparison. According to Figure 3-5, the estimation-based results had lower probability levels than the observation-based results. However, the spatial distribution trend was similar (see Figure 3-5 (a) and (b)); the probability was higher in the eastern areas, whereas it was lower in the western areas. This result indicates that a general trend can be replicated using the logistic regression model, as discussed in the previous chapter on the logistic regression model. However, the spatial differences at a higher spatial resolution, as observed in Figure 3-5 (a), could be replicated. The kriging-based model can replicate such differences as it allows one state to have multiple probability levels. With respect to detailed spatial variations, a relatively clear band-shaped distribution pattern in the west-east direction was observed in Figure 3-5 (a) and (b). The spatial variations, however, were weakened in the estimation-based result. The results did not accurately reflect the probabilities for the highest-level areas, such as the northeast part of the U.S., including Michigan and Minnesota. The error for such areas ranged from 6% to 11% partly owing to the limitations of the logistic regression-based approach in reflecting spatial variations.



(a) Observation-based probabilities.

(b) Estimation-based probabilities.

Figure 3-5. Comparison of the kriging results.

3.4 Conclusion

The objective of this study was three-fold: (1) to confirm the existence of spatial variation and historical change in OB, (2) to find significant variables for representing the spatial and temporal variations, and (3) to evaluate the performance of a logistic regression-based approach to consider the spatial variations in OB. First, based on the analysis from the ATUS data, we confirmed the existence of spatial variation and historical change in the watching TV activity for the subpopulation of women with full-time jobs. We observed a clear historical transition as the probability of watching TV during the time interval from 21:00 to 21:59 decreased from 2009 to 2014 and 2014 to 2019. This result may be due to the fact that many people changed the time at which they watched TV or they participated in other emerging entertainment activities, such as playing games on computers or smartphones during this time interval. We also observed spatial variation and historical change wherein higher probability levels first shifted from the north to the south (2009 to 2014) and then gradually moved to the eastern part (2014 to 2019) of the country.

Then, the significant variables were determined using the logistic regression model. We obtained a regression model that fits well with the TUD sample as the developed model passed the goodness-of-fit test and the error was small enough as well. Socio-demographic and spatiotemporal variables were selected for the model. With respect to the temporal variables, the day of the week, month, and year were significant. Regarding spatial variables, the probability for people living in the western part of the U.S. was lower than that in other regions. Some specific states were also found to be significant (e.g., Florida). Based on these results, the logistic regression method was partly proved to be able to replicate the spatial variation and historical change in OB modeling. However, not all the considered spatiotemporal variables were significant, such as metropolitan status. One possible reason is that the penetration rate of TV in the U.S. is high and the metropolitan status may not have much impact on watching TV activity. Further analysis is required to determine the types of variables that can represent spatial variations and historical change and the types of formats that should be used for variables in OB modeling.

Subsequently, we applied the ordinary kriging method to evaluate the spatial variation of the probability estimated using the developed logistic regression model. The results indicated that a general trend can be replicated using the logistic regression-based approach, but this approach is not as effective for the replication of spatial variation and historical change. The kriging-based model in our study showed a strong advantage for representing spatial variation and historical change. Moreover, the kriging-based model can predict the undertaking probability for locations without data that cannot be estimated only based on the logistic regression model. It indicates that the kriging method is a possible prediction approach that can contribute to the field of TUD-based OB modeling in the case of a lack of data. About 51 internal points, however, were considered to conduct interpolation for the entire U.S., therefore, the spatial interpolation performance may not be ideal for some marginal areas which had fewer neighbor observations (e.g., Alaska). The identification of locations at higher spatial resolution would contribute to replicating spatial variation and historical change because more accurate and realistic interpolation predictions could be conducted. More importantly, the kriging-based model in this study can be applied to other energy-related activities of interest, which may benefit energy demand modeling in the fields of OB and building energy efficiency.

4 Modeling of occupant behavior involving spatial variation: geostatistical analysis and application based on American Time Use Survey data

4.1 Purpose

Numerous OB models that simulate occupancy, activity, and action at home have been developed to improve the accuracy and quality of energy demand estimations (Osman & Ouf, 2021; Yan et al., 2017). As OBs play the main role in shaping the residential energy demand profiles. Various methods have been applied to TUD integrated with additional survey data that cover social, economic, and building aspects, to develop representative OB models. Previous studies have revealed that the consideration of diversity improves the performance of OB models (Li et al., 2022). However, obtaining a reliable source of data and developing representative models for these OBs remain a key challenge, especially for developing energy models with consideration of OBs at high spatial resolution.

Based on the background mentioned in Chapter 1.5.2 together with the findings in Chapter 3, we found that the previous studies have seldom confirmed the existence of spatial variation instead most of them assumed the spatial variation existed for the targeted research objective for the entire modeling time. Also, existing models ignore spatial variation in OBs or partially consider it using a simple method without evaluating whether it is sufficient. Spatial variation is commonly treated as the difference among the research objectives represented by different measured locations in previous studies which cannot be fully reproduced. Moreover, the modeling method to reproduce the spatial variation is missing.

Hence, the study in this chapter proposes and evaluates methods to model OB with the consideration of spatial variation. The research gaps were addressed through three research questions: 1) when does spatial variation exist in OB, 2) how can spatial variation in OB be represented quantitatively, and 3) how can spatial methods reproduce spatial variations in OB. We selected a spatial logistic regression model as the spatial method in this study as it is an extension of one of the most frequently used OB models. The remainder of this chapter presents the methodology, results, and discussion, followed by our conclusions.

4.2 Data material and method

4.2.1 Data material

The multi-year ATUS0319 collected the activity diaries and sociodemographic conditions of the survey participants. The data collected between 2009 and 2019 were used to ensure the consistent coding of the variables. We selected women aged 30–59 years in the U.S. because the sample size of women was large in the ATUS dataset and women conduct various activities including both paid and unpaid work (Anxo et al., 2011; Gentry et al., 2003; Li & Tilahun, 2020; Sayer, 2005). 70% of the data was used as the training dataset, whereas the remaining was used as the test dataset.

The predefined activities were summarized into the 16 categories as listed in Table 4-1 such that: 1) the sub-categories in a group exhibited similar appliance usage and 2) the activity locations could be divided into indoor and outdoor groups. Four typical activities were considered: sleeping, cooking and washing up, watching television, and commuting. Additionally, the 1 min resolution data were converted to 1 h time interval data.

The location of each occupant was defined by the internal point of the state in which the occupant lived on the U.S. mainland. We considered the states as the unit for modeling as it was the only available data with respect to space use for the entire nation. The cartographic boundary shapefile of the U.S. as of 2018 was used to visualize the spatial distribution of the probability on the map. The spatial distribution of the probability of undertaking the activities at each time interval is referred to as the spatial probability in this study.

Table 4-1. Activity code.

Code for all activities in this study					
Code	**Activity**	**Location (indoor)**	**Code**	**Activity**	**Location (outdoor)**
1	Sleeping	Home and yard	10	Work-relted	Workplace
2	Grooming		11	Education	School, library
3	Laundry		12	Commuting	Transporptaion
4	Caring		13	Other travelling	Transportation
5	Cooking and washing up		14	Consumer purchase	Store, mall
6	Eating & drinking		15	Other 2	Not at home or yard
7	Watching Television		16	Other 3	Missing
8	Listen to music				
9	Other 1				

4.2.2 Method

The methodology of this study is shown in Figure 4-1. Steps 1–3 address research questions discussed in Chapter 4.1.

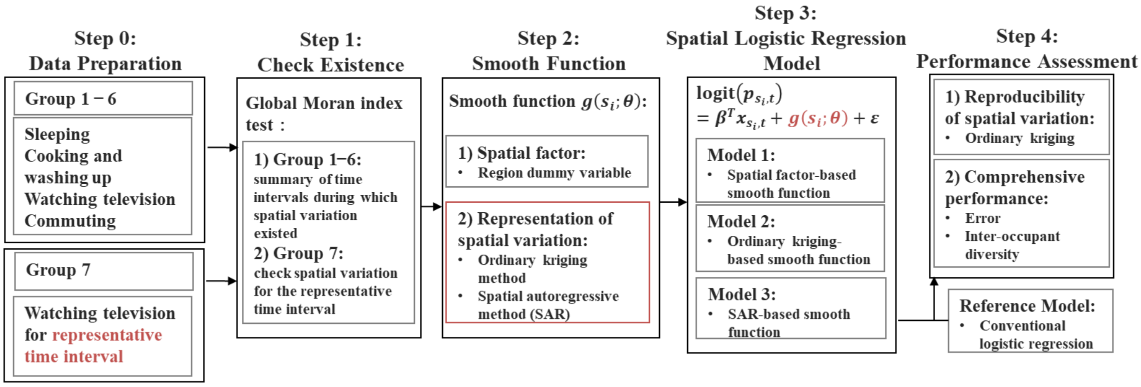

Step 0:
Data Preparation
Group 1 − 6
Sleeping
Cooking and washing up
Watching television
Commuting
Group 7
Watching television for representative time interval
Step 1:
Check Existence
Global Moran index test :
1) Group 1−6:
summary of time intervals during which spatial variation existed
2) Group 7:
check spatial variation for the representative time interval
Step 2:
Smooth Function
Smooth function $g(s_i;\theta)$:
1) Spatial factor:
• Region dummy variable
2) Representation of spatial variation:
• Ordinary kriging method
• Spatial autoregressive method (SAR)
Step 3:
Spatial Logistic Regression Model
$\mathrm{logit}(p_{s_i,t}) = \beta^T x_{s_i,t} + g(s_i;\theta) + \varepsilon$
Model 1:
• Spatial factor-based smooth function
Model 2:
• Ordinary kriging-based smooth function
Model 3:
• SAR-based smooth function
Step 4:
Performance Assessment
1) Reproducibility of spatial variation:
• Ordinary kriging
2) Comprehensive performance:
• Error
• Inter-occupant diversity
Reference Model:
• Conventional logistic regression

Figure 4-1. Study methodology.

4.2.2.1 Step1: Existence of the spatial variation

We applied the Global Moran Index (Moran's *I*) test to confirm the time intervals that spatial variation existed for selected four activities. Moran's *I* test is used to check the significance of the random distribution of qualitative determination on areas of a map (Moran, 1948). Moran's *I* ranges from –1 to 1 and its definition is:

$$I = \frac{\sum_{i=1}^{N}\sum_{j=1}^{N} \mathrm{w}_{s_i,s_j}(p_i - \bar{p})(p_j - \bar{p})}{S^2 \sum_{i=1}^{N}\sum_{j=1}^{N} \mathrm{w}_{s_i,s_j}} \qquad \text{4-1}$$

where I is the Moran's *I*, p_i is the probability for state s_i, $\bar{p}$ is the average probability for all states, S^2 is the sample variance and w_{s_i,s_j} is the element for state s_i and s_j in the weighting matrix. *Z* score was calculated to evaluate the significance of the Moran's *I*:

$$Z = \frac{I - E(I)}{\sqrt{var(I)}} \qquad \text{4-2}$$

If the *Z* score is not statistically significant, (P-value > 0.05), it is probable that the objectives are randomly distributed in space; if the *Z* score is positive and significant, the objectives display a clustered distribution (similar tendency); if the *Z* score is negative and significant, the objectives display a dispersed distribution (competitive tendency). The subsequent steps only considered the time intervals during which spatial variation existed.

4.2.2.2 Step 2: Methods to represent spatial variation

Two representations of spatial variation that quantify the average probability of undertaking an activity in each state s_i at each time interval were designed using the ordinary kriging and spatial autoregressive (SAR) methods. However, as they measure the probability of undertaking an activity, their values were restricted from 0 to 1. Furthermore, the ordinary kriging and SAR methods can generate representations at higher resolutions if detailed location data are available.

A) Ordinary kriging method

The ordinary kriging method uses the observations of the surroundings to predict the values of unmeasured locations (Cressie, 1988) which has been simply introduced in Chapter 3.2. Here give a more detailed explanation. For a certain time interval that the spatial variation existed, the prediction G_{s_0} for location $s_0(u_0, v_0)$ was given by:

$$\tilde{G}_{s_0(u_0,v_0)} = \sum_{i=1}^{N} \lambda_i \, G_{s_i(u_i,v_i)} \tag{4-3}$$

where $G_{s_i(u_i,v_i)}$ was the average undertaking probability of activity at the state s_i represented by the internal point (u_i, v_i), and λ_i was unknown weight subject to $\sum_{i=1}^{N} \lambda_i = 1$ for obtaining the unbiased estimation of G_{s_0}. λ can be estimated by Equation 4-4 to achieve the minimum variance estimation of G_{s_0}:

$$\text{agrmin}_{\lambda} \;\; f: E\left\{\left[G_{s_0(u_0,v_0)} - \tilde{G}_{s_0(u_0,v_0)}\right]^2\right\} - 2\mu\left(\sum_{i=1}^{N} \lambda_i - 1\right) \tag{4-4}$$

The widely used approach to dealing with the first item of Equation 4-4 was to apply the theoretical semivariogram which is defined as $\gamma(s_i, s_j) = \gamma(s_i - s_j) = \frac{1}{2}E\left\{\left[G_{s_i(u_i,v_i)} - G_{s_j(u_j,v_j)}\right]^2\right\}$ to fit the experiment variogram. A commonly considered theoretical semivariogram — spherical model defined as Equation 3-4 was mentioned in Chapter 3.2.

B) Spatial autoregressive method

The SAR allows us to examine the impact that the undertaking probability of an activity for one state has on other neighboring states by including other variables in the modeling process. It is generated based on the cross-sectional spatial model which is defined as Equation 4-5:

$$\tilde{G}_{s_0(u_0,v_0)} = \mathrm{y}_{s_0} = \beta^T x + \lambda^T \mathrm{W}\mathrm{y}_{s_0} + \varepsilon \tag{4-5}$$

where $\tilde{G}_{s_0(u_0,v_0)}$ is average undertaking probability of activity at the state s_o. x are variables and W is the weighting matrix that is constructed in the form of adjacent edges or points of each state. λ is scalar autoregressive parameters. The variable $\mathrm{W}\mathrm{y}_{s_0}$is referred to as the spatial lag of y_{s_0}.

4.2.2.3 Step 3: Spatial logistic regression

In this subchapter, we developed three spatial logistic regression models through Equation 4-6:

$$\text{logit}(p_{s_i,t}) = \ln\frac{p_{s_i,t}}{1-p_{s_i,t}} = \beta^T x_{s_i,t} + g(s_i;\theta) + \varepsilon \qquad 4\text{-}6$$

where $p_{s_i,t}$ is the probability for ith individual of location s at t time interval. β is a coefficient of the variable $\mathrm{x}_{s_i,t}$. $g(s_i;\theta)$ is the smooth function parameterized by θ over location s. We used one spatial factor variable $\tilde{G}_r$ and two representations of the spatial variation $\tilde{G}_s$ as the smooth function $g(s_i;\theta)$ which the θ were considered as the averaged probability of undertaking activities at the regional and state level. The conventional logistic regression model without considering spatial variation served as the reference model for comparison. Stepwise analysis was applied to all the models to statistically test the significance of the considered variables, including the spatial factor variables and representations.

A) Reference model

For the conventional logistic regression model, the probability of undertaking activity at time *t* for an individual *i* is modeled by Equations 4-7 and 4-8:

$$\text{logit}(p_{i,t}) = \ln\frac{p_{i,t}}{1-p_{i,t}} = \beta^T x_{i,t} + \varepsilon \qquad 4\text{-}7$$

$$p_{i,t} = \frac{1}{1+e^{-(\beta^T x_{i,t}+\varepsilon\ )}} \qquad 4\text{-}8$$

where β is a coefficient of the socio-demographic variable $x_{i,t}$.

B) Model 1

Model 1 introduces the spatial factor —three region dummy variables to represent the spatial variations. Therefore, Equation 4-6 can be rewritten as:

$$\text{logit}(p(x_i, r_i)_t) = \beta^T x_{i,t} + \tilde{G}_{r_{i,t}} + \varepsilon \qquad 4\text{-}9$$

$$\tilde{G}_{r_{i,t}} = \gamma_1 R_{1,i,t} + \gamma_2 R_{2,i,t} + \gamma_3 R_{3,i,t} \qquad 4\text{-}10$$

where R_1, R_2 and R_3 indicates the northeast, mid-west, and west region respectively with the region of the south being the reference group. γ is the corresponding coefficient to each region dummy variable.

C) Model 2

We extracted the estimations $\tilde{G}_{s_{i,t}}$, the average undertaking probability of activities for each state s_i at t time interval, from the ordinary kriging interpolation results to represent the spatial variations. Then the probability of undertaking an activity estimated by Model 2 is given by Equation 4-11:

$$\mathrm{logit}(p(x_i, s_i)_t) = \beta^T x_{i,t} + \gamma \tilde{G}_{s_{i,t}} + \varepsilon \quad 4\text{-}11$$

where the $x_{i,t}$ are the variables of the individual i at t time interval.

D) Model 3

The calculation of Model 3 is as same as Model 2. The only difference is that estimations $\tilde{G}_{s_{i,t}}$, the average undertaking probability of activities for each state, was extracted from the SAR results to represent the spatial variations.

4.2.2.4 Segmentation

To develop the smooth functions and the corresponding spatial logistic regression models, six groups were designed to represent different subpopulations of women. Each group was homogenized by avoiding the influence of spatial variation in the socio-demographic factor, as shown in Table 4-2. The variables used in the developed models are also listed in Table 4-2. The grouping conditions used were the type of day (i.e., weekdays and weekends) and employment status, which have been commonly used in previous studies for segmentation (Kleinebrahm et al., 2021; Marín-Restrepo et al., 2020; C. Wang et al., 2011; Wilke et al., 2013; Zhou et al., 2022). Groups 1 and 4 represent women with full-time jobs, Groups 2 and 5 represent women with part-time jobs, and Groups 3 and 6 represent unemployed women. Groups 1–3 comprise activities that were performed during the weekdays, and Groups 4–6 comprise activities that were performed during the weekends.

Table 4-2. Designed groups and their details.

Group	Subpopulation	Type of day	Employment status	Items of interest	Observations
1	Women aged 30–59	Weekdays	Full-time	Survey year, age, presence of children, family income, carer, education, ownership of the housing unit, number of people in the household, region, and state	8888
2			Part-time		3571
3			Unemployed		5753
4		Weekends	Full-time		9086
5			Part-time		3571
6			Unemployed		5753
7	Entire population of women aged 30–59			Items in Groups 1–6, as well as employment status and type of day	36622

In addition to Groups 1–6, Group 7 was considered for representing the entire population of women aged between 30–59 including Groups 1–6. Group 7 was designed to examine whether the developed spatial logistic regression model 1) can be applied to larger and more complex populations and 2) can be used to determine the superior approach for OB modeling, using segmentation (Hayn et al., 2014) or using grouping conditions as variables. Furthermore, many previous OB studies designed segmentations for simulated occupants (Li et al., 2022). This analysis was conducted considering the watching television activity.

4.3 Performance assessment

The performance of the models was assessed in terms of the reproducibility of the spatial variations in OBs and the comprehensive performance. The ordinary kriging method was applied to visualize the spatial probability, thereby assessing the reproducibility of the spatial variation. The comprehensive performance was evaluated by indicators to assess the error and diversity considering the training and test datasets.

4.3.1 Error indicator

Error indicators measure the errors between the estimations obtained from all models and the observations. The first indicator is the TAE which quantified the cumulative value of the errors observed in all time intervals that spatial variation existed for each of the selected activities. TAEs assess the ability to model the total averaged undertaking probability at the national and state level respectively which is crucial for obtaining a more realistic averaged energy demand. TAEs are quantified as Equations 4-12 and 4-13 show:

$$\mathrm{TAE}_{\mathrm{nation}} = \sum\nolimits_{t=1}^{T} \left| \bar{p}_{est_t} - \bar{p}_{obs_t} \right| \qquad 4\text{-}12$$

$$\mathrm{TAE}_{\mathrm{state}} = \sum\nolimits_{t=1}^{T} \sum\nolimits_{s=1}^{S} \left| \bar{p}_{est_{t,s}} - \bar{p}_{obs_{t,s}} \right| \qquad 4\text{-}13$$

where t indicates the selected time interval that spatial variation exists, s indicates the state of the U.S., and $\bar{p}_{est}$ and $\bar{p}_{obs}$ is the average probability of the estimation and the observation respectively.

The second indicator is the RMSE. RMSEs are quantified to measure the averaged errors of each selected time interval and averaged errors of combinations of the state and time interval. They are sensitive to individual outliers of the estimations which are shown by Equations 4-14 and 4-15:

$$\mathrm{RMSE}_{\mathrm{nation}} = \sqrt{\frac{\sum_{t=1}^{T} \left(\bar{p}_{est_t} - \bar{p}_{obs_t} \right)^2}{T}} \qquad 4\text{-}14$$

$$\mathrm{RMSE_{state}} = \sqrt{\frac{\sum_{t=1}^{T}\sum_{s=1}^{S}\left(\bar{p}_{est_{t,s}} - \bar{p}_{obs_{t,s}}\right)^2}{S * T}} \tag{4-15}$$

4.3.2 Diversity indicator

Diversity indicators assess how well the model represents the variation in OBs among simulated occupants. We followed the Hosmer-Lemeshow test to design subgroups to check the diversity of the model estimations which measures the root mean squared error between the averaged estimated probability and averaged probability of observations of subgroups shown by Equation 4-16:

$$\mathrm{RMSE_{GA}} = \sqrt{\frac{\sum_{t=1}^{T}\sum_{d=1}^{D}\left(Mean_{t,d}\left(P_{pred}\right) - Mean_{t,d}\left(P_{obs}\right)\right)^2}{T * D}} \tag{4-16}$$

where d indicates the subgroup ($D = 10$) and the number 10 is commonly used in the Hosmer-Lemeshow test. RMSE_GA is only quantified at the national level because of the data limitation.

To compare the diversity performance at the state level, another indicator — MSD was considered. MSDs measure the mean of the distance that each estimation deviates from the mean at the national and state level respectively. Their definitions are shown by Equations 4-17 and 4-18:

$$\mathrm{MSD_{nation}} = \frac{\sum_{t=1}^{T} SD_{pred_t}}{T} \tag{4-17}$$

$$\mathrm{MSD_{state}} = \frac{\sum_{t=1}^{T}\sum_{s=1}^{S} SD_{pred_{t,s}}}{T * S} \tag{4-18}$$

where SD_{pred} is the standard deviation of estimated probability among the sample.

4.4 Results

4.4.1 Confirmation of the existence of the spatial variation

Figure 4-2 shows the representative probabilities of the women in Group 4 sleeping, those in Group 3 cooking and washing up, those in Group 6 watching television, and those in Group 1 commuting based on all observations. As shown in Figure 4-2, the probability of undertaking activities exhibited certain variation among the states at different times of the day. Such variation results from the combination of the difference in demographic factor variables, the spatial variation, and the random bias, according to Equation 4-6. The effect of the first element (i.e., the difference in demographic factor variables) is decreased by the segmentation.

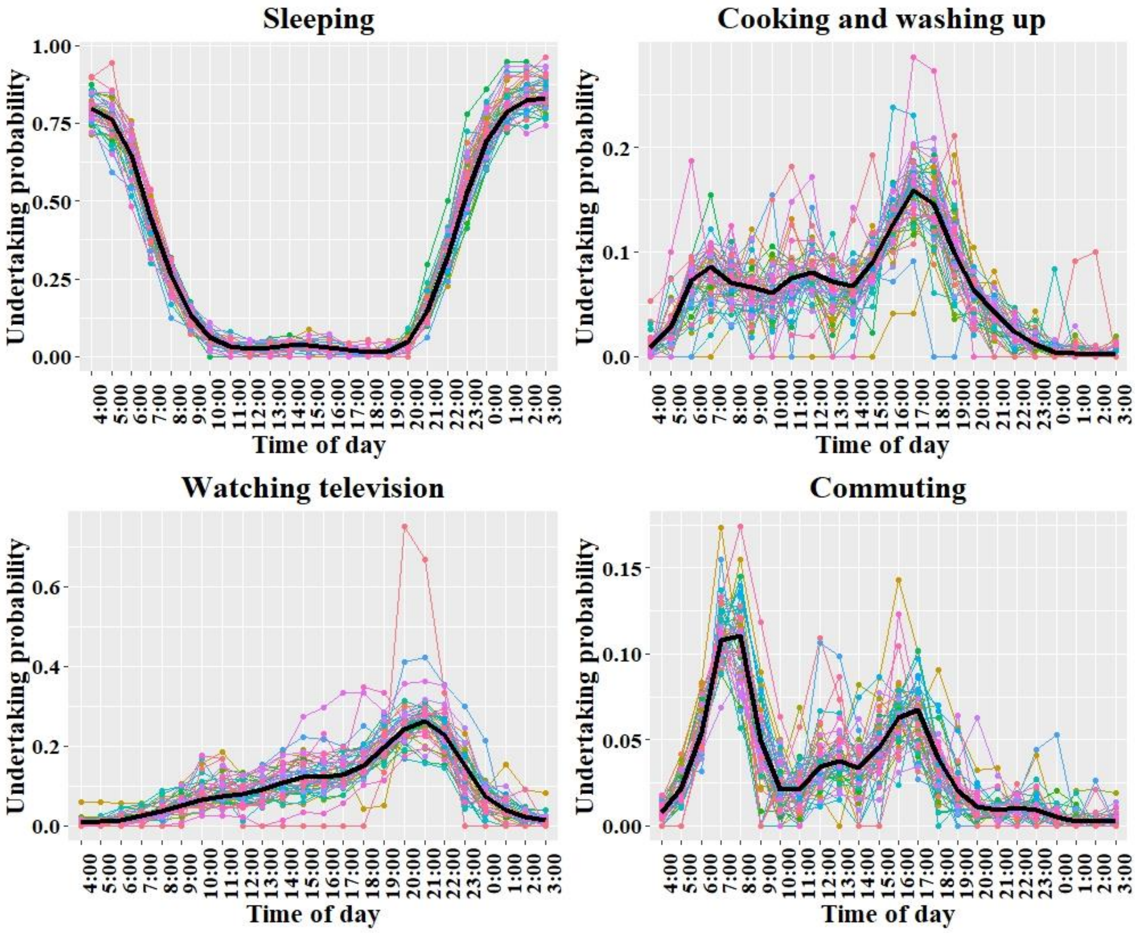


Figure 4-2. Probability of undertaking activities for representative groups. Lines in different colors indicate different states and the black one indicates the national level estimations.

Figure 4-3 summarizes the results of the Moran's *I* tests applied for all combinations of the groups and activities. Spatial variation existed only during limited time intervals. The spatial variation during different time intervals varied with the type of day (weekdays or weekends), subpopulations with different employment statuses, and activities. As shown in Figure 4-3, considering sleeping, employed women in Group 1 exhibited lesser spatial variation than unemployed women in Group 3 during the relevant time interval on weekdays. On weekends, women exhibited the same number of spatial variations during the relevant time intervals, irrespective of their employment statuses. Considering cooking and washing up, unemployed women in Group 3 exhibited more spatial variation during the weekdays, whereas women with full-time jobs in Group 4 exhibited more spatial variation during the weekends. No spatial variations existed for women with full-time jobs in Group 1 on weekdays and unemployed women in Group 6 on the weekends. Considering watching television, women with part-time jobs in

Group 5 did not exhibit any spatial variation during the weekends. Women with part-time jobs in Group 2 further exhibited a low spatial variation during the weekdays. Considering commuting, irrespective of their employment status, women in Groups 1–3 exhibited more spatial variation during the weekdays than those in Groups 4–6 during the weekends. Women with part-time jobs in Group 5 did not exhibit any spatial variation during the weekends.

In most time intervals, the spatial variation exhibited a clustered distribution, with only limited time intervals exhibiting a dispersed distribution. Figure 4-4 illustrates the probability of the women in Group 6 watching television at 13:00. An obvious clustered distribution can be observed at the state level. The observed spatial probability ranged from 0–21%.

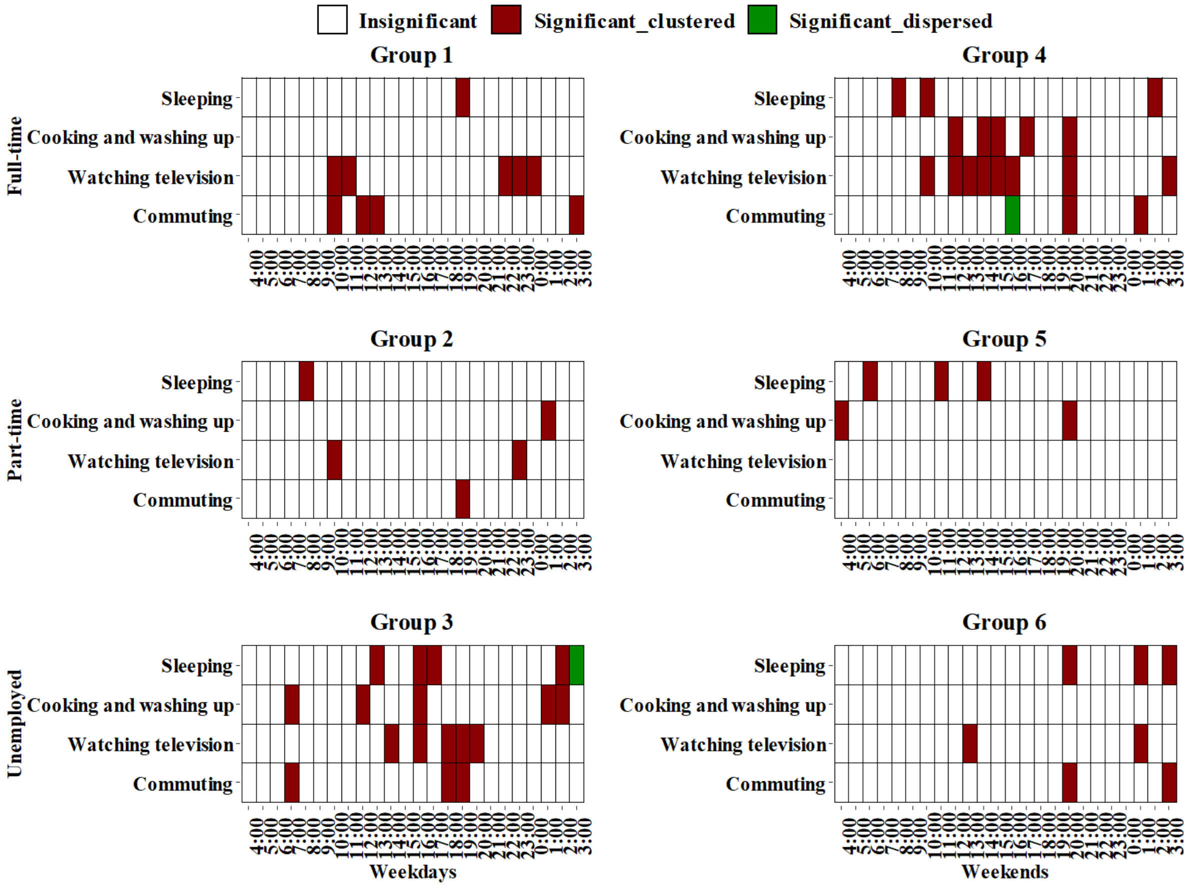


Figure 4-3. Results of the Moran’s I test considering the representative activities for each group. The time intervals listed in the table are the intervals during which spatial variation existed.

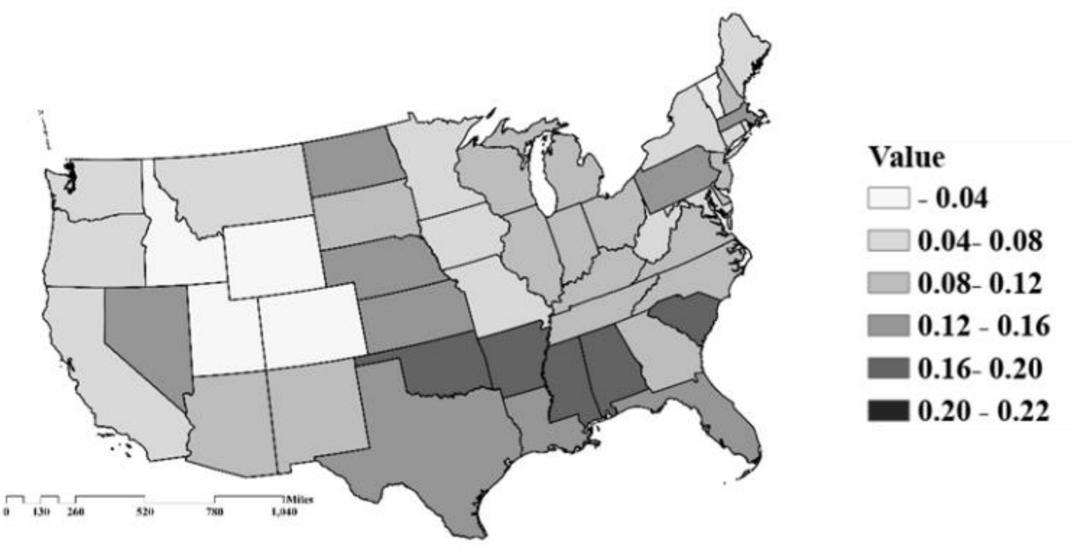


Figure 4-4. The spatial distribution of the probability of undertaking watching television for Group 6 at the 13:00 time interval at the state level based on observations.

4.4.2 Representations of spatial variation

Figure 4-5 shows the spatial probability of the women in Group 6 watching television at 13:00 based on the representations of the spatial variation generated by the ordinary kriging and SAR methods. The representation generated by ordinary kriging (i.e., kriging-based representation) ranges from 6 to 14% whereas the representation generated by SAR (i.e., SAR-based representation) ranges from 4 to 17%. The variation was narrower than the observation shown in Figure 4-5. The kriging-based representation can simulate the changing tendencies of spatial probabilities. However, the clustered pattern was not identified. The SAR-based representation can provide more accurate estimations for certain states, simultaneously providing a better representation of the cluster areas.

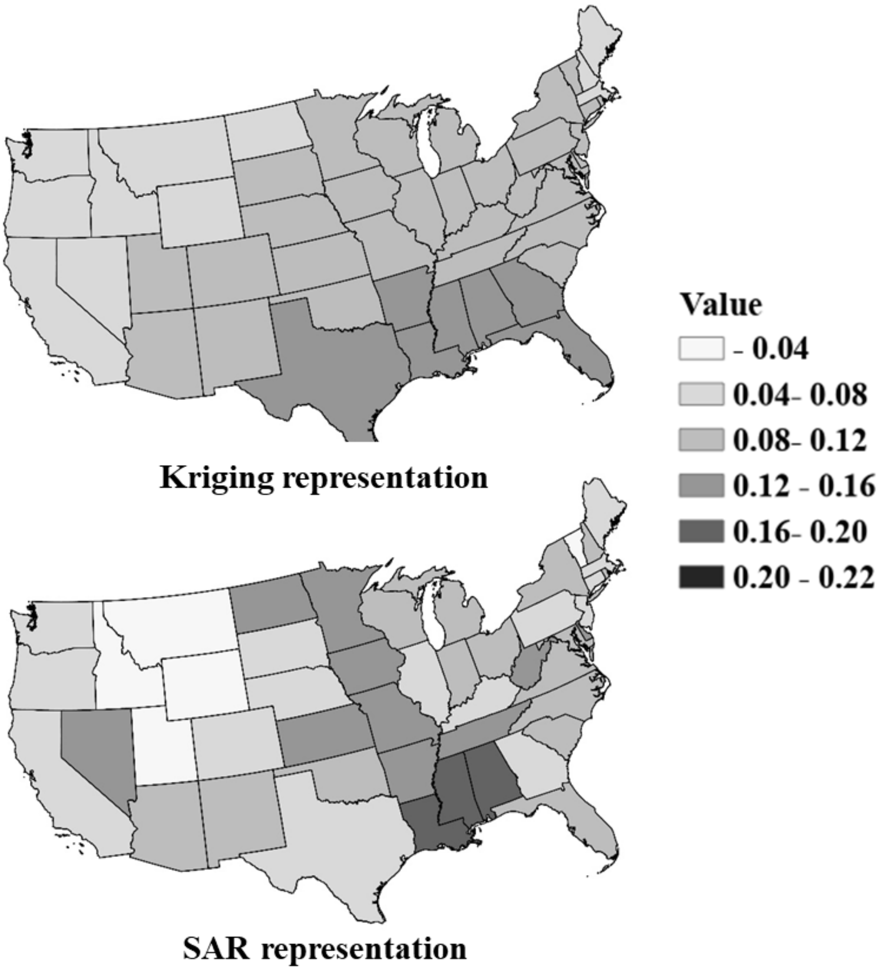


Figure 4-5. The spatial distribution of the probability of undertaking watching television for Group 6 at 13:00 based on representations of spatial variation obtained from the ordinary kriging and SAR.

Furthermore, we also compared the two representations considering all the combinations of groups, activities, and states. Figure 4-6 shows the comparison between the kriging-based and SAR-based representations for all combinations of the group, activity, and state. Two representations and the observations were conducted with the base-10 logarithmic transformation. Two R^2 values with and without the logarithmic transformation were also presented in the figures. According to Figure 4-6, the kriging-based representations are more scattered than the SAR-based. The R^2 without and with logarithmic transformation (R^2 and R_{log}^2 respectively) is 82.5% and 24.0% for kriging-based representations and 98.3% and 84.9% for SAR-based representations. Regarding TAE and RMSE at the state level, the kriging-based representation was 126.5 and

9.9%, and the SAR-based representation was 61.2 and 3.0% respectively. These results implied that the SAR method can generate more accurate representations of the spatial variation at the state level than ordinary kriging.

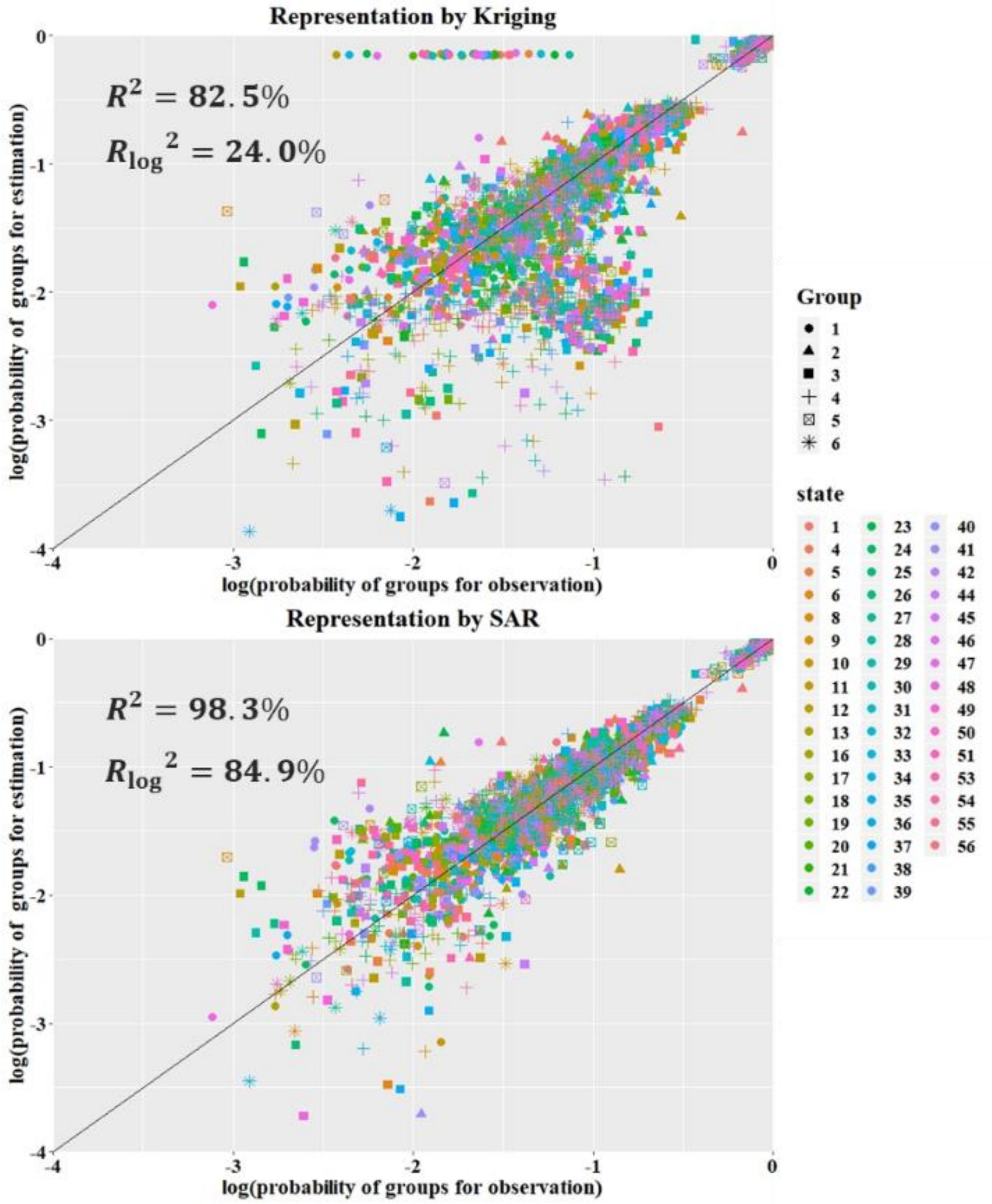


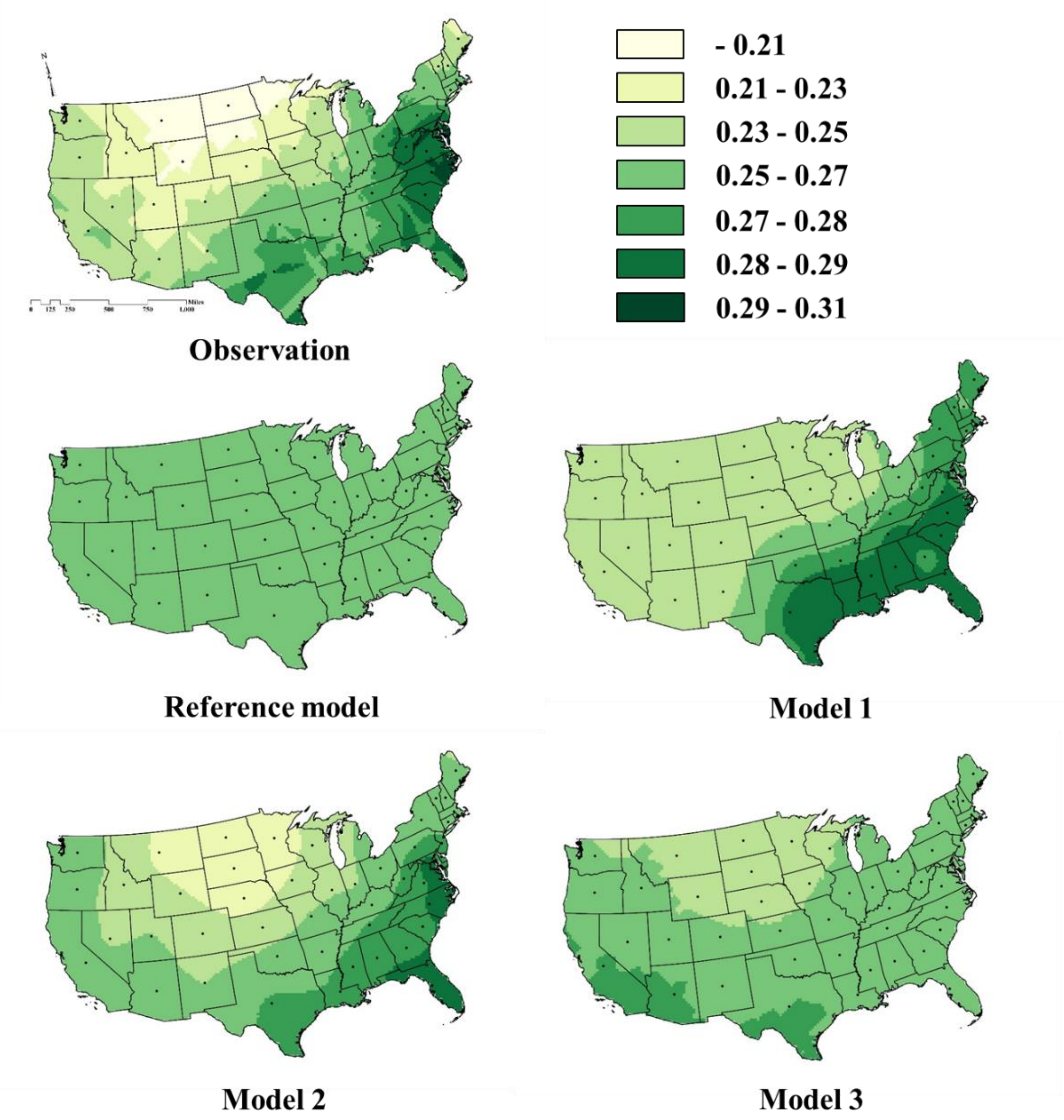
Figure 4-6. The accuracy of the representations of the spatial variation at the state level. The horizontal axis shows the observation probabilities of combinations of group, activity, and state. The vertical axis shows the representations. The black line indicates the reference line $y = x$. Logarithmic transformation was conducted in the range (–4, 0)×(–4, 0).

4.4.3 Spatial logistic regression models

4.4.3.1 Reproduction of the spatial variation

The reproducibility of the spatial variation by the developed spatial logistic regression models was evaluated based on four representative cases: (a) sleeping at 8:00 in Group 4; (b) cooking and washing up at 12:00 in Group 3; (c) watching television at 13:00 in Group 6; (d) commuting at 10:00 in Group 1. Figure 4-7 illustrates the spatial probability of activity in each of the four cases, based on the observations and estimations. The visualization of the spatial variations in all the subfigures was interpolated using the ordinary kriging method. Considering the reproduction of the spatial variations in these four cases, the spatial distributions determined by the three spatial logistic regression models were more consistent with the observations than those determined by the reference model. However, Model 2 for Case (b) and Model 3 for Case (c) yielded the same results as that of the reference model. This is because the spatial representations, $g(s_i;\theta)$, were eliminated during the stepwise process. The reference model also showed limited spatial variations (see subfigures in Figure 4-7 for Cases (b) and (c)), which is attributed to the variations in demographic factor variables.

Case (a)



Observation

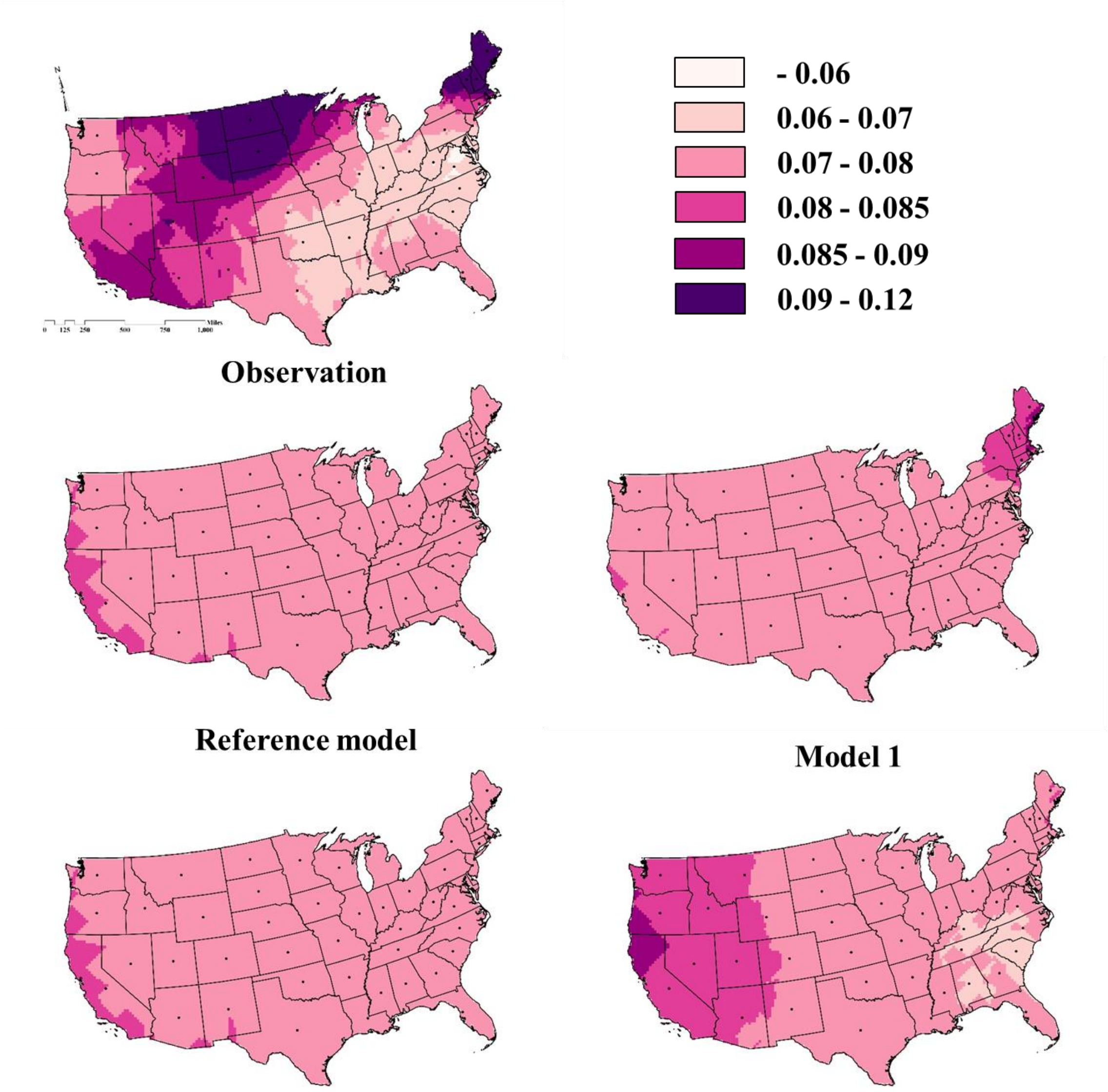
Reference model

Model 1

Model 2

Model 3

Case (b)



Observation

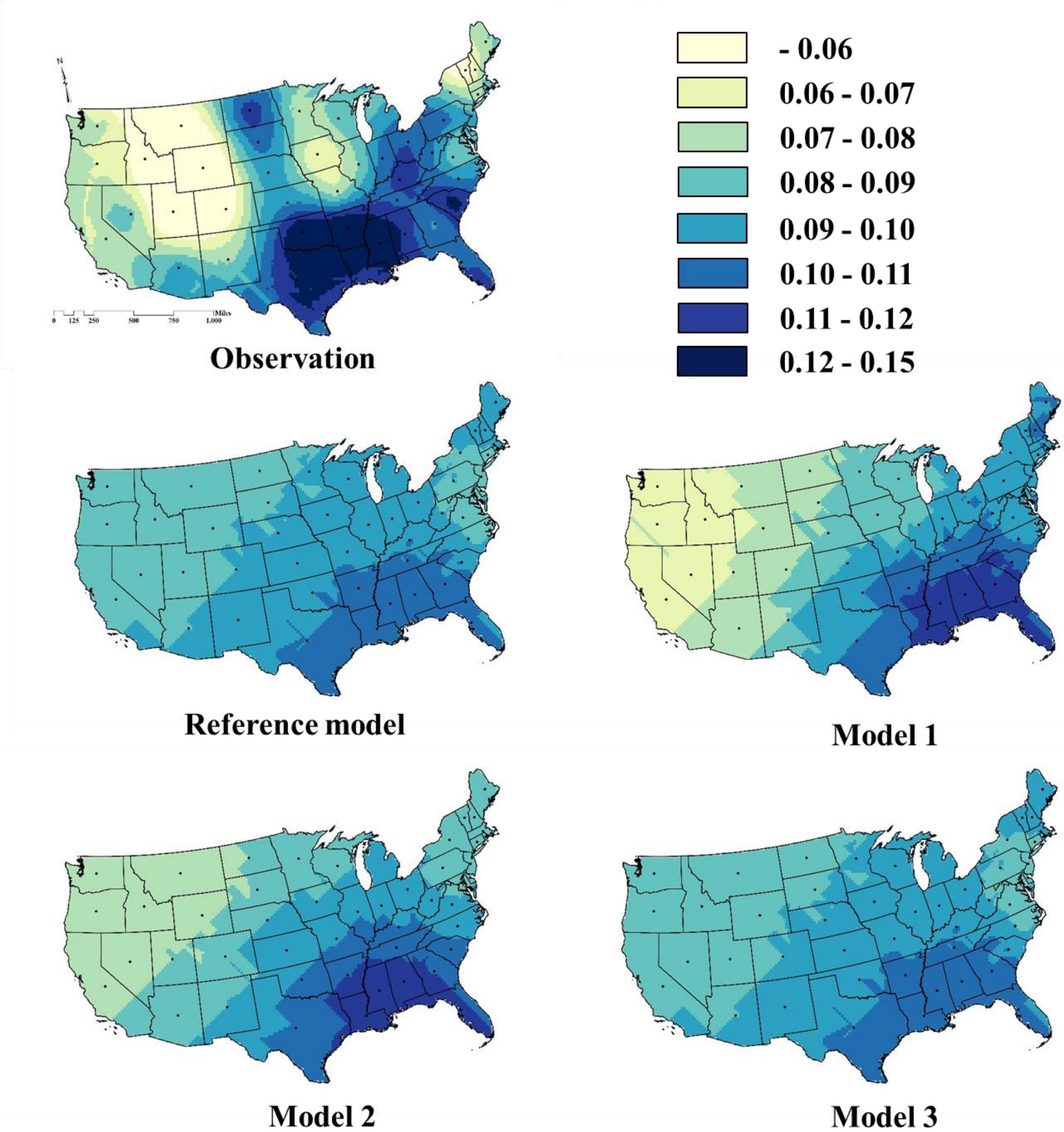
Reference model

Model 1

Model 2

Model 3

Case (c)



Observation

Reference model

Model 1

Model 2

Model 3

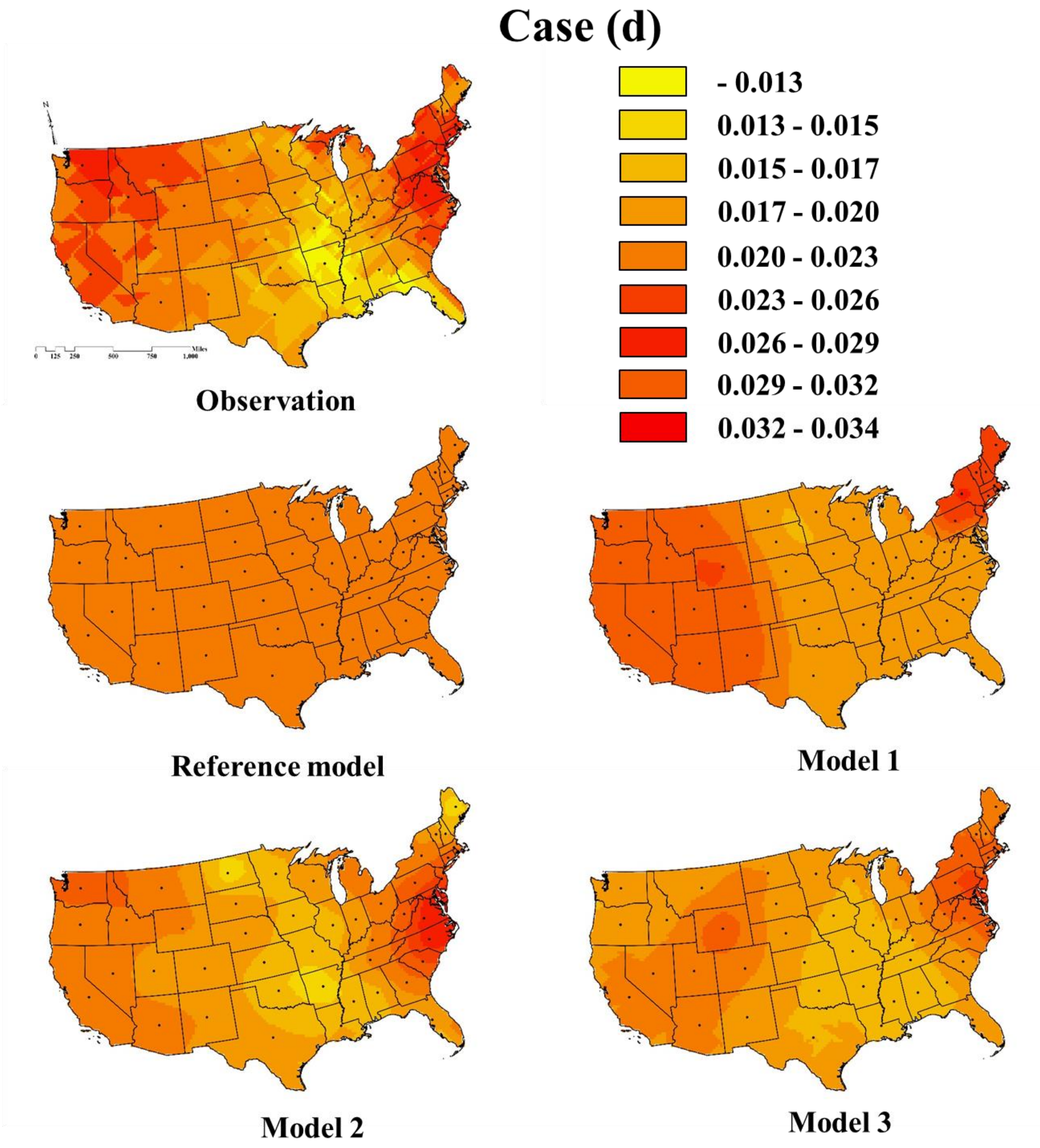


Figure 4-7. The spatial distribution of the undertaking probability based on observations and the reproductions of the spatial variation by the reference model and three developed spatial logistic regression models for Case (a)–Case (d) respectively.

4.4.3.2 Comprehensive performance

Figure 4-8 shows the stacked values of performance indicators quantified at the national level,

$\text{TAE}_{\text{nation}}$, $\text{RMSE}_{\text{nation}}$, $\text{MSD}_{\text{nation}}$, and $\text{RMSE\_GA}_{\text{nation}}$, for all the models considering the six groups in the training and test datasets. The indicators are the cumulative values quantified for each activity and group combination. As shown, the error indicators exhibited similar performances with all the models for almost all of the combinations in the training and test datasets. Considering the diversity, Models 1 and 3 exhibited 7% higher $\text{MSD}_{\text{nation}}$ than the reference model. Considering $\text{RMSE\_GA}_{\text{nation}}$, all the models exhibited similar results with both the training and test datasets.

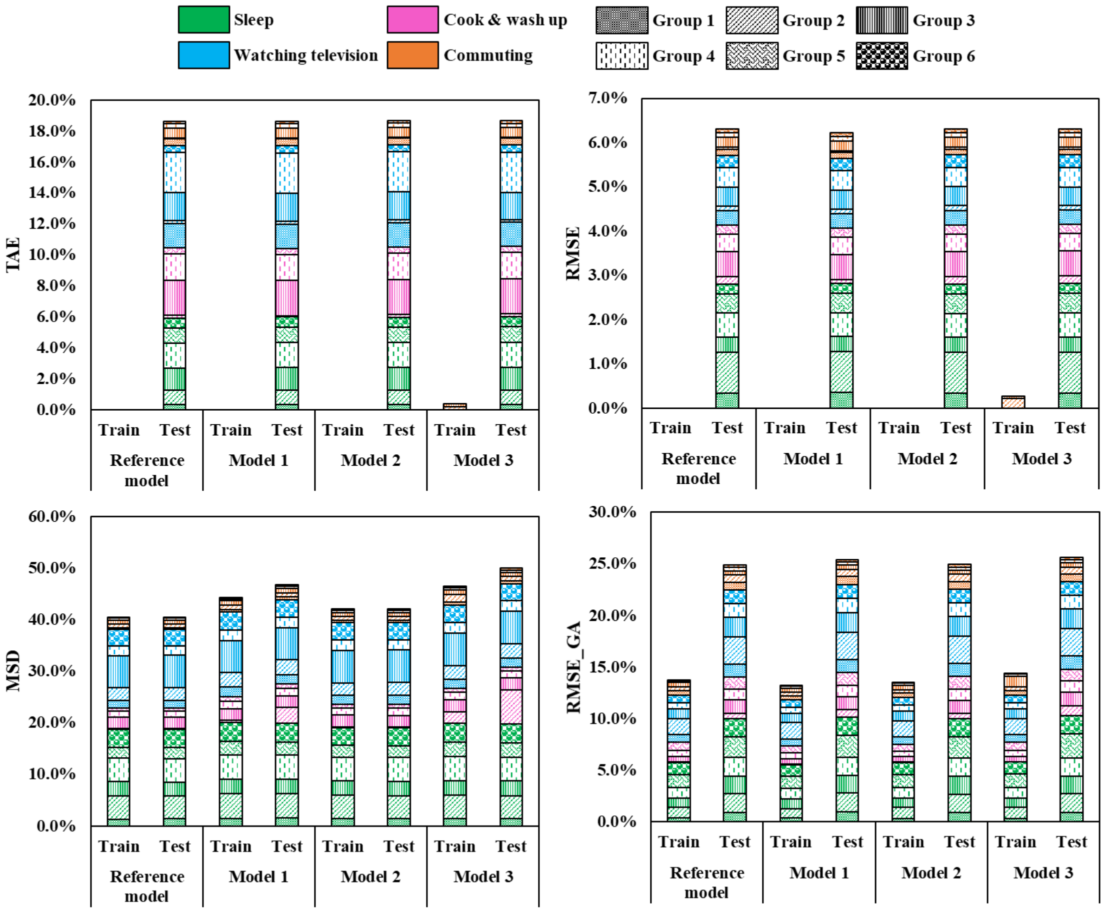


Figure 4-8. The results of indicators at the national level for all models in the training and test sets.

Figure 4-9 illustrates the TAE, RMSE, and MSD values of the models for the six groups quantified at the state level. As shown in Figure 4-8 and Figure 4-9, the magnitudes of the error indicators increased from the national level. However, MSD exhibited the opposite trend. Considering the error indicators, improvements were observed in the spatial logistic regression models compared to the reference model. Model 3 exhibited the greatest improvement compared to the reference model, reducing the stacked $\text{TAE}_{\text{state}}$ value by 9.9 and the stacked $\text{RMSE}_{\text{state}}$

value by 11% for the training dataset. This was followed by Model 1 (stacked $\mathrm{TAE}_{\mathrm{state}}$ decreased by 4.4 and stacked $\mathrm{RMSE}_{\mathrm{state}}$ decreased by 3.6%) and Model 2 (stacked $\mathrm{TAE}_{\mathrm{state}}$ decreased by 3.2 and stacked $\mathrm{RMSE}_{\mathrm{state}}$ decreased by 2.1%). However, the spatial logistic regression models, particularly Model 3, did not provide such advantages with the test dataset. The unexpected performance in the test set was mainly due to the sample size for each state being small. Although the test set had the same probabilities as the training set at the national level, the probabilities among states were quite different as shown in Figure 4-13 and Figure 4-14 in Appendix 4.7. Considering MSD, the spatial logistic regression models, particularly Models 1 and 3, performed better than the reference model with both the training and test datasets.

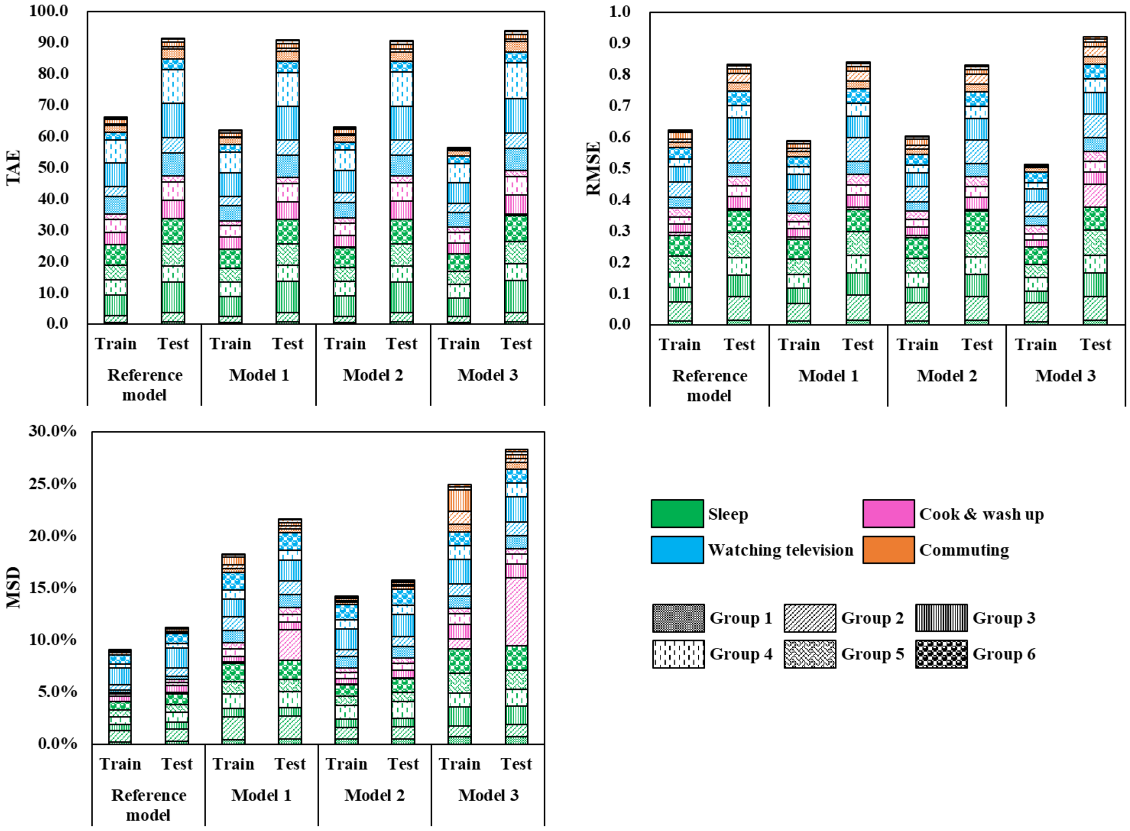


Figure 4-9. The results of indicators at the regional level and state level for all models in the training set.

These results are confirmed in Figure 4-10, which shows the accuracy evaluations of each model at the state level. The estimations and observations were obtained using the base-10 logarithmic transformation. Two R^2 values, with and without logarithmic transformation, were quantified. All the models exhibited high accuracies. However, the points in the reference model were relatively scattered compared to those in the spatial logistic regression models. Considering the values of R^2, the spatial logistic regression models, especially Model 3, exhibited relatively higher R^2 values than the reference model.

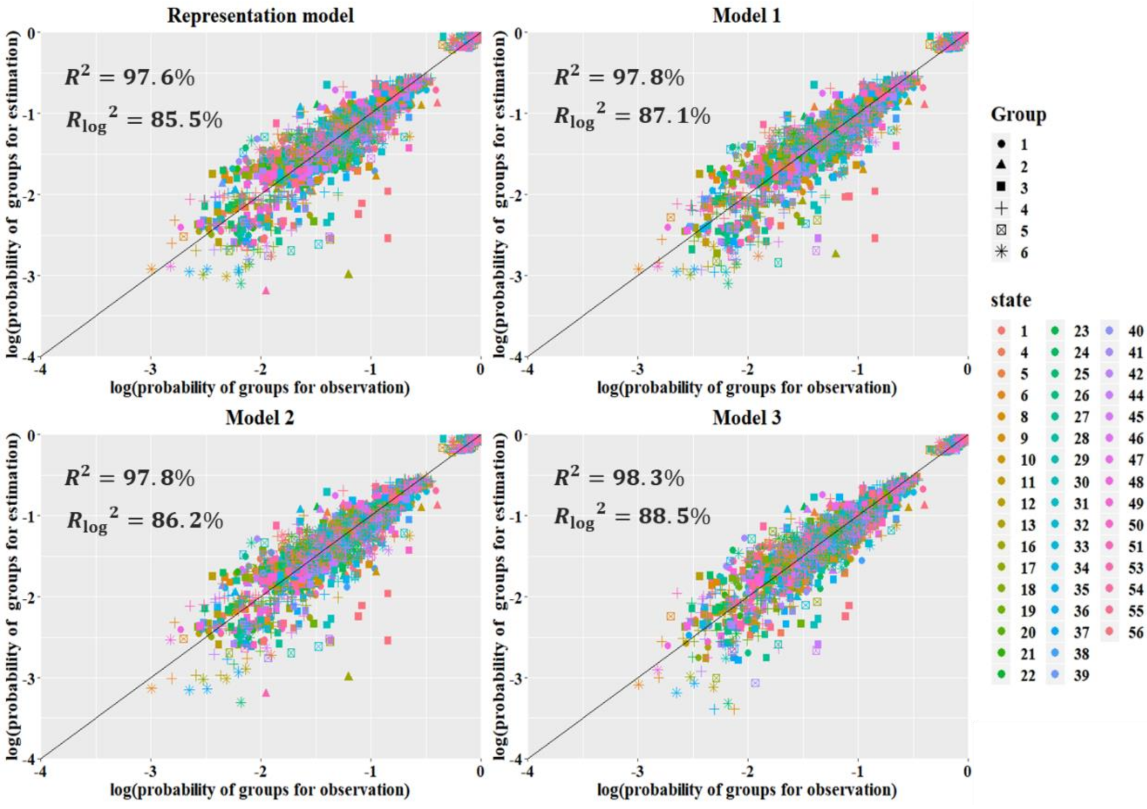


Figure 4-10. The accuracy of the spatial logistic regression model at the state level. The horizontal axis shows the observation probabilities of combinations of the group, state, and activity. The vertical axis shows the estimations. The black line indicates the reference line $y = x$. Logarithmic transformation was conducted in the range (–4, 0)×(–4, 0).

4.4.4 Evaluation of spatial logistic regression models applied to the entire population

4.4.4.1 Application of Group 7

In this subchapter, the spatial logistic regression model was applied to Group 7 (i.e., the entire population of women aged between 30–59 years) for watching television activity. The Moran's *I* test results indicated that spatial variation existed during the time intervals 9:00–17:00 and 22:00–0:00. Therefore, the spatial logistic regression models were developed and assessed only for these time intervals.

Figure 4-11 shows the same visualization maps (see Figure 4-7) of the spatial probability of watching television at 13:00 based on the observations and estimations of Group 7. The range of probability is narrower than Figure 4-7 for Group 6 because Groups 1–6 were combined. The spatial logistic regression models, especially Model 3, showed a more accurate spatial distribution relative to the observations than the reference model. Table 4-3 shows the performance of all of the models evaluated by the indicators, considering all the time intervals that exhibited spatial variation. The models performed effectively with Group 7. At the national level, all the models exhibited the same performance in terms of errors and MSD. However, the reference model showed a relatively lower $\mathrm{RMSE\_GA}$ compared to the spatial logistic regression models. At the state level, the spatial logistic regression models exhibited lower TAE and RMSE values, and similar MSD values to the reference model.

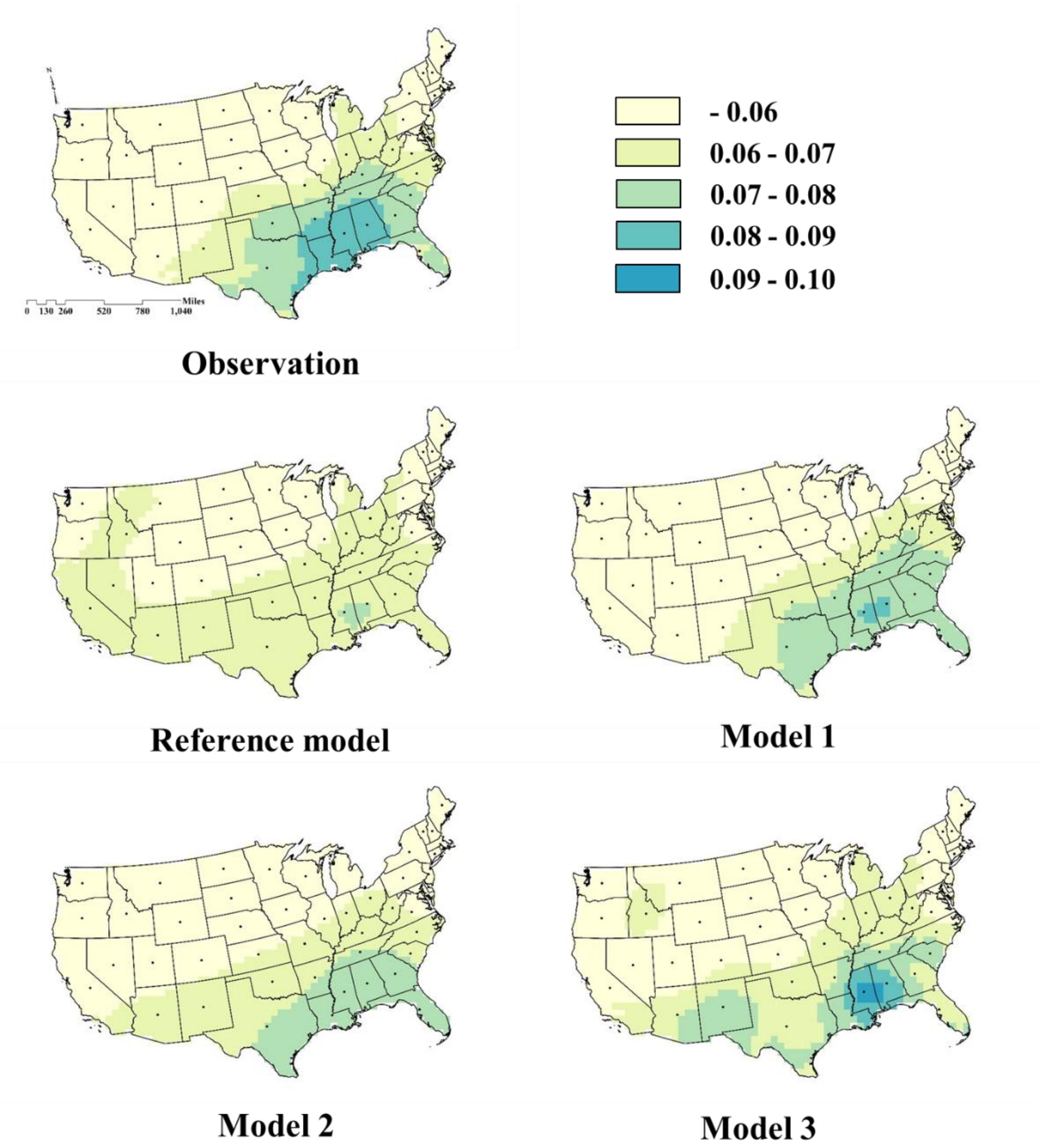


Figure 4-11. The spatial distribution of the undertaking probability based on observations and estimations for watching television at the 13:00 time interval of Group 7.

Table 4-3. Results of indicators considering all the models with Group 7 at the national and state levels. RMSE_GA was calculated only at the national level.

Level	Indicator	Reference model	Model 1	Model 2	Model 3
National	TAE	0.0%	0.0%	0.0%	0.0%
	RMSE	0.0%	0.0%	0.0%	0.0%
	MSD	0.6%	0.6%	0.6%	0.6%
	RMSE_GA	3.3%	3.4%	3.5%	3.4%
State	TAE	7.6	6.3	6.0	6.1
	RMSE	1.7%	1.5%	1.4%	1.4%
	MSD	3.2%	3.2%	3.1%	3.1%

4.4.4.2 Comparison approach of segmentation and using conditions as variables

Figure 4-12 depicts the accuracy in the base-10 logarithmic transformation of Model 3 for watching television, considering Group 7 and different subpopulations at the state level. Only the time intervals that exhibited spatial variation considering Group 7 and the subpopulations of Groups 1–6 have been considered in this analysis. Model 3 developed for Group 7 was applied to certain subpopulations Groups 1–6 corresponding to the different time intervals to represent estimations based on Group 7. The thick black line shown in the two subfigures of Figure 4-12 represents the fitted line of the estimations obtained from Model 3 considering Group 7, which indicates the approach that uses variables, and the thick dashed line represents the estimations obtained from Model 3 considering the subpopulations, which indicates the approach using segmentation. The thin black line is the reference line, $y = x$. Model 3 considering both the entire population and the subpopulations fitted significantly with the observations. However, the thick dashed line was slightly closer to the reference line than the thick black line, which implies that the estimations obtained from Model 3 through segmentation were more accurate than those obtained from the variable-based approach.

Table 4-4 shows the comprehensive performance comparison through the statistical indicators of the two approaches for all models at the state level. According to Table 4-4, all models performed adequately for both approaches. However, the segmentation-based approach yielded smaller TAE and RMSE for all the models. In contrast, for the diversity assessed by MSD, the variable-based approach showed a relatively better performance.

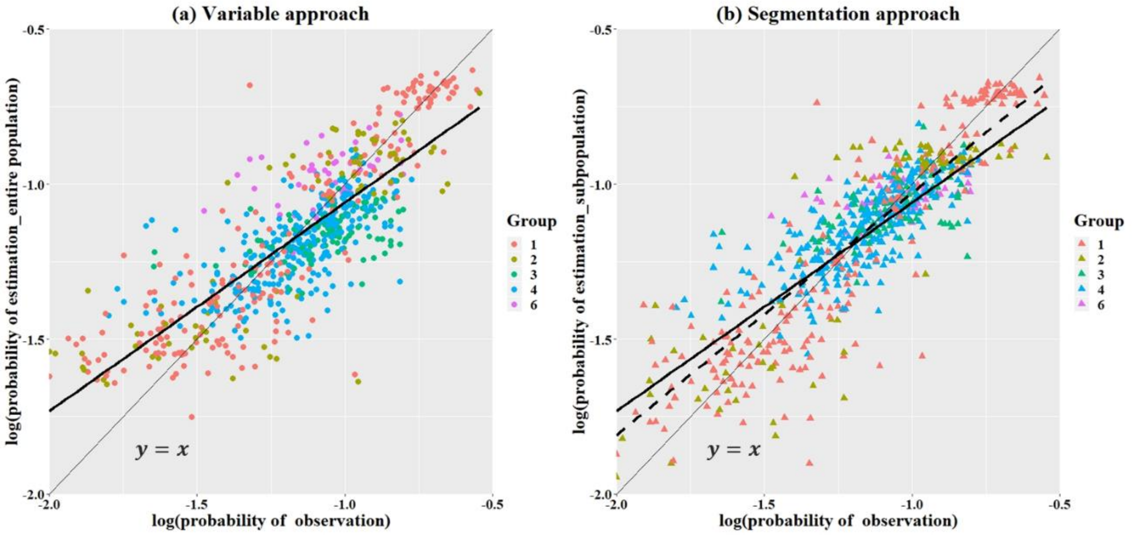


Figure 4-12. Accuracy of Model 3 at the state level, considering two approaches (variables and segmentation). The different colors in the figure represent different groups. The circular and triangular shapes represent the entire population and the subpopulations, respectively. Logarithmic transformation was performed in the range of (−2, −0.5) × (−2, −0.5).

Table 4-4. Comparison of the approaches through statistical indicators at the state level.

Approach	Group	Indicator	Reference model	Model 1	Model 2	Model 3
Segmentation-based	Subpopulation Group 1–6	TAE	17.8	16.4	16.6	15.3
		RMSE	3.3%	3.1%	3.1%	2.9%
		MSD	0.4%	1.1%	1.0%	1.3%
Variable-based	Entire population Group 7	TAE	18.8	17.6	17.3	17.4
		RMSE	3.4%	3.2%	3.1%	3.1%
		MSD	0.5%	1.0%	1.2%	1.3%

4.5 Discussion

The Moran’s *I* tests in Chapter 4.4.1 showed that spatial variation exists and it differed according to the time of day and activity for different study populations. Therefore, spatial variation should be carefully considered in OB modeling. To this end, SAR-based and kriging-based spatial representations were developed to better represent spatial variation empirically and used in subsequent spatial logistic regression models. The results in Chapter 4.4.2 showed the SAR-based representation was superior to the kriging-based representation because the former accounts for the variation in other demographic factor variables. If the location data required to develop a spatial representation is insufficient, spatial factor variables can be used to represent spatial variation for model development, as in the case of Model 1.

As discussed in Chapter 4.4.3, the developed spatial logistic regression models improved the diversity, as the single-activity MSD for subpopulations improved by 0.6%, and the stacked MSD for all combinations improved by 12.5% at the state level with the training dataset compared to the reference model. In particular, the developed models better reproduced the spatial variation of OB, as the error was further reduced (RMSE decreased by 0.3%, and stacked RMSE decreased by 5.6%). Furthermore, we compared the two approaches for model development: variable-based and segmentation-based. As discussed in Chapter 4.4.4.2, the variable-based approach can be an effective substitute for the segmentation-based approach for further grouping, because it can approximately reflect the diversity, and the error was only marginally larger than the segmentation approach (the stacked TAE and RMSE increased by 1.3 and 0.1%, respectively).

This study showed the existence of spatial variation in OBs and established a new modeling method to consider spatial variations in OBs, which may contribute to better reproducing the spatial variation in building energy demand while maintaining high accuracy. A limited sample extracted from ATUS data representing women from states of the U.S. mainland and the inaccurate low-resolution location data were used. Therefore, Models 2 and 3 exhibited the same performance as the reference model in several cases, whereas the developed model showed no significant improvement with the test dataset. Nevertheless, the developed model can easily be applied to different regions or countries, as the national level time use data have been collected in many countries. However, the detailed information relevant to the housing, households, and environment should be supplemented by combining the data collected at the local level. In addition, reliable new samples ought to be generated to enrich the sample size similarly. Furthermore, the advancements in geographic information systems allow the higher resolution location data to become more and more available. Thus, if adequate data is available (i.e., rich information of occupants, higher resolution location data, sufficient sample size), spatial representations can be generated with higher accuracy at the zip code or even household level. Therefore, subsequent spatial logistic regression models can facilitate further improvements.

4.6 Conclusion

Existing OB models lack a comprehensive and systematic consideration of spatial variation. These models were primarily established within limited locations based on geo-referenced data to determine space use or to simulate occupant mobility. Some studies used spatial factor variables to insufficiently consider the spatial variation in OBs or energy demand. However, the real spatial distribution of OBs has not been comprehensively investigated, and modeling methods that reproduce spatial variation in OBs are yet to be developed.

This study showed that spatial variation exists in OBs and developed new OB models that can incorporate spatial variation. The developed models significantly enhanced the reproducibility of spatial variations in OBs and generated smaller errors at the state level than the conventional logistic regression model. Particularly, the developed models can be applied in different countries for any application context (i.e., any spatial scale and population). However, our results were obtained with limited samples from the ATUS data and low-resolution location data. The performance may be improved when the following requirements are satisfied: the high-resolution location data, behavioral data with richer information, and sufficient sample sizes. Therefore, with more comprehensive considerations of spatial variation in the new OB model, location-based OB patterns can be generated, which can be used in future studies to simulate more realistic energy demand profiles and to develop region-sensitive energy policies.

4.7 Appendix

Figure 4-13 and Figure 4-14 show the representative probabilities with the training set and the test set for the women in Group 4 sleeping, those in Group 3 cooking and washing up, those in Group 6 watching television, and those in Group 1 commuting. According to these two figures, we found the probabilities at the national level are similar, however, the probabilities among the states are quite different between the training set and the test set.

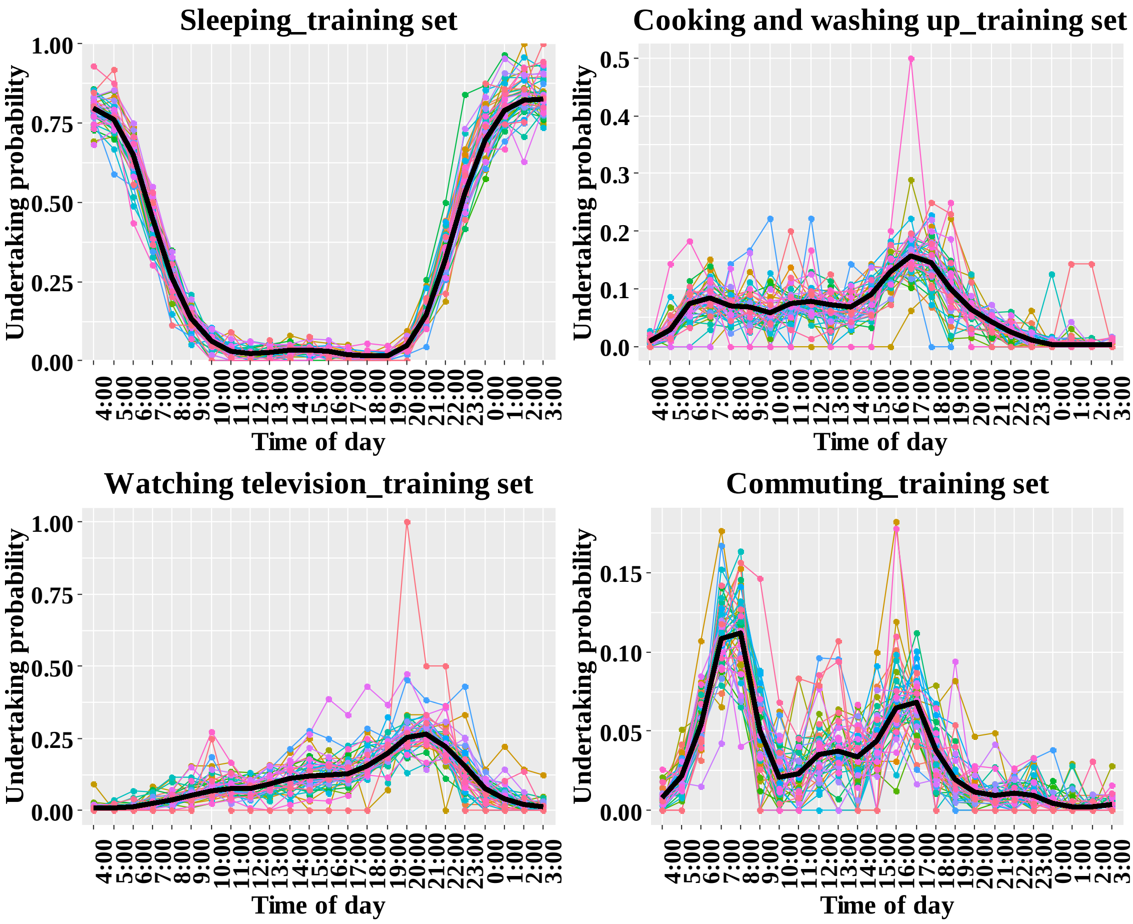


Figure 4-13. Probability of undertaking activities for representative groups with the training set. Lines in different colors indicate different states and the black one indicates the national level estimations.

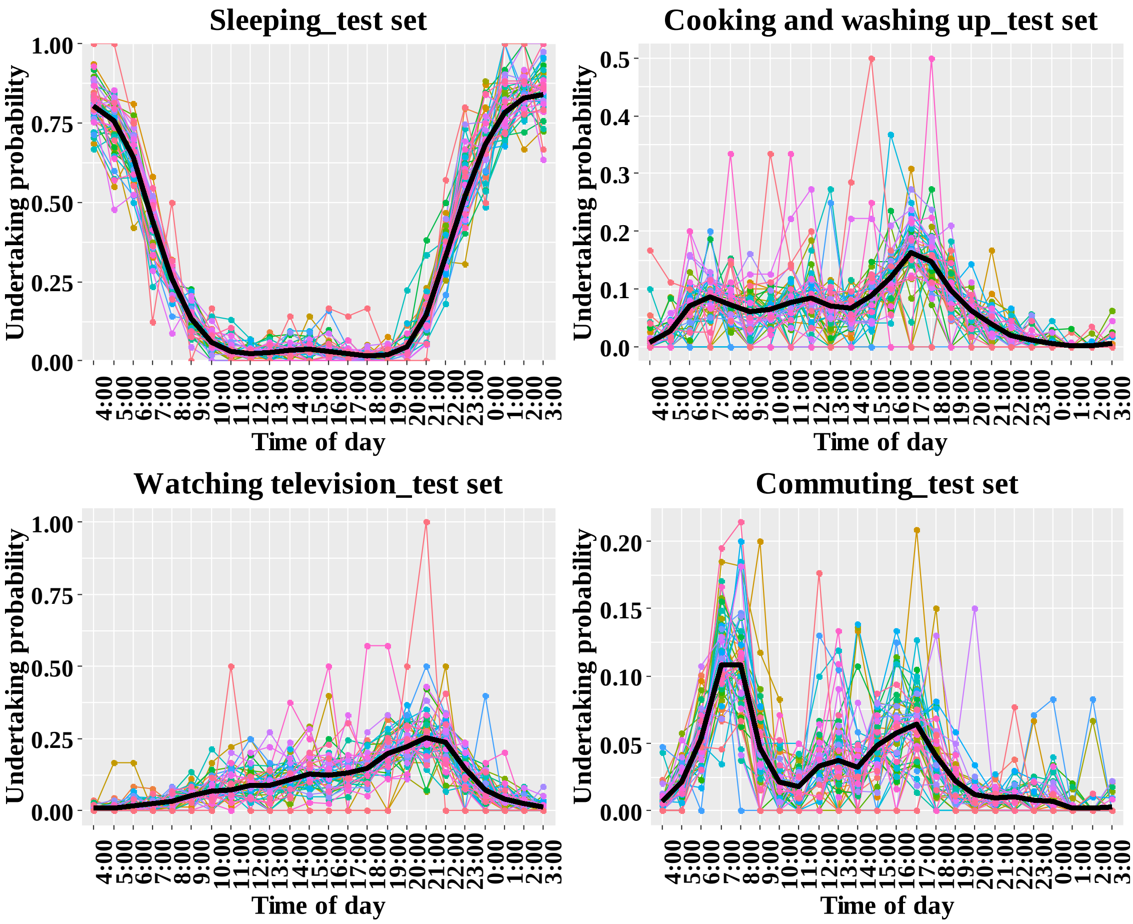


Figure 4-14. Probability of undertaking activities for representative groups with the test set. Lines in different colors indicate different states and the black one indicates the national level estimations.

5 Integrated discussion

This chapter presents an integrated discussion of this thesis in Chapter 5.1. In particular, the discussion focuses on the extended scientific knowledge in the field in terms of the modeling process, diversity represented by model input, and the modeling method. The limitations of the new OB model developed in this thesis and the potential for future research are discussed in Chapter 5.2

5.1 Discussion

This thesis conducted three studies to accomplish the research aims. By analyzing the overall development of the OB model underpinning the energy demand simulation, the significance of pre-simulation was revealed. Choices including the selection of the variable, selection of the parameter preparation method, and the model engine had large influences on model estimations. Logistic regression models with an appropriate number of significant variables can provide relatively good performance in terms of the error in the actual application context. However, it is incapable of enabling the reproduction of spatial variation in OBs. There are several elements to the proposed new OB model — spatial logistic regression model which favors it over the current predominant logistic regression models. First, the error and accuracy can be further improved at a higher spatial level. Second, the diversity among simulated occupants can be enhanced. Finally, the spatial probability of OBs can be better reproduced. The further discussion that might help address the development of energy demand simulation with consideration of OBs and encourage the investigation of diversity are outlined below.

5.1.1 Pre-simulation process

The pre-simulation process of OB-based energy demand simulation contains the data processing, variable selection, and parameter preparation which was mentioned in Chapter 1.3. Previous studies have paid less attention to this process and the majority of them focused on the modeling engine part. When considering the pre-simulation process, various methods were applied to different data with items of interest to prepare the parameters for the engine. More importantly, sample distribution is dominantly used to prepare the parameters in which diversity is ignored. In addition, no detailed explanation was made to guide other researchers to start as the first step for developing models in the previous studies.

However, our study in Chapter 2 showed that the pre-simulation process needed to be carefully considered thus developing more comprehensive research frameworks for simulating the OB patterns. In our study, two key sub-processes which were variable selection and parameter preparation method selection were analyzed based on the commonly used TUD. Nine models which are combinations of three parameters preparing methods (i.e., multinomial log-linear regression (MLR), support vector machines (SVMs), and feedforward artificial neural network (ANN)) and three cases of different variable conditions (i.e., type and number of the variable) were designed to assess the impact of the pre-simulation process on OB modeling. The selections of the methods and the variables were based on the concrete literature review and available data. More specifically, we assessed the model performances by three aspects (i.e., average performance, diversity, and individual specificity) which were represented by well-designed indicators.

The analysis suggested that the OB modeling performance can be affected by the pre-simulation process. Data with significant variables and enough sample size can better support the model simulation. Regarding the parameter preparation method, all selected methods had a quite good and similar performance. However, the running time of each method varied. Therefore, the selection of the method should consider the actual situation (data in hand and the speed of the computer). As for the variable, compared to the factors such as demographic and time factors that were commonly used in the previous study, our finding suggested considering the appliance and activity factor in the modeling process because variables from these factors are proved to be significant. Simply increasing the variable may not improve the model performance as we observed in one of our cases. Only considering more significant variables can improve the diversity among the simulated occupants.

5.1.2 Enhancement of diversity

As mentioned in Chapter 1.4, some researchers point out that diversity among the occupants can explain a great part of the gap between reality and simulations in both OB modeling and energy demand modeling. Previous studies have tried to deal with the diversity issue by conducting segmentation in terms of different conditions such as type of day, housing unit type, and other sociodemographic information. Such segmentation techniques are quite useful and efficient to enhance the diversity among all the simulated occupants as proved by the findings in Chapter 4. Also, as shown in Figure 5-1, different women subpopulation groups represented by different colors have different density distributions in conducting the activity of watching television. Therefore, developing a specific model for each group can better capture the characteristics of each group and reflect the discrepancies among the groups. However, the drawback of segmentation is that it requires enough sample size otherwise it will draw unrealistic conclusions. The findings in Chapter 4 illustrate that only basic segmentation is needed if the sample size is enough. Further detailed segmentation is unnecessary as it can be replaced by using variables in the model. However, we also found that increasing the variable may deteriorate the model performance as shown in Chapter 2. Only considering significant variables can improve the diversity while maintaining the model accuracy. Therefore, the selection of the variables is a crucial part of the modeling which also needs to be carefully designed.

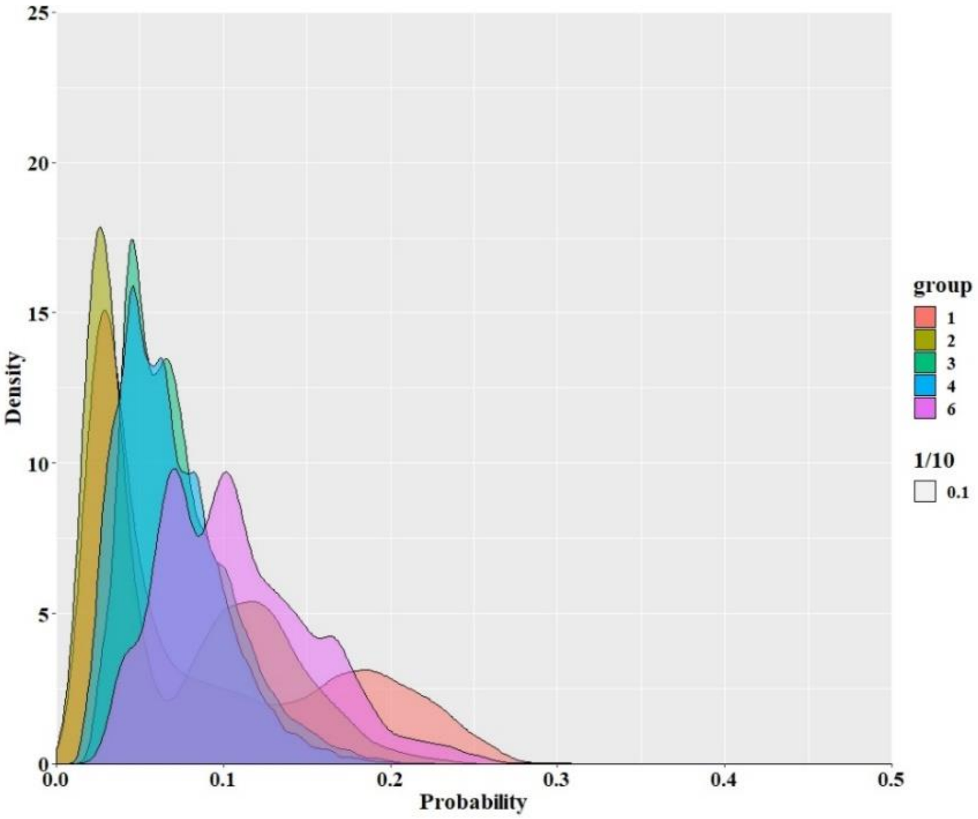


Figure 5-1. The density of probability of undertaking watching television for each group in the case study used in Chapter 4.

Here should be noted that diversity refers to the total variation among simulated occupants. However, diversity contains various aspects such as the difference in attributes of the people, spatial variation, temporal variation, and socioeconomic cultural differences. These aspects are not all independent, but there is a certain connection between them. For example, Figure 5-2 shows the spatial probability distribution of undertaking watching TV at the 13:00 time interval for Group 6 which is defined in Chapter 4, where the time zone divided by the red line is marked. Although the undertaking probability was simulated by the local time, the time zone may still affect occupants conducting watching television, especially watching some live shows. Thus spatial and temporal variations (i.e., spatiotemporal variation) may be considered and processed simultaneously to better reflect reality. As mentioned before, researchers have realized that diversity is one of the important causes resulting in the gap between simulation and reality in recent decades. However, their focus was on the simple differences in the attributes of the research objectives (i.e., occupant, building, and type of day). The long-term temporal variation (e.g., historical change) and spatial variation were seldom fully analyzed although some researchers revealed that these variations should be highlighted. In Chapter 3, our study verified that the historical change and spatial variation existed for certain OBs. The findings also pointed out that certain OBs may have a clustered or dispersed probability distribution over space at a specific time. Such integrated variations should be paid more attention to in the modeling in future research to improve the accuracy and quality of the simulation estimations.

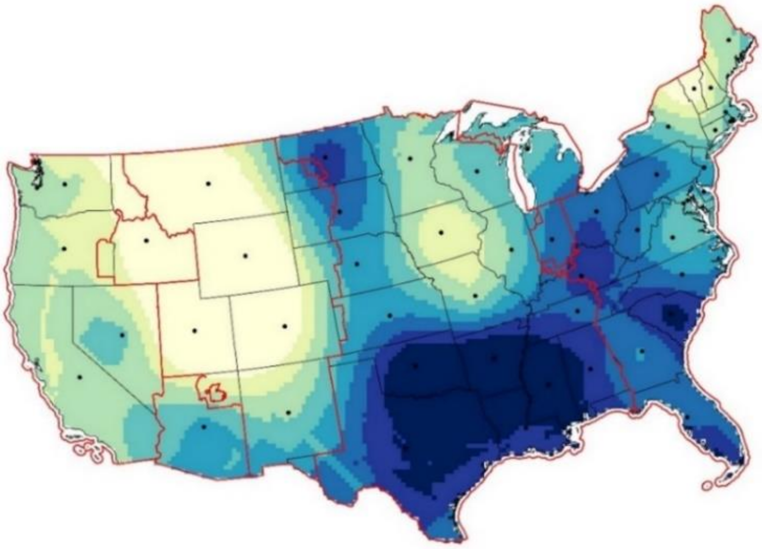


Figure 5-2. The spatial probability of undertaking watching television for Group 6 at 13:00 in the case study used in Chapter 4. The red line divides several major time zones for the mainland of the U.S.

5.1.3 Conventional logistic regression and spatial logistic regression model

Previous OB studies considering the discrete-event approach commonly used the logistic regression model to quantify the probability of starting or undertaking an activity. The logistic regression model has been proved very useful and efficient. It can take various factors to distinguish the simulated occupants in the modeling process. However, it cannot fully reproduce the variation exhibited in the activity conducted by occupants as we analyzed in this thesis. An obvious example is that only limited variation in OBs was reproduced in a given space as mentioned in Chapters 3 and 4.

Compared to the conventional logistic regression model, the developed spatial logistic regression model based on knowledge from the geostatistical field can involve the consideration of spatial variation. Actually, the spatial logistic regression model is an extension of logistic regression. It incorporates a smooth function (e.g., estimation of the spatial autoregressive model or ordinary kriging model, spatial factor variable which are introduced in Chapter 4) that represents the spatial variation in the modeling process. In Chapter 4, our study revealed that the spatial logistic regression model can better reproduce the spatial variation in OBs. Moreover, the spatial logistic regression model can generate estimations with smaller errors and higher diversity at higher spatial resolution. As no sufficient location was available for this study, the spatial logistic regression model had the same results as the logistic regression model for several cases. But we believe that the developed spatial logistic regression model may obtain further improvement once adequate data especially higher resolution spatial data is available.

In addition, the developed spatial logistic regression model only involved spatial variation. The impact of the temporal aspect on the OBs is undeniable, no matter the long-term temporal variation (i.e., historical change, annual change) or mid-term temporal variation (i.e., seasonal change, monthly change). More importantly, as mentioned in Chapter 5.1.2, spatial variation and temporal variation may not independently influence the OBs. Thus, a novel model is urgently required to consider the variation from both spatial and temporal aspects for modeling OB and energy demand in further research.

5.2 Limitations and future work

This thesis aims to investigate how can an OB model be developed and improve the OB model from a spatial perspective. It would contribute to the dynamic building energy demand estimation thereby providing useful references in both industry and academia. Summarised below are the limitation of the studies in this thesis:

1. The studies are limited to demographic and time factors and don’t include environmental and psychological factors. The physical environmental factor variables are usually considered in the energy demand models. Building’s physical conditions such as the wall material and the neighbor's environmental conditions such as trees can largely influence people's daily behavior. However, considering these factors again in the OB model may be redundant. Therefore, the significant variables of the OB model should be selected based on a compromise of variables considered in all sub-modules of energy demand modeling. Likewise, the psychological health of an occupant influences their comfort and enthusiasm for activities such as working. However, these psychological factor variables are challenging to collect and may fluctuate rapidly and vary from person to person.

2. All studies are conducted based on ATUS. Literature suggests that an occupant tends to

respond positively to questions. However, we only extracted the neutral question (i.e., demographic factor variables and activity records) from the survey. Still, ATUS contains the TUD only recording the representative individual in each household. The activity patterns for the unit of the household cannot be obtained to better estimate the energy consumption for each household. Moreover, the applicability of the developed models and corresponding findings should be checked by other countries or regions

3. This thesis is only targeted for energy demand modeling with consideration of OBs in the residential sector. Its results are only applicable to buildings in similar conditions. In general, OB models for modeling occupancy and action are more likely to be applied in the commercial sector. Regarding the transport sector, electrical vehicle models are widely used. These models had their own modeling logic.

4. The developed spatial logistic regression model requires the spatial location data for each simulated occupant. The studies used low-resolution location data which makes the people in one state of the U.S. share the same location. Therefore, the results obtained by the new OB model did not show obvious progress compared to the conventional logistic regression model for several cases. Thus, sufficient higher resolution location data is needed for the model development. In addition, this thesis only focuses on the spatial variation in OBs. As mentioned in Chapter 5.1.2, the spatial and temporal variation should be considered simultaneously in the modeling to achieve a more realistic dynamic energy demand simulation.

The overall work of this thesis provides the starting point for investigating the development of the OB model underpinning the bottom –up based energy demand simulation and enhancing the diversity in OBs. In addition, this work is one of the first studies to develop new OB models that can robustly incorporate spatial variation. Hence, future work can be identified from the following point:

1. Integrate variables from a wider range of factors such as the time, environmental, and psychological factors to represent the variation shown by occupants. Design metrics to characterize some hard-to-measure factors for OB models. At the same time, comprehensively considering variables in various factors for developing the OB model to further tackle the trade-off between error and diversity when increasing the model complexity.

2. The results of the developed OB model should be easily explained. The reason for the spatial variation in OBs should be further discussed and investigated. Moreover, the new OB model should generate more realistic OB patterns in terms of different spatial and time scales.

3. Test new OB models on different sectors as well as different geographical regions to verify their applicability range. The modeling method for the residential sector in America should be extended to other application contexts.

4. Efficiently combine other sub-modules with the developed OB model in energy demand simulation to generate reliable dynamic energy demand profiles. Evaluate the performance of the energy demand model along with the newly developed OB model.

5. The design of the division of a region should be rethought when adequate location data is available. More location data contributes to simulating the spatial autocorrelation among these locations to represent the spatial trend. However, the amount of supplement data containing other factors such as the demographic factor for one location will be decreased as the number of location data increases. Therefore, the number and position of the location data to represent a certain area needs to meet the requirements of 1) enough to reflect the trend change of the space, and 2) the amount of other data in each location is sufficient.

6. Derive strategies or policies as well as adjust programs (i.e., demand response program) based on simulated energy demand profiles.

6 Conclusion

There is an increasing interest in reducing energy consumption as well as reducing the associated greenhouse gas emissions in every sector of all nations. Among all sectors, the residential sector is a substantial consumer of energy and therefore draws the focus of energy consumption reduction efforts. Due to the complex characteristics of energy consumption in the residential sector, comprehensive models are needed to assess the impact of adopting energy efficiency and renewable energy technologies suitable for residential applications (Swan & Ugursal, 2009). In particular, models to capture dynamic changes in the energy demand are urgently required.

Based on this context, researchers have recognized that occupant behavior (OB) has a significant impact on building energy consumption. As occupants influence energy consumption through their direct interaction with the building systems and devices and influence the indoor environment by their presence just in terms of sources of heat and carbon dioxide production (Naspi et al., 2018). Numerous OB models for capturing the occupancy, activities, and actions of building occupants have been developed for understanding, modeling, and analyzing OBs and be integrated into current building energy simulation tools to quantify the effect of OBs on building energy use. However, existing OB studies seldom paid attention to the model development process, especially for the pre-simulation process (i.e., the modeling methods to prepare the parameter and variables used in the model). In addition, modeling OB with spatial variation which underpins building energy simulation is rare in general.

This thesis seeks to analyze and evaluate the modeling process for the existing OB models thereby summarizing the experiences that support researchers to shape their decision-making for designing the energy demand model with consideration of OBs. Furthermore, this thesis proposes new OB models that can incorporate spatial variation into the modeling process thereby better enhancing the model diversity.

The overall target of this thesis is to better understand and design the OB model based on ATUS data. This thesis conducted three studies to address the research gap. The following are sets of research questions relevant to the research gap and corresponding answers.

1) Address research questions on which variables should be considered and what is the most appropriate parameter preparation method to improve the pre-simulation process of OB modeling.

   Answer: all three parameter preparation methods (i.e., multinomial log-linear regression (MLR), support vector machines (SVMs), and feedforward artificial neural network (ANN)) with the same variables inputs had similar performances which are evaluated by indicators representing three aspects: average performance, diversity, and individual specificity. However, we recommend the use of MLR when applying basic TUD. Regarding the variable, we found that only including more significant variables, especially from the activity and appliance-related factors, contribute to enhancing the diversity. Therefore, setting a reference group using basic variables and a simple method helps to assess the feasibility of the designed methodology framework is recommended.

2) Address research questions on whether spatial variation and historical change exist in OB, whether variables can represent spatial and temporal variations, and whether the conventional modeling method can reproduce the spatial variation.

   Answer: the historical change was confirmed to exist in watching television activity conducted by the targeted subpopulation of women with full-time jobs based on comprehensive descriptive analysis. The spatial distribution of the activity probability can be visualized through the ordinary kriging interpolation method. The interpolation results revealed that the probability of undertaking watching television had an obvious spatiotemporal variation over the mainland of the U.S. at a five year-period time slot. More importantly, the analysis implied that the conventional logistic regression model can only reproduce limited spatial variation. A new model is urgently required to better reproduce the detailed spatial variation.

3) Address research questions on when spatial variation exists in OB, how can spatial variation in OB be represented quantitatively, and how can spatial methods reproduce spatial variations in OB to develop a new method for OB modeling with consideration of spatial variation.

   Answer: developed three spatial logistic regression models involving one spatial factor variable and two spatial representations (kriging-based and SAR-based representations) successfully incorporating the spatial variation into the modeling process based on the case of four representative activities for six subpopulations groups of women. By comparing with the conventional logistic regression model in terms of reproducibility and comprehensive performance (i.e., error and diversity), the results showed that the developed spatial logistic regression model improves the reproducibility of the spatial variation in OBs, at the same time it is able to generate estimations with smaller errors and higher diversity at a higher spatial resolution level. Moreover, the finding revealed that basic segmentation is recommended when the data is sufficient. Segmentation can enhance the diversity among simulated occupants as it helps to develop targeted models to better grasp and describe the differences between research objectives in different scenarios. However, further segmentation to run the model is unnecessary especially when only limited data is available. In this case, diversity can be enhanced by using appropriate variables in the modeling process.

To conclude, the overall work of this thesis has contributed to developing advanced OB models incorporating spatial variation. The thesis also has shown that indicators can be drawn from various aspects, including error, diversity, and reproducibility to assess the model performance more comprehensively. To sum up, this thesis has broadened the knowledge of the pre-simulation and successfully enhanced the model diversity thus improving the understanding of the systems and identifying areas to support sustainable decision-making depending on the time-use of people in different regions. Going forward, our findings in this thesis can be extended to generate more reliable energy demand profiles thus guiding to improve energy efficiency, saving energy cost, and reducing greenhouse gas emissions for heterogeneous regions.

Reference

Abbasabadi, N., Ashayeri, M., Azari, R., Stephens, B., & Heidarinejad, M. (2019). An integrated data-driven framework for urban energy use modeling (UEUM). *Applied Energy*, *253*(February), 113550. https://doi.org/10.1016/j.apenergy.2019.113550

Aerts, D., Minnen, J., Glorieux, I., Wouters, I., & Descamps, F. (2014). A method for the identification and modelling of realistic domestic occupancy sequences for building energy demand simulations and peer comparison. *Building and Environment*, *75*, 67–78. https://doi.org/10.1016/j.buildenv.2014.01.021

Al-Mumin, A., Khattab, O., & Sridhar, G. (2003). Occupants' behavior and activity patterns influencing the energy consumption in the Kuwaiti residences. *Energy and Buildings*, *35*(6), 549–559. https://doi.org/10.1016/S0378-7788(02)00167-6

An, J., Yan, D., Hong, T., & Sun, K. (2017). A novel stochastic modeling method to simulate cooling loads in residential districts. *Applied Energy*, *206*(May), 134–149. https://doi.org/10.1016/j.apenergy.2017.08.038

Anderson, B. (2016). Laundry, energy and time: Insights from 20 years of time-use diary data in the United Kingdom. *Energy Research and Social Science*, *22*, 125–136. https://doi.org/10.1016/j.erss.2016.09.004

Anxo, D., Mencarini, L., Pailhé, A., Solaz, A., Tanturri, M. L., & Flood, L. (2011). Gender differences in time use over the life course in France, Italy, Sweden, and the US. *Feminist Economics*, *17*(3), 159–195. https://doi.org/10.1080/13545701.2011.582822

Arraiz, I., Drukker, D. M., Kelejian, H. H., & Prucha, I. R. (2010). A spatial cliff-ord-type model with heteroskedastic innovations: Small and large sample results. *Journal of Regional Science*, *50*(2), 592–614. https://doi.org/10.1111/j.1467-9787.2009.00618.x

Belessiotis, V., & Mathioulakis, E. (2002). Analytical approach of thermosyphon solar domestic hot water system performance. *Solar Energy*, *72*(4), 307–315. https://doi.org/10.1016/S0038-092X(02)00011-7

Berke, O. (1999). Estimation and prediction in the spatial linear model. *Water, Air, and Soil Pollution*, *110*(3–4), 215–237. https://doi.org/10.1023/a:1005035509922

Berke, O. (2001). Modified median polish kriging and its application to the Wolfcamp-Aquifer data. *Environmetrics*, *12*(8), 731–748. https://doi.org/10.1002/env.495

Bivand, R., Millo, G., & Piras, G. (2021). A review of software for spatial econometrics in r. *Mathematics*, *9*(11), 1–40. https://doi.org/10.3390/math9111276

Buttitta, G., Turner, W., & Finn, D. (2017). Clustering of Household Occupancy Profiles for Archetype Building Models. *Energy Procedia*, *111*(September 2016), 161–170. https://doi.org/10.1016/j.egypro.2017.03.018

Calis, G., Atalay, S. D., Kuru, M., & Mutlu, N. (2017). Forecasting occupancy for demand driven HVAC operations using time series analysis. *Journal of Asian Architecture and Building Engineering*, *16*(3), 655–660. https://doi.org/10.3130/jaabe.16.655

Candanedo, L. M., Feldheim, V., & Deramaix, D. (2017). Data driven prediction models of energy use of appliances in a low-energy house. *Energy and Buildings*, *140*, 81–97. https://doi.org/10.1016/j.enbuild.2017.01.083

Carlucci, S., Lobaccaro, G., Li, Y., Catto Lucchino, E., & Ramaci, R. (2016). The effect of spatial and temporal randomness of stochastically generated occupancy schedules on the energy performance of a multiresidential building. *Energy and Buildings*, *127*, 279–300. https://doi.org/10.1016/j.enbuild.2016.05.023

Chasco, C., García, I., & Vicéns, J. (2007). Modeling Spatial Variations in Household Disposable Income with Geographically Weighted Regression. *Munich Personal RePEc Archive*, *50*(June 2014), 31.

Chiou, Y. S., Carley, K. M., Davidson, C. I., & Johnson, M. P. (2011). A high spatial resolution residential energy model based on American Time Use Survey data and the bootstrap sampling method. *Energy and Buildings*, *43*(12), 3528–3538. https://doi.org/10.1016/j.enbuild.2011.09.020

Concato, J., Peduzzi, P., Holford, T. R., & Feinstein, A. R. (1995). Importance of events per independent variable in proportional hazards analysis I. Background, goals, and general strategy. *Journal of Clinical Epidemiology*, *48*(12), 1495–1501. https://doi.org/10.1016/0895-4356(95)00510-2

Corgnati, S. P., Fabrizio, E., Filippi, M., & Monetti, V. (2013). Reference buildings for cost optimal analysis: Method of definition and application. *Applied Energy*, *102*, 983–993. https://doi.org/10.1016/j.apenergy.2012.06.001

Cressie, N. (1988). Spatial prediction and ordinary kriging. *Mathematical Geology*, *20*(4), 405–421. https://doi.org/10.1007/BF00892986

De Lauretis, S., Ghersi, F., & Cayla, J. M. (2017). Energy consumption and activity patterns: An analysis extended to total time and energy use for French households. *Applied Energy*, *206*(September), 634–648. https://doi.org/10.1016/j.apenergy.2017.08.180

Degré, A., Tech, G. A., & Passage, S. S. (2015). Different methods for spatial interpolation of rainfall data for operational hydrology and hydrological modeling at watershed scale : a review PoPuPS | Different methods for spatial interpolation of rainfall data for ... *Base*, *17*(2013), 1–10. lake

Deng, Z., & Chen, Q. (2019). Simulating the impact of occupant behavior on energy use of HVAC systems by implementing a behavioral artificial neural network model. *Energy and Buildings*, *198*, 216–227. https://doi.org/10.1016/j.enbuild.2019.06.015

Diao, L., Sun, Y., Chen, Z., & Chen, J. (2017). Modeling energy consumption in residential buildings: A bottom-up analysis based on occupant behavior pattern clustering and stochastic simulation. *Energy and Buildings*, *147*, 47–66. https://doi.org/10.1016/j.enbuild.2017.04.072

Druckman, A., & Jackson, T. (2008). Household energy consumption in the UK: A highly geographically and socio-economically disaggregated model. *Energy Policy*, *36*(8), 3177–3192. https://doi.org/10.1016/j.enpol.2008.03.021

Dziedzic, J. W., Yan, D., Sun, H., & Novakovic, V. (2020). Building occupant transient agent-based model – Movement module. *Applied Energy*, *261*(7491), 114417. https://doi.org/10.1016/j.apenergy.2019.114417

Esteban Ortiz-Ospina, C. G., & Roser, M. (2020). Time Use. *Our World in Data*.

Fabi, V., Buso, T., Andersen, R. K., Corgnati, S. P., & Olesen, B. W. (2013). Robustness of building design with respect to energy related occupant behaviour. *Proceedings of BS 2013: 13th Conference of the International Building Performance Simulation Association*, 1999–2006.

Faisal.N, & A.Gaffar. (2012). Development of Pakistan ’ s New Area Weighted Rainfall Using Thiessen Polygon Method. *Pakistan Journal of Meteorology*, *9*(17), 107–116.

Feng, X., Yan, D., & Hong, T. (2015). Simulation of occupancy in buildings. *Energy and Buildings*, *87*, 348–359. https://doi.org/10.1016/j.enbuild.2014.11.067

Fischer, D., Härtl, A., & Wille-Haussmann, B. (2015). Model for electric load profiles with high time resolution for German households. *Energy and Buildings*, *92*, 170–179. https://doi.org/10.1016/j.enbuild.2015.01.058

Gentry, J., Commuri, S., & Jun, S. (2003). " Review of Literature on Gender in the Family. *Academy of Marketing Science Review*, *2003*(January 2003), 1.

Goh, A. T. C. (1994). Some civil engineering applications of neural networks. *Proceedings of the Institution of Civil Engineers: Structures and Buildings*, *104*(4), 463–469. https://doi.org/10.1680/istbu.1994.27204

Haldi, F., Calì, D., Andersen, R. K., Wesseling, M., & Müller, D. (2017). Modelling diversity in building occupant behaviour: a novel statistical approach. *Journal of Building Performance Simulation*, *10*(5–6), 527–544. https://doi.org/10.1080/19401493.2016.1269245

Halleck Vega, S., van Leeuwen, E., & van Twillert, N. (2022). Uptake of residential energy efficiency measures and renewable energy: Do spatial factors matter? *Energy Policy*, *160*(March 2021), 112659. https://doi.org/10.1016/j.enpol.2021.112659

Happle, G., Fonseca, J. A., & Schlueter, A. (2018). A review on occupant behavior in urban building energy models. *Energy and Buildings*, *174*, 276–292. https://doi.org/10.1016/j.enbuild.2018.06.030

Hayn, M., Bertsch, V., & Fichtner, W. (2014). Electricity load profiles in Europe: The importance of household segmentation. *Energy Research and Social Science*, *3*(C), 30–45. https://doi.org/10.1016/j.erss.2014.07.002

Heinrich, M., Ruellan, M., Oukhellou, L., Samé, A., & Lévy, J.-P. (2022). From energy behaviours to lifestyles: Contribution of behavioural archetypes to the description of energy consumption patterns in the residential sector. *Energy and Buildings*, *269*, 112249. https://doi.org/10.1016/j.enbuild.2022.112249

Hoes, P., Hensen, J. L. M., Loomans, M. G. L. C., de Vries, B., & Bourgeois, D. (2009). User behavior in whole building simulation. *Energy and Buildings*, *41*(3), 295–302. https://doi.org/10.1016/j.enbuild.2008.09.008

Ibrahim, A., Ali, H., Abuhendi, F., & Jaradat, S. (2020). Thermal seasonal variation and occupants' spatial behaviour in domestic spaces. *Building Research and Information*, *48*(4), 364–378. https://doi.org/10.1080/09613218.2019.1681928

Jeong, B., Kim, J., & de Dear, R. (2021). Creating household occupancy and energy behavioural profiles using national time use survey data. *Energy and Buildings*, *252*, 111440. https://doi.org/10.1016/j.enbuild.2021.111440

Jiawei Han, Micheline Kamber, J. P. (2014). Data mining: Data mining concepts and techniques. In *Proceedings - 2013 International Conference on Machine Intelligence Research and Advancement, ICMIRA 2013*. https://doi.org/10.1109/ICMIRA.2013.45

Jones, R. V., Fuertes, A., Gregori, E., & Giretti, A. (2017). Stochastic behavioural models of occupants' main bedroom window operation for UK residential buildings. *Building and Environment, 118,* 144–158. https://doi.org/10.1016/j.buildenv.2017.03.033

Kavgic, M., Mavrogianni, A., Mumovic, D., Summerfield, A., Stevanovic, Z., & Djurovic-

Petrovic, M. (2010). A review of bottom-up building stock models for energy consumption in the residential sector. *Building and Environment*, *45*(7), 1683–1697. https://doi.org/10.1016/j.buildenv.2010.01.021

Kelejian, H. H., & Prucha, I. R. (1998). A Generalized Spatial Two-Stage Least Squares Procedure for Estimating a Spatial Autoregressive Model with Autoregressive Disturbances. *Journal of Real Estate Finance and Economics*, *17*(1), 99–121. https://doi.org/10.1023/A:1007707430416

Kim, T. W., & Cha, S. H. (2019). Empirical validation of the spatial-choice modelling approach to user simulation. *Architectural Science Review*, *62*(4), 313–322. https://doi.org/10.1080/00038628.2019.1625299

Kleinebrahm, M., Torriti, J., McKenna, R., Ardone, A., & Fichtner, W. (2021). Using neural networks to model long-term dependencies in occupancy behavior. *Energy and Buildings*, *240*, 110879. https://doi.org/10.1016/j.enbuild.2021.110879

Li, M., & Tilahun, N. (2020). A comparative analysis of discretionary time allocation for social and non-social activities in the U.S. between 2003 and 2013. *Transportation*, *47*(2), 893–909. https://doi.org/10.1007/s11116-018-9924-1

Li, Y., Yamaguchi, Y., & Shimoda, Y. (2022). Impact of the pre-simulation process of occupant behaviour modelling for residential energy demand simulations. *Journal of Building Performance Simulation*, *15*(3), 287–306. https://doi.org/10.1080/19401493.2021.2022759

Liisberg, J., Møller, J. K., Bloem, H., Cipriano, J., Mor, G., & Madsen, H. (2016). Hidden Markov Models for indirect classification of occupant behaviour. *Sustainable Cities and Society*, *27,* 83–98. https://doi.org/10.1016/j.scs.2016.07.001

Lim, H., & Zhai, Z. J. (2017). Review on stochastic modeling methods for building stock energy prediction. *Building Simulation*, *10*(5), 607–624. https://doi.org/10.1007/s12273-017-0383-y

Lu, G. Y., & Wong, D. W. (2008). An adaptive inverse-distance weighting spatial interpolation technique. *Computers and Geosciences*, *34*(9), 1044–1055. https://doi.org/10.1016/j.cageo.2007.07.010

Marín-Restrepo, L., Trebilcock, M., & Gillott, M. (2020). Occupant action patterns regarding spatial and human factors in office environments. *Energy and Buildings*, *214*. https://doi.org/10.1016/j.enbuild.2020.109889

Martinaitis, V., Zavadskas, E. K., Motuziene, V., & Vilutiene, T. (2015). Importance of occupancy information when simulating energy demand of energy efficient house: A case study. *Energy and Buildings*, *101*, 64–75. https://doi.org/10.1016/j.enbuild.2015.04.031

Mastrucci, A., Pérez-López, P., Benetto, E., Leopold, U., & Blanc, I. (2017). Global sensitivity analysis as a support for the generation of simplified building stock energy models. *Energy and Buildings*, *149*, 368–383. https://doi.org/10.1016/j.enbuild.2017.05.022

McKenna, E., Higginson, S., Grunewald, P., & Darby, S. J. (2018). Simulating residential demand response: Improving socio-technical assumptions in activity-based models of energy demand. *Energy Efficiency*, *11*(7), 1583–1597. https://doi.org/10.1007/s12053-017-9525-4

Mcmillen, D. P. (1996). One hundred fifty years of land values in Chicago: A nonparametric approach. *Journal of Urban Economics*, *40*(1), 100–124. https://doi.org/10.1006/juec.1996.0025

McMillen, D. P., & McDonald, J. F. (1997). A nonparametric analysis of employment density in a polycentric city. *Journal of Regional Science*, *37*(4), 591–612. https://doi.org/10.1111/0022-4146.00071

Mohammadi, N., & Taylor, J. E. (2017). Urban energy flux: Spatiotemporal fluctuations of building energy consumption and human mobility-driven prediction. *Applied Energy*, *195*, 810–818. https://doi.org/10.1016/j.apenergy.2017.03.044

Monestiez, P., Courault, D., Allard, D., & Ruget, F. (2001). Spatial interpolation of air temperature using environmental context: Application to a crop model. *Environmental and Ecological Statistics*, *8*(4), 297–309. https://doi.org/10.1023/A:1012726317935

Moran, P. A. P. (1948). The Interpretation of Statistical Maps. *Journal of the Royal Statistical Society: Series B (Methodological)*, *10*(2), 243–251. https://doi.org/10.1111/j.2517-6161.1948.tb00012.x

Mosteiro-Romero, M., Fonseca, J. A., & Schlueter, A. (2017). Seasonal effects of input parameters in urban-scale building energy simulation. *Energy Procedia*, *122*, 433–438. https://doi.org/10.1016/j.egypro.2017.07.459

Murakami, D., Yoshida, T., Seya, H., Griffith, D. A., & Yamagata, Y. (2017). A Moran coefficient-based mixed effects approach to investigate spatially varying relationships. *Spatial Statistics*, *19*, 68–89. https://doi.org/10.1016/j.spasta.2016.12.001

Naspi, F., Arnesano, M., Stazi, F., D'orazio, M., & Revel, G. (2018). Measuring Occupants' Behaviour for Buildings' Dynamic Cosimulation. *Journal of Sensors*, *2018*, 1–17. https://doi.org/10.1155/2018/2756542

Naspi, F., Arnesano, M., Stazi, F., D'Orazio, M., & Revel, G. M. (2018). Measuring occupants' behaviour for buildings’ dynamic cosimulation. *Journal of Sensors*, *2018*. https://doi.org/10.1155/2018/2756542

Nassar, K., & Elnahas, M. (2007). Occupant dynamics: Towards a new design performance measure. *Architectural Science Review*, *50*(2), 100–105. https://doi.org/10.3763/asre.2007.5015

Nath, R. (2014). Trend Surface Analysis of Spatial Data. *Gondwana Geological Magazine*, *29*(1–2), 39–44.

Nguyen, B. V. D., Wang, T. H., & Peng, C. (2020). Integration of agent-based modelling of social-spatial processes in architectural parametric design. *Architectural Science Review*, *63*(2), 119–134. https://doi.org/10.1080/00038628.2019.1640107

O’Brien, W., & Gunay, H. B. (2015). Mitigating office performance uncertainty of occupant use of window blinds and lighting using robust design. *Building Simulation*, *8*(6), 621–636. https://doi.org/10.1007/s12273-015-0239-2

O’Brien, W., Gunay, H. B., Tahmasebi, F., & Mahdavi, A. (2017a). A preliminary study of representing the inter-occupant diversity in occupant modelling. *Journal of Building Performance Simulation*, *10*(5–6), 509–526. https://doi.org/10.1080/19401493.2016.1261943

O’Brien, W., Gunay, H. B., Tahmasebi, F., & Mahdavi, A. (2017b). A preliminary study of representing the inter-occupant diversity in occupant modelling. *Journal of Building Performance Simulation*, *10*(5–6), 509–526. https://doi.org/10.1080/19401493.2016.1261943

Okada, T., Yamaguchi, Y., & Shimoda, Y. (2020). Data Preparation to Address Heterogeneity in Time Use Data Based Activity Modelling. *Proceedings of Building Simulation 2019: 16th Conference of IBPSA*, *16*, 2356–2363. https://doi.org/10.26868/25222708.2019.211095

Oliver, M. A., & Webster, R. (2007). International journal of geographical information systems Kriging : a method of interpolation for geographical information systems. *Geographical*, *October 2011*, 37–41.

Osman, M., & Ouf, M. (2021). A comprehensive review of time use surveys in modelling occupant presence and behavior: Data, methods, and applications. *Building and Environment*, *196*, 107785. https://doi.org/https://doi.org/10.1016/j.buildenv.2021.107785

Paciorek, C. J. (2007). Computational techniques for spatial logistic regression with large data sets. *Computational Statistics and Data Analysis*, *51*(8), 3631–3653.

https://doi.org/10.1016/j.csda.2006.11.008

Paul, P., Pennell, M. L., & Lemeshow, S. (2013). Standardizing the power of the Hosmer-Lemeshow goodness of fit test in large data sets. *Statistics in Medicine*, *32*(1), 67–80. https://doi.org/10.1002/sim.5525

Piselli, C., & Pisello, A. L. (2019). Occupant behavior long-term continuous monitoring integrated to prediction models: Impact on office building energy performance. *Energy*, *176*, 667–681. https://doi.org/10.1016/j.energy.2019.04.005

Rafiee, A., Dias, E., & Koomen, E. (2019). Analysing the impact of spatial context on the heat consumption of individual households. *Renewable and Sustainable Energy Reviews*, *112*(May), 461–470. https://doi.org/10.1016/j.rser.2019.05.033

Rafiq, M. Y., Bugmann, G., & Easterbrook, D. J. (2001). Neural network design for engineering applications. *Computers and Structures*, *79*(17), 1541–1552. https://doi.org/10.1016/S0045-7949(01)00039-6

Ramírez-mendiola, J. L., Grünewald, P., & Eyre, N. (2019). Residential activity pattern modelling through stochastic chains of variable memory length. *Applied Energy*, *237*(July 2018), 417–430. https://doi.org/10.1016/j.apenergy.2019.01.019

Ramírez-Mendiola, J. L., Grünewald, P., & Eyre, N. (2019). Residential activity pattern modelling through stochastic chains of variable memory length. *Applied Energy*, *237*(January), 417–430. https://doi.org/10.1016/j.apenergy.2019.01.019

Ranstam, J., & Cook, J. A. (2018). LASSO regression. *British Journal of Surgery*, *105*(10), 1348. https://doi.org/10.1002/bjs.10895

Raudys, S. J., & Jain, A. K. (1991). Small Sample Size Effects in Statistical Pattern Recognition: Recommendations for Practitioners. In *IEEE Transactions on Pattern Analysis and Machine Intelligence* (Vol. 13, Issue 3, pp. 252–264). https://doi.org/10.1109/34.75512

Richardson, I., Thomson, M., & Infield, D. (2008). A high-resolution domestic building occupancy model for energy demand simulations. *Energy and Buildings*, *40*(8), 1560–1566. https://doi.org/10.1016/j.enbuild.2008.02.006

Ruan, Y., Cao, J., Feng, F., & Li, Z. (2017). The role of occupant behavior in low carbon oriented residential community planning: A case study in Qingdao. *Energy and Buildings*, *139*, 385–394. https://doi.org/10.1016/j.enbuild.2017.01.049

Sayer, L. C. (2005). Gender, time and inequality: Trends in women's and men's paid work, unpaid work and free time. *Social Forces*, *84*(1), 285–303.

https://doi.org/10.1353/sof.2005.0126

Shahzad, S., Calautit, J. K., Hughes, B. R., Satish, B. K., & Rijal, H. B. (2019). Patterns of thermal preference and Visual Thermal Landscaping model in the workplace. *Applied Energy*, *255*(July). https://doi.org/10.1016/j.apenergy.2019.113674

Sheela, K. G., & Deepa, S. N. (2014). Selection of number of hidden neurons in neural networks in renewable energy systems. *Journal of Scientific and Industrial Research*, *73*(10), 686–688.

Shorrock, L. D., & Dunster, J. E. (1997). The physically-based model BREHOMES and its use in deriving scenarios for the energy use and carbon dioxide emissions of the UK housing stock. *Energy Policy*, *25*(12), 1027–1037. https://doi.org/10.1016/S0301-4215(97)00130-4

Silva-Palacios, D., Ferri, C., & Ramírez-Quintana, M. J. (2017). Improving Performance of Multiclass Classification by Inducing Class Hierarchies. *Procedia Computer Science*, *108*, 1692–1701. https://doi.org/10.1016/j.procs.2017.05.218

Stazi, F., Naspi, F., & D'Orazio, M. (2017). A literature review on driving factors and contextual events influencing occupants' behaviours in buildings. *Building and Environment*, *118*, 40–66. https://doi.org/10.1016/j.buildenv.2017.03.021

Swan, L. G., & Ugursal, V. I. (2009). Modeling of end-use energy consumption in the residential sector: A review of modeling techniques. *Renewable and Sustainable Energy Reviews*, *13*(8), 1819–1835. https://doi.org/10.1016/j.rser.2008.09.033

Tabak, V. (2009). *User Simulation of Space Utilisation: System for Office Building Usage Simulation* (Issue 2009). https://doi.org/10.6100/IR640457

Tahmasebi, F., & Mahdavi, A. (2018). On the utility of occupants' behavioural diversity information for building performance simulation: An exploratory case study. *Energy and Buildings*, *176*, 380–389. https://doi.org/10.1016/j.enbuild.2018.07.042

Tanimoto, J., Hagishima, A., & Sagara, H. (2008a). A methodology for peak energy requirement considering actual variation of occupants' behavior schedules. *Building and Environment*, *43*(4), 610–619. https://doi.org/10.1016/j.buildenv.2006.06.034

Tanimoto, J., Hagishima, A., & Sagara, H. (2008b). Validation of probabilistic methodology for generating actual inhabitants' behavior schedules for accurate prediction of maximum energy requirements. *Energy and Buildings*, *40*(3), 316–322. https://doi.org/10.1016/j.enbuild.2007.02.032

Toftum, J. (2010). Central automatic control or distributed occupant control for better indoor environment quality in the future. *Building and Environment*, *45*(1), 23–28.

https://doi.org/10.1016/j.buildenv.2009.03.011

Torriti, J. (2012). Demand Side Management for the European Supergrid: Occupancy variances of European single-person households. *Energy Policy*, *44*, 199–206. https://doi.org/10.1016/j.enpol.2012.01.039

Torriti, J. (2017). Understanding the timing of energy demand through time use data: Time of the day dependence of social practices. *Energy Research and Social Science*, *25*, 37–47. https://doi.org/10.1016/j.erss.2016.12.004

Ueda, R., & Mita, A. (2015). Homeostasis lighting control based on relationship between lighting environment and human behavior. *Sensors and Smart Structures Technologies for Civil, Mechanical, and Aerospace Systems 2015*, *9435*(March 2015), 94352S. https://doi.org/10.1117/12.2083672

Varouchakis, E. A. (2021). Median polish kriging and sequential gaussian simulation for the spatial analysis of source rock data. *Journal of Marine Science and Engineering*, *9*(7). https://doi.org/10.3390/jmse9070717

Wang, C., Yan, D., & Jiang, Y. (2011). A novel approach for building occupancy simulation. *Building Simulation*, *4*(2), 149–167. https://doi.org/10.1007/s12273-011-0044-5

Wang, Y.-C. (2012). Examining landscape determinants of Opisthorchis viverrini transmission. *EcoHealth*, *9*(3), 328–341. https://doi.org/10.1007/s10393-012-0789-z

Widén, J., Nilsson, A. M., & Wäckelgård, E. (2009). A combined Markov-chain and bottom-up approach to modelling of domestic lighting demand. *Energy and Buildings*, *41*(10), 1001–1012. https://doi.org/10.1016/j.enbuild.2009.05.002

Wilke, U., Haldi, F., Scartezzini, J., & Robinson, D. (2013). A bottom-up stochastic model to predict building occupants ' time-dependent activities. *Building and Environment*, *60*, 254–264. https://doi.org/10.1016/j.buildenv.2012.10.021

Xie, C., Huang, B., & Claramunt, C. (2000). *Spatial logistic regression and GIS to model rural^ urban land conversion'' Estimation of ubiquitous air quality View project Maritime Big Data Workshop 2020 View project SEE PROFILE*. *March 2014*. https://www.researchgate.net/publication/228904456

Xie, X., Semanjski, I., Gautama, S., Tsiligianni, E., Deligiannis, N., Rajan, R. T., Pasveer, F., & Philips, W. (2017). A review of urban air pollution monitoring and exposure assessment methods. *ISPRS International Journal of Geo-Information*, *6*(12), 1–21. https://doi.org/10.3390/ijgi6120389

Xu, X., & Chen, C. fei. (2019). Energy efficiency and energy justice for U.S. low-income households: An analysis of multifaceted challenges and potential. *Energy Policy*, *128*(January), 763–774. https://doi.org/10.1016/j.enpol.2019.01.020

Yamaguchi, Y, Yilmaz, S., Prakash, N., Firth, S. K., & Shimoda, Y. (2019). A cross analysis of existing methods for modelling household appliance use. *Journal of Building Performance Simulation*, *12*(2), 160–179. https://doi.org/10.1080/19401493.2018.1497087

Yamaguchi, Yohei, & Shimoda, Y. (2017). A stochastic model to predict occupants' activities at home for community-/urban-scale energy demand modelling. *Journal of Building Performance Simulation*, *10*(5–6), 565–581. https://doi.org/10.1080/19401493.2017.1336255

Yan, D., Hong, T., Dong, B., Mahdavi, A., D'Oca, S., Gaetani, I., & Feng, X. (2017). IEA EBC Annex 66: Definition and simulation of occupant behavior in buildings. *Energy and Buildings*, *156*, 258–270. https://doi.org/10.1016/j.enbuild.2017.09.084

Yan, D., O'Brien, W., Hong, T., Feng, X., Burak Gunay, H., Tahmasebi, F., & Mahdavi, A. (2015). Occupant behavior modeling for building performance simulation: Current state and future challenges. *Energy and Buildings*, *107*, 264–278. https://doi.org/10.1016/j.enbuild.2015.08.032

Yang, Z., Li, N., Becerik-Gerber, B., & Orosz, M. (2012). A multi-sensor based occupancy estimation model for supporting demand driven HVAC operations. *Simulation Series*, *44*(8 BOOK), 100–107.

Yu, W., Li, B., Lei, Y., & Liu, M. (2011). Analysis of a residential building energy consumption demand model. *Energies*, *4*(3), 475–487. https://doi.org/10.3390/en4030475

Zhao, J., Lasternas, B., Lam, K. P., Yun, R., & Loftness, V. (2014). Occupant behavior and schedule modeling for building energy simulation through office appliance power consumption data mining. *Energy and Buildings*, *82*, 341–355. https://doi.org/10.1016/j.enbuild.2014.07.033

Zhou, M., Li, J., Basu, R., & Ferreira, J. (2022). Creating spatially-detailed heterogeneous synthetic populations for agent-based microsimulation. *Computers, Environment and Urban Systems*, *91*(September 2021), 101717. https://doi.org/10.1016/j.compenvurbsys.2021.101717

Zhu, G., Liu, J., Tan, Q., & Shi, B. (2016). Inferring the Spatio-temporal Patterns of Dengue Transmission from Surveillance Data in Guangzhou, China. *PLoS Neglected Tropical Diseases*, *10*(4). https://doi.org/10.1371/journal.pntd.0004633