Modeling of occupant behavior considering spatial variation: geostatistical analysis and application based on American time use survey data

Energy & Buildings 281 (2023) 112754

Contents lists available at ScienceDirect

Energy & Buildings

journal homepage: www.elsevier.com/locate/enb

Modeling of occupant behavior considering spatial variation: Geostatistical analysis and application based on American time use survey data

Yuanmeng Li [a,*], Yohei Yamaguchi [a,*], Jacopo Torriti [b], Yoshiyuki Shimoda [a]

[a] Graduate School of Engineering, Osaka University, Osaka, Japan
[b] School of Built Environment, University of Reading, RG6 6DF, UK

ARTICLE INFO

Article history:
Received 29 April 2022
Revised 10 November 2022
Accepted 27 December 2022
Available online 28 December 2022

Keywords:
Occupant behavior
Time use
Spatial variation
Spatial logistic regression model

ABSTRACT

Numerous occupant behavior (OB) models that simulate occupancy, activity and action at home have been developed to improve the accuracy and quality of energy demand estimations. Previous studies have revealed that the consideration of inter-occupant diversity improves the performance of OB models. However, existing models ignore spatial variation in OBs or partially consider it using a simple method without evaluating whether it is sufficient. Moreover, the modeling method to reproduce the spatial variation is missing. This study aims to develop a modeling method that can effectively reproduce spatial variation in OBs using American time use data. The global Moran's index test confirmed that spatial variations exist in OBs; however, they differ by time of day and activities for studied population. Subsequently, two spatial variation representations were generated using the ordinary kriging and spatial autoregressive methods. Finally, three spatial logistic regression models that consider spatial variations were developed and evaluated. The developed models generated smaller errors and higher inter-occupant diversity than the conventional logistic regression models at the state level. The established method is applicable to any country and region. Using higher spatial resolution and richer time use datasets may further improve OB models to model region-specific characteristics of building energy demand.

1. Introduction

Over the past few decades, numerous occupant behavior (OB) models have been established to simulate the occupancy, activity, and behavior of building residents, to improve the prediction accuracy of building energy demands. OB is a major source of uncertainty in building energy demand modeling because energy-consuming appliances are generally operated to meet people's daily needs in response to activities performed by occupants, and building energy systems and indoor environments are adjusted by occupants for comfort [1]. Various methods have been applied to time use data integrated with additional survey data that cover social, economic, and building aspects, to develop representative OB models [2]. However, a significant gap exists between simulation and reality [3] owing to (1) the use of oversimplified assumptions, such as a fixed schedule rather than a dynamic schedule; (2) assumptions on when and how residents use appliances and building systems; and (3) ignorance of inter-occupant diversity [4]. Although some studies have attempted to address the first two gaps, inter-occupant diversity, particularly in terms of spatial variation, has not been thoroughly investigated [5,6]. Druckman and Jackson [7] demonstrated that household energy use and the associated carbon emissions vary significantly with household socioeconomic conditions and locations. Rural/urban environments are another important factor in devising policies for a low-carbon society. Vega et al. [8] 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 spatial factors are important, and ignoring them 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 [9–11]. Esteban et al. [12] found that

Abbreviations: ATUS, American Time Use and Leisure Activity Survey; Moran's I test, Moran's index test; MSD, Mean standard deviation; OB, Occupant behavior; RMSE, Root mean squared error; SAR, Spatial autoregressive; TAE, Total absolute error.

* Corresponding authors.
E-mail addresses: Lym19940224@yahoo.co.jp (Y. Li), yohei@see.eng.osaka-u.ac.jp (Y. Yamaguchi).

https://doi.org/10.1016/j.enbuild.2022.112754
0378-7788/

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.

This study proposes and evaluates various methods learned from other fields for modeling OB considering 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? (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 paper presents the methodology, results, and discussion, followed by our conclusions.

2. Literature review

2.1. Review of methods for considering spatial variation in OB and energy modeling

Spatial variation essentially refers to the rules or tendencies of objects of the research exhibited in a given space. Spatial variation can be represented and considered in the modeling in different ways. There is a significant development in OB modeling that addresses space use. These space use studies considered spatial choice or individual preference based on geo-referenced data to determine space use [13,14]. Tabak [15] 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 [16–21]. However, the variation of OBs over space has not been considered in these studies.

Some studies have used spatial factors as independent variables to consider spatial variation during the modeling process to enhance the inter-occupant diversity of the model [22]. Vega et al. [8] assessed various factors, including seven spatial factors (e.g., urban–rural gradient, city center, and village center), to develop a suitable policy for increasing the uptake of carbon emission reduction measures, and highlighted the importance of using spatial factors for designing energy policy frameworks. Marín-Restrepo et al. [23] 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 factors. Wilke et al. [24] 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. [25] applied the same method by considering city size as an independent variable to simulate the probability of undertaking activities. Rafiee et al. [26] 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. [27] 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).

Therefore, less focus has been paid to spatial variation in the OB modeling. Spatial variation has been insufficiently represented based on the actual data in previous studies. Although some studies used spatial factors, there is a lack of modeling methods to better reproduce spatial variation in OB.

2.2. Review of methods for spatial analysis and modeling

Disciplines associated with the fields of epidemiology, environmental meteorology, and econometrics have applied sound spatial analysis methods to solve subject-specific problems [28–35]. This section summarized such methods used to either empirically represent the spatial variation or simulate the research object with the consideration of the spatial variation. Fig. 1 shows the summary of the methods.

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 objects or to generate spatial predictions for unmeasured areas. Berke [36] applied the trend surface analysis and universal kriging to simulate acid-precipitation in Lower Saxony. Berke [37] also developed the modified median polish kriging method to generate more robust spatial predictions for Wolfcamp-Aquifer. Varouchakis [38] 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 variables, into the modeling process. Chasco et al. [35] analyzed the spatially varying impacts of some conventional factors, such as unemployment rate and average housing price, on the per capita household income in Spanish provinces based on geographically weighted regression. Xie et al. [39] 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 [40] 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.

These spatial analysis methods may be useful in OB modeling, however have been sparsely considered and applied in the energy field.

3. Methodology

3.1. Data

The multiyear American Time Use and Leisure Activity Survey (ATUS0319) collected the activity diaries and sociodemographic conditions of the survey participants. These activity diaries were recorded for a 24-hour period beginning at 4:00 am on the survey day. The data collected between 2009 and 2019 were used to ensure the consistent coding of the variables. Although there were 124,941 participants in total, we used those of women aged 30–59 living in mainland U.S. to homogenize the sample and better observe the effects of spatial variation in time use. This subpopulation was selected because women generally conduct various activities involving both paid and unpaid work [41–44]. In particular, these unpaid activities (e.g., housework) may affect the operating conditions of many home appliances and building systems, thereby affecting residential energy demand. We checked statistically that this subpopulation features the highest level of unpaid work in the ATUS. In addition, this is supported by other empirical research on

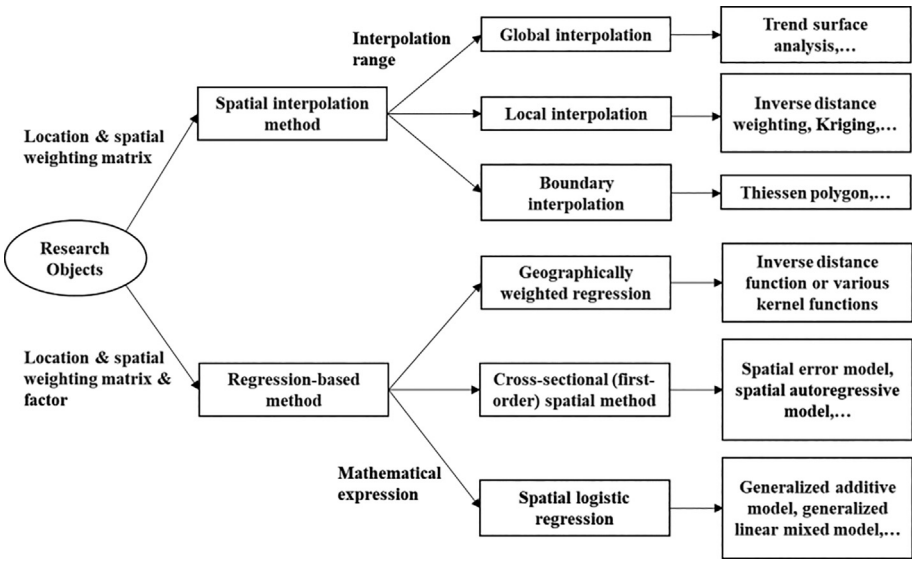

Fig. 1. Summary of the methods for spatial analysis and modeling.

time use showing that women trigger residential peak electricity demand due to caring activities and unpaid work [45]. We also selected four typical activities—sleeping, cooking and washing up, watching television, and commuting—to evaluate the proposed method. Sleeping is a basic in-home activity while commuting to work or school is one of the main out-of-home activities. These two activities had little influence on the use of home appliances. Watching television, cooking, and washing up are among the main indoor activities for women [46], and generally involve using appliances. Although the results are not sufficiently comprehensive to evaluate the applicability and usefulness of the proposed method to model the entire population and all activities, this design is sufficient to address the research questions described in the Introduction.

As a result of the selection of the subpopulation, the sample size was reduced to 36,438. However, this sample size is relatively large compared to many previous studies because we used the eleven-year data, whereas single-year time use data have often been used in previous studies [47]. Notably, the ATUS sample is distributed approximately proportionally to each state's population, with the number of samples varying considerably from state to state, ranging from 55 to 3652. In addition, 70 % of each state's data were randomly selected as the training dataset and the remaining were used as the test dataset. The split between percentages for training dataset and test dataset is in line with modelling practices associated with models requiring data training. Also, many previous studies used this split [22,48,49].

We considered the states as the unit for modeling as it was the only available data with respect to space for the entire nation. The location of each occupant was defined by the internal point of the state in which the occupant lived. Therefore, only one location point was used to represent the entire state to smooth the spatial variations for the entire U.S. mainland. The cartographic boundary shapefile of the U.S. of 2018 was used to visualize the spatial distribution of the probability on the map.

The spatial distribution of the activity probability at each time interval is referred to as the spatial probability in this study. Note that the 1 min resolution data in the time use diary were converted to 1 hourly binary data by assigning 1 when an activity was conducted within a 1-hour interval distinguished by clock times and 0 otherwise. Based on this principle, we quantified the probability of activity frequency within an hour. In this paper, we refer to this probability as the "activity probability".

3.2. Method

When simulating OB, numerous stochastic models use several modeling parameters (e.g., probability of undertaking an activity, probability of starting an activity and corresponding duration) [22]. These modeling parameters were prepared during the presimulation process. Li et al. [22] revealed that many previous studies conducted segmentation of sample time use data and applied the logistic regession method to model the modeling parameters to better enhance the inter-occupant diversity originating from demographic and other influencing factors. Our developed modeling method followed this approach but involved a smooth function that representative of the spatial variation in the modelling parameters. The whole methodology of this study is shown in Fig. 2. Steps 1–3 address the three research questions discussed in Section 1. Section 3.2.1 describes the segmentation of the time use data. The following Sections 3.2.2 to 3.2.4 give a more detailed introduction to each step.

3.2.1. Segmentation

In the presimulation process, six groups were designed to represent different subpopulations of women. Each group was homogenized to avoid the influence of sociodemographic factors in the spatial variation as shown in Table 1. The conditions for segmentation were the type of day (i.e., weekdays and weekends) and employment status—commonly used parameters in previous studies [19,20,23,24,50]. Groups 1 and 4 represent women with full-time jobs; Groups 2 and 5 represent women with part-time jobs; Groups 3 and 6 represent unemployed women. Groups 1–3 and Groups 4–6 comprise activities performed during weekdays and weekends, respectively.

Table 1 presents the total sample size for each group. The national-level sample size for each case satisfies the Whitemore formula [51] for most of the time intervals. However, some states did not have a large sample sizes, as shown in Fig. 3. Some states, such as Delaware, District of Columbia, and Wyoming (numbers 10, 11, and 56) had small sample sizes.

One approach to avoid a decrease in sample size is to use the group conditions as variables. To evaluate this approach, Group 7 was considered to represent the entire population of women aged 30–59 years, including Groups 1–6, using dummy variables representing each group in the developed spatial logistic regression model. The comparison enables the determination of a superior

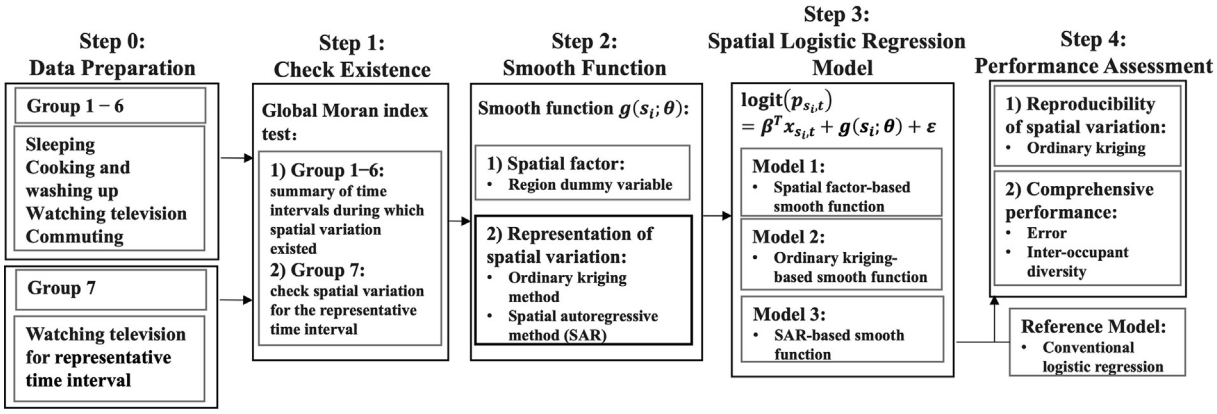

Fig. 2. Study methodology.

Table 1
Groups and their details.

Group	Subpopulation	Type of day	Employment status	Items of interest	Sample size
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	8849
2			Part-time		3551
3			Unemployed		5724
4		Weekends	Full-time		9041
5			Part-time		3548
6			Unemployed		5725
7	Entire population of women aged 30–59			Items in Groups 1–6, as well as employment status and type of day	36,438

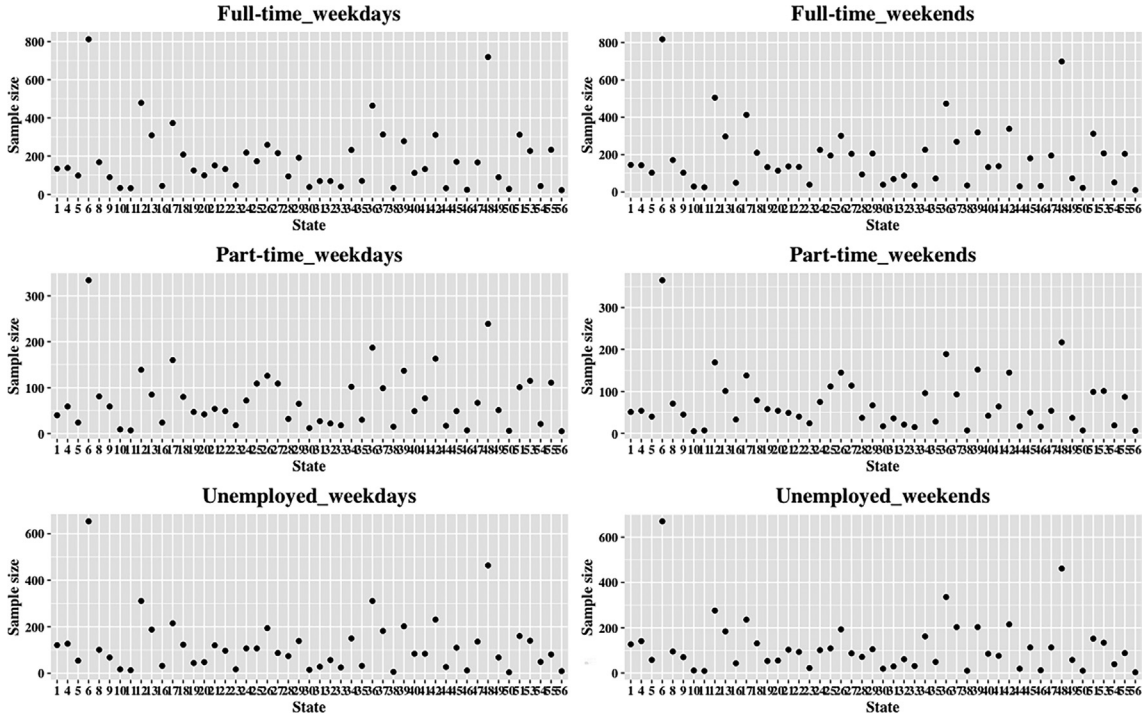

Fig. 3. Sample size of each group for each state.

approach for OB modeling, using the segmentation [52] or using grouping conditions as variables. This analysis was conducted by considering the watching television activity.

3.2.2. Step 1: Existence of spatial variation

We employed the global Moran's index (Moran's I) test to confirm the time intervals during which the selected four activities exhibited spatial variation. The Moran's I test is used to verify the significance of the random distribution of qualitative determination in the areas of a map [53]. The Z score was calculated to evaluate the significance of Moran's I. If the Z score is not statistically significant ($p > 0.05$), 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.

3.2.3. Step 2: Methods to represent spatial variation

Two representations of spatial variation that quantify the average probability of 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 an activity, their values were restricted within 0 and 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 [54]. Considering a certain time interval during which spatial variation exists, the prediction G_{s_0} for the location $s_0(u_0, v_0)$ is given by:

$$\tilde{G}_{s_0(u_0,v_0)} = \sum_{j=1}^{N} \lambda_j G_{s_j(u_j,v_j)}, \tag{1}$$

where $G_{s_j(u_j,v_j)}$ is the average probability of an activity in the state s_j represented by the internal points (u_j, v_j); and λ_j is the unknown weight subjected to $\sum_{j=1}^{N} \lambda_j = 1$, for obtaining an unbiased estimation of G_{s_0}. We considered the commonly used theoretical semivariogram–spherical model to estimate λ.

B) Spatial autoregressive method.

The SAR method is used to examine the impact of the probability of an activity in one state on the neighboring states by including other factors in the modeling process. It is generated based on the cross-sectional spatial model defined by Equation (2):

$$\tilde{G}_{s_0(u_0,v_0)} = \mathrm{y}_{s_0} = \beta^T x + \lambda^T \mathrm{W}\mathrm{y}_{s_0} + \varepsilon, \tag{2}$$

where $\tilde{G}_{s_0(u_0,v_0)}$ is the average probability of an activity in the state s_o; x represents the variables; W is the weighting matrix constructed in the form of adjacent edges or points corresponding to each state; and λ is a scalar autoregressive parameter. The variable $\mathrm{W}\mathrm{y}_{s_0}$ is the spatial lag of y_{s_0}.

3.2.4. Step 3: Spatial logistic regression

In this study, we developed three spatial logistic regression models through Equation (3):

$$\mathrm{logit}\left(p_{s_i,t}\right) = \ln \frac{p_{s_i,t}}{1 - p_{s_i,t}} = \beta^T x_{s_i,t} + g(s_i;\theta) + \varepsilon, \tag{3}$$

where $p_{s_i,t}$ is the probability of the i th individual at a location s at a time interval t; β is the coefficient of the variable $\mathrm{x}_{s_i,t}$; and $g(s_i;\theta)$ is a smooth function parameterized by θ over the location s_i. One spatial factor $\tilde{G}_r$ and two representations of the spatial variation $\tilde{G}_s$, explained below, were used as $g(s_i;\theta)$ herein. The conventional logistic regression model ignoring $g(s_i;\theta)$ served as the reference model for comparison. In Model 1, $\tilde{G}_r$ is modeled as $\tilde{G}_{r_{i,t}} = \gamma_1 R_{1,i,t} + \gamma_2 R_{2,i,t} + \gamma_3 R_{3,i,t}$ for i th individual, where R_1, R_2, and R_3 indicate the northeast, mid-west, and west, respectively, with the southern region being the reference group; γ is the corresponding coefficient for each regional dummy variable. In Models 2 and 3, the estimations of $\tilde{G}_{s_{i,t}}$ was extracted from the ordinary kriging and SAR results in Step 2 to represent the spatial variation.

Stepwise analysis was applied to all the models to statistically test the significance of the considered variables, including the spatial factors and representations.

3.3. Performance assessment

The performance of the models was assessed in terms of the reproducibility of the spatial variations in OB 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 inter-occupant diversity considering the training and test datasets.

3.3.1. Error indicators

Total absolute error (TAE) and root mean squared error (RMSE) were considered to measure the error between the estimations obtained from the models and the observations. These two indicators were quantified at national and state levels. Previous studies only considered the national level, which measures the error for each time interval. At the state level, the errors were quantified based on the combinations of the time interval and state, thereby introducing an error because of spatial variations.

3.3.2. Inter-occupant diversity indicators

Inter-occupant diversity indicators assess the ability of the model to represent the total variations of OB among the simulated occupants. The indicator RMSE_GA [22] based on the Hosmer–Lemeshow test [55], which measures the root mean squared error between the averaged estimated probability and averaged observed probability of different subdivisions, is provided by Equation (4):

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

where d denotes the subdivision ($D = 10$). RMSE_GA was only quantified at the national level owing to data limitations. To compare the inter-occupant diversity at the state level, another indicator—the mean standard deviation (MSD), was used to measure the deviation of each estimation from the mean at the national and state levels.

4. Results

4.1. Confirmation of the existence of spatial variation

Fig. 4 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. As shown in Fig. 4, the probability of activities exhibited certain variation among the states at different times of the day. Such variation results are obtained from the combination of the differences in sociodemographic variables, the spatial variation, and sampling error [56], according to Equation (3). The effect of the first element was reduced by segmentation.

Spatial variations for each time interval for all combinations of groups and activities were confirmed using Moran's I test. Fig. 5 summarizes the results of the Moran's I tests. The results showed that spatial variation existed only during limited time intervals and varied with the type of day (weekdays or weekends), subpopulations with different employment statuses, and activities. For example, spatial variation in sleep existed at different time inter-

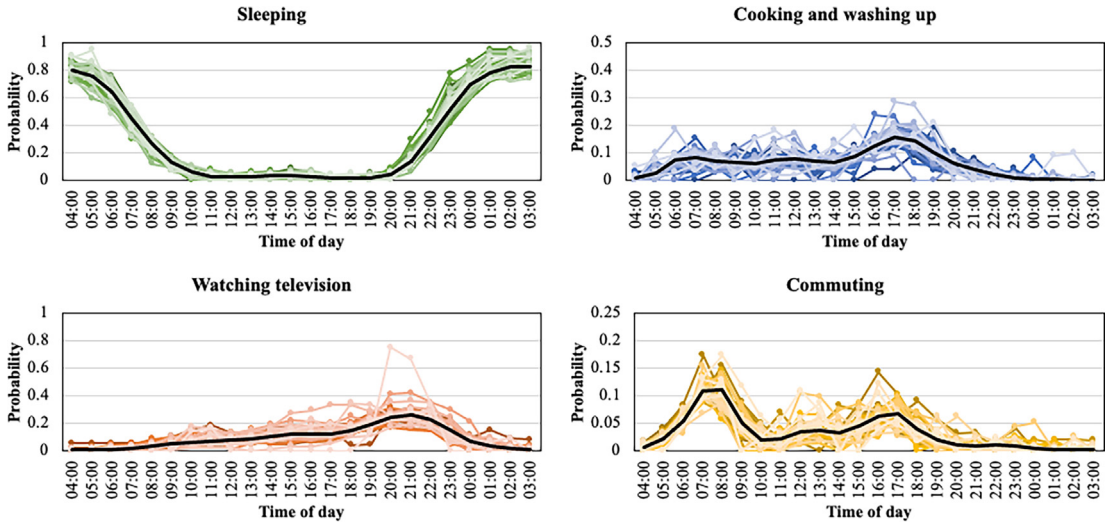

Fig. 4. Probability of activities considering representative groups. The differently colored lines represent the different states and the black line represents the national estimate.

vals for different groups because these groups may have different sleeping styles, containing various sub-activities, such as lying awake and napping, in the ATUS dataset. As shown in Fig. 5, considering sleeping, on weekdays, employed women in Group 1 exhibited lesser spatial variation than unemployed women in Group 3 during the relevant time intervals. 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 for 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

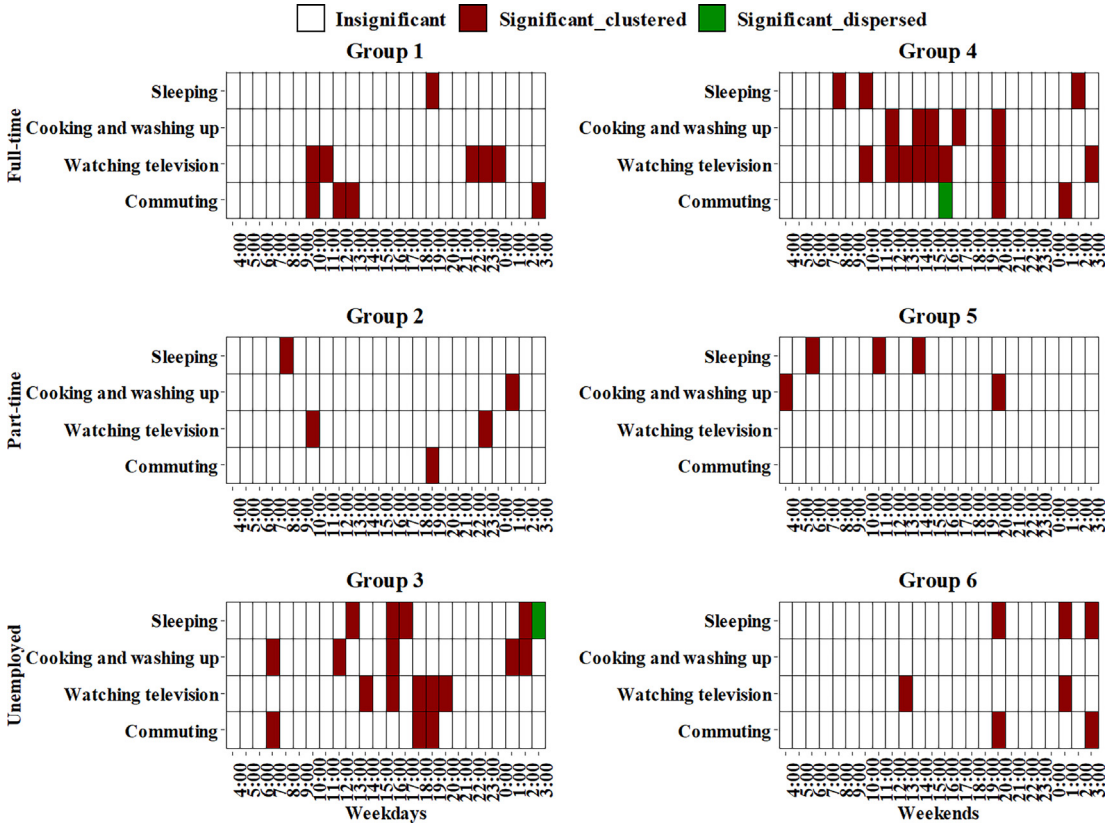

Fig. 5. Results of the Moran's I test considering the representative activities for each group. The time intervals filled in red or green are the intervals with spatial variations; red and green cells indicate clustered and dispersed distribution, respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

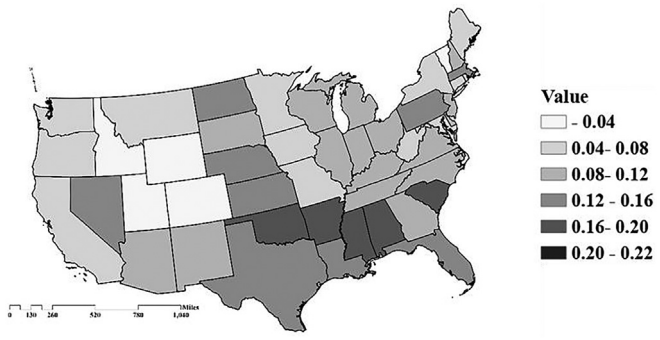

Fig. 6. Spatial probability of the women in Group 6 watching television at 13:00 at the state level based on observations.

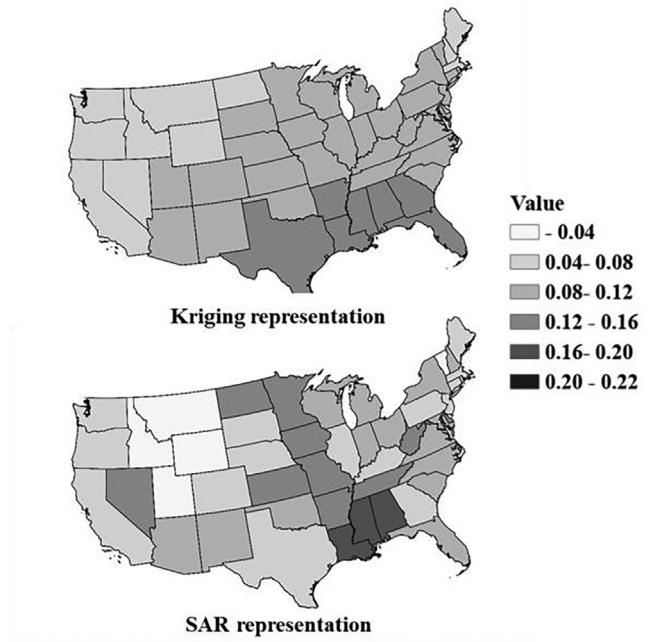

Fig. 7. Spatial probability of the women in Group 6 watching television at 13:00 based on representations of the spatial variation generated by the ordinary kriging and SAR methods.

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. Fig. 6 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 to 21 %.

4.2. Representations of spatial variation

Fig. 7 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 kriging-based representation ranges from 6 to 14 %, whereas SAR-based representation ranges from 4 to 17 %. The variation was narrower than the observation shown in Fig. 6. 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. Furthermore, we also compared the two representations considering all the combinations of groups, activities, and states. 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.

To better understand the cause of the error, Fig. 8 show the observed probability, a 95 % confidence interval, and the estimated probabilities of each state. Note that the observations of some states contain large sampling errors because of their small sample size. Some states had a probability of 0 because activity occurrence was not observed, which could also be attributed to the small sample size. As the error indicators were quantified based on the difference from the observed probabilities, they were at the scale described above. However, as shown in the figure, the two representations are within the 95 % confidence intervals of most states.

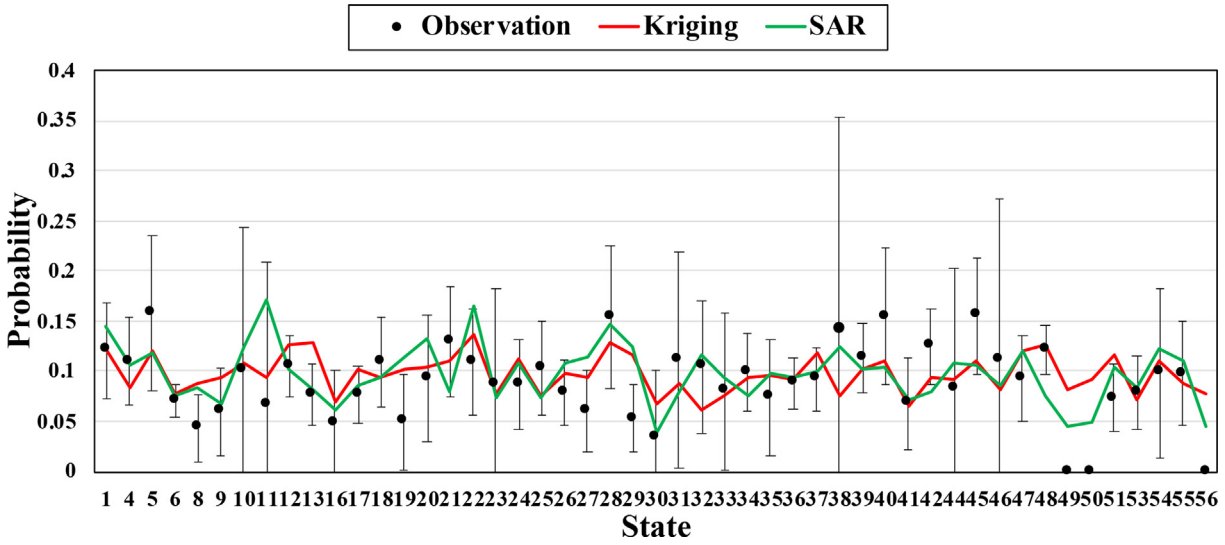

Fig. 8. Probability of activity of the women in Group 6 watching television at 13:00 based on representations and observations. Error bars indicate the 95% confident intervals of the observations.

4.3. Spatial logistic regression models

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; and (d) commuting at 10:00 in Group 1. These representative cases were selected among the time intervals with a spatial variation to compare the model performance. The four representative cases were selected based on their high probability compared to other intervals. Fig. 9 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 Fig. 9 for Cases (b) and (c)), which is attributed to the variations in sociodemographic variables.

As shown in Fig. 9, neither the reference model nor the spatial logistic regression models were not adequately consistent with the observations. To understand the reason, Fig. 10 shows the probabilities of each state. As shown in the figure, most of the estimated probabilities shown by the lines fall within the 95 % confidence intervals of the observations. The observation of states with a small sample size had either larger error bars or no error bars (probability = 0). Thus, the estimations were different from the observations for these states. However, spatial logistic

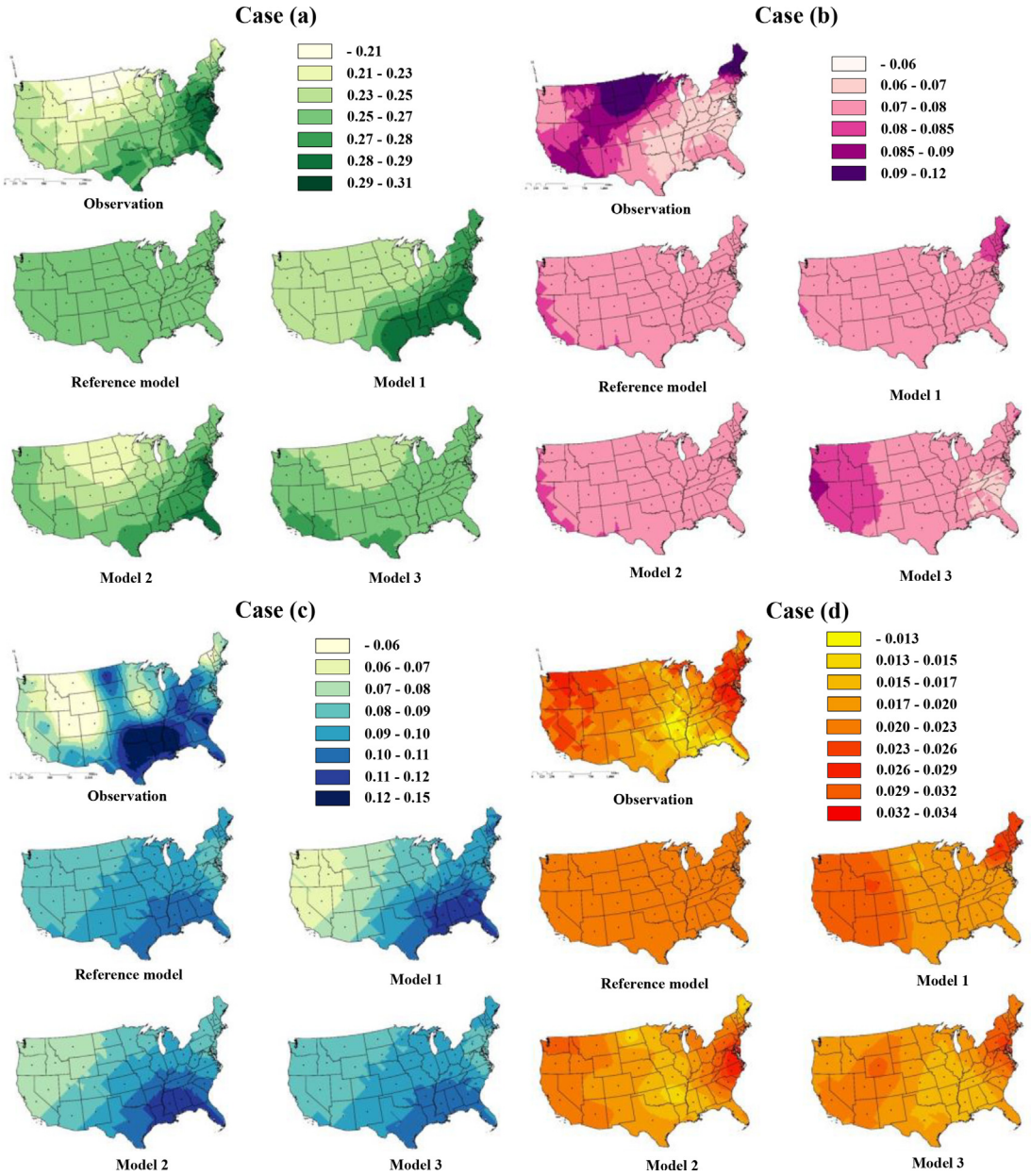

Fig. 9. Comparison of the spatial probability of activities based on the observations and reproductions of the spatial variation by the reference model and the three spatial logistic regression models for Cases (a)–(d), respectively. The spatial distribution results were interpolated by the ordinary kriging method.

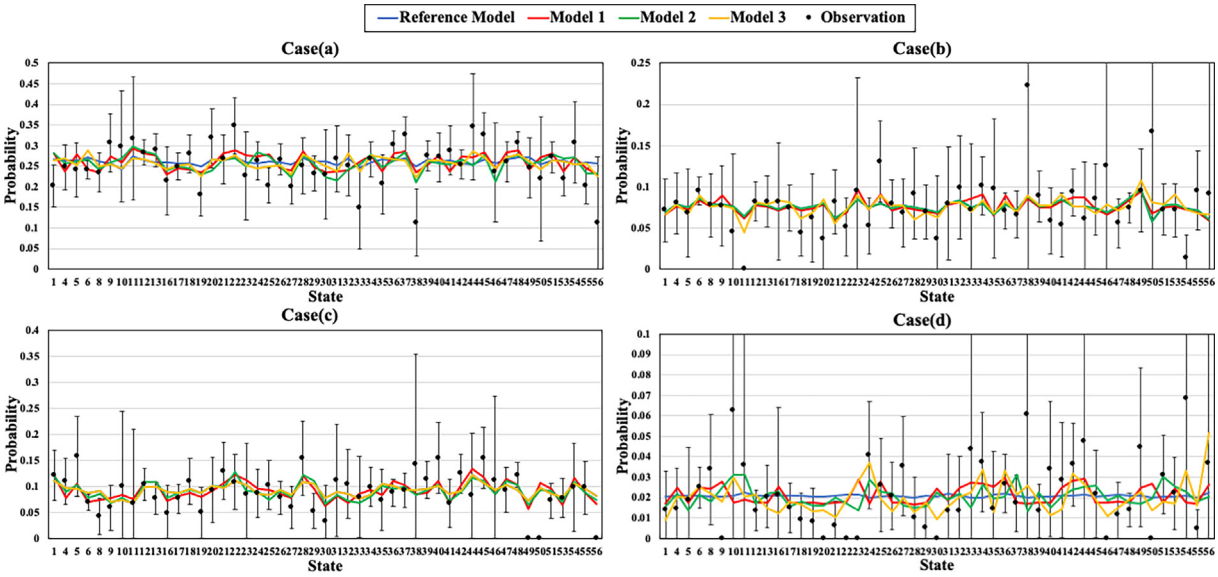

Fig. 10. Probability of activity of each state for cases used in Section 4.3.1. Error bar was quantified by the 95% confident intervals.

models are still more consistent with observations than reference models.

4.3.2. Comprehensive performance

Fig. 11 shows the stacked values of performance indicators quantified at the national level, TAE_{nation}, $RMSE_{nation}$, MSD_{nation}, and $RMSE\_GA_{nation}$, for all the models considering the six groups in 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 the combinations of the training and test datasets. Considering the inter-occupant diversity, Models 3 and 1 exhibited a 7 % higher MSD_{nation} than the reference model. Considering $RMSE\_GA_{nation}$, all the models exhibited similar results with both the training and test datasets.

Fig. 12 illustrates the TAE, RMSE, and MSD values of the models for the six groups quantified at the state level. As shown in Fig. 11 and Fig. 12, 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 TAE_{state} value by 9.9 and the stacked $RMSE_{state}$ value by 11 % for the training dataset. This was followed by Model 1 (stacked TAE_{state} decreased by 4.4 and stacked $RMSE_{state}$ decreased by 3.6 %) and Model 2 (stacked TAE_{state} decreased by 3.2 and stacked $RMSE_{state}$ decreased by 2.1 %). However, the spatial logistic regression models, particularly Model 3, did not provide such advantages with the test dataset. 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.

The above results are confirmed in Fig. 13, 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.

4.4. Evaluation of spatial logistic regression models applied to the entire population

4.4.1. Application of Group 7

In this section, the spatial logistic regression model was applied to Group 7 the entire population of women aged between 30 and 59 years, for watching television. 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.

Fig. 14 shows the same visualization maps of the spatial probability of watching television at 13:00 (Fig. 9) based on the observations and estimations of Group 7. The range of probability is narrower than Fig. 9 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 2 shows the performance of all 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 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.

4.4.2. Comparison approach of segmentation and using grouping conditions as variables

Fig. 15 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 Group 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 Fig. 15 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 con-

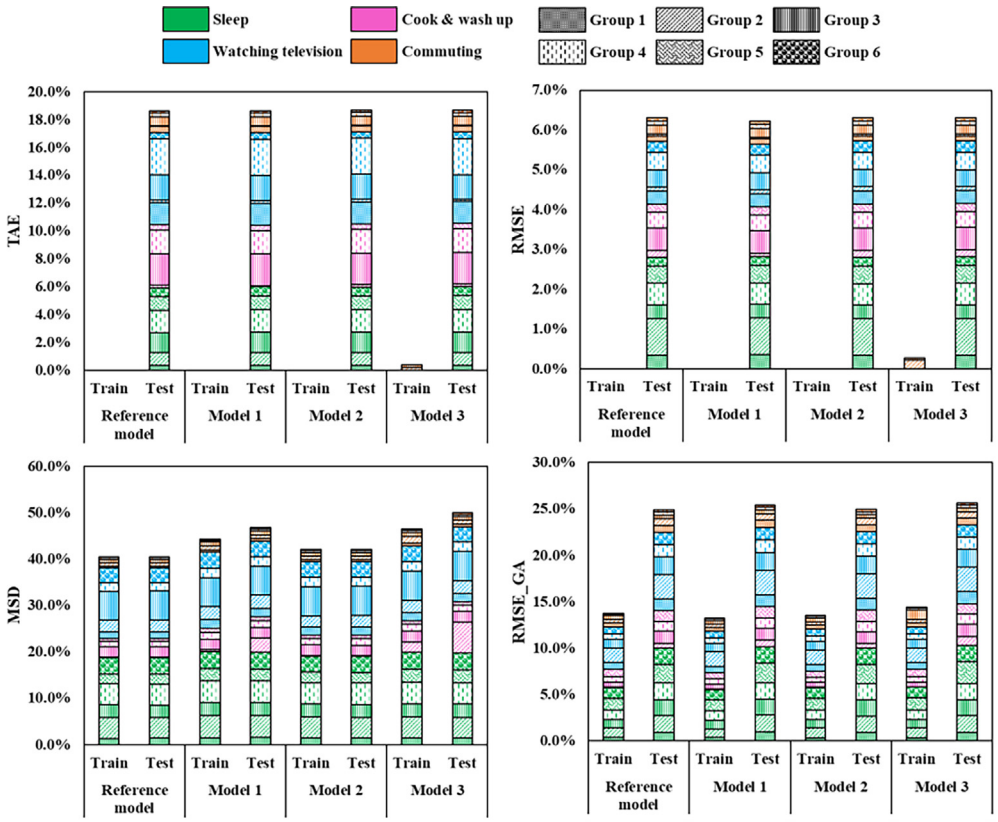

Fig. 11. Results of indicators at the national level for all the models in the training and test datasets.

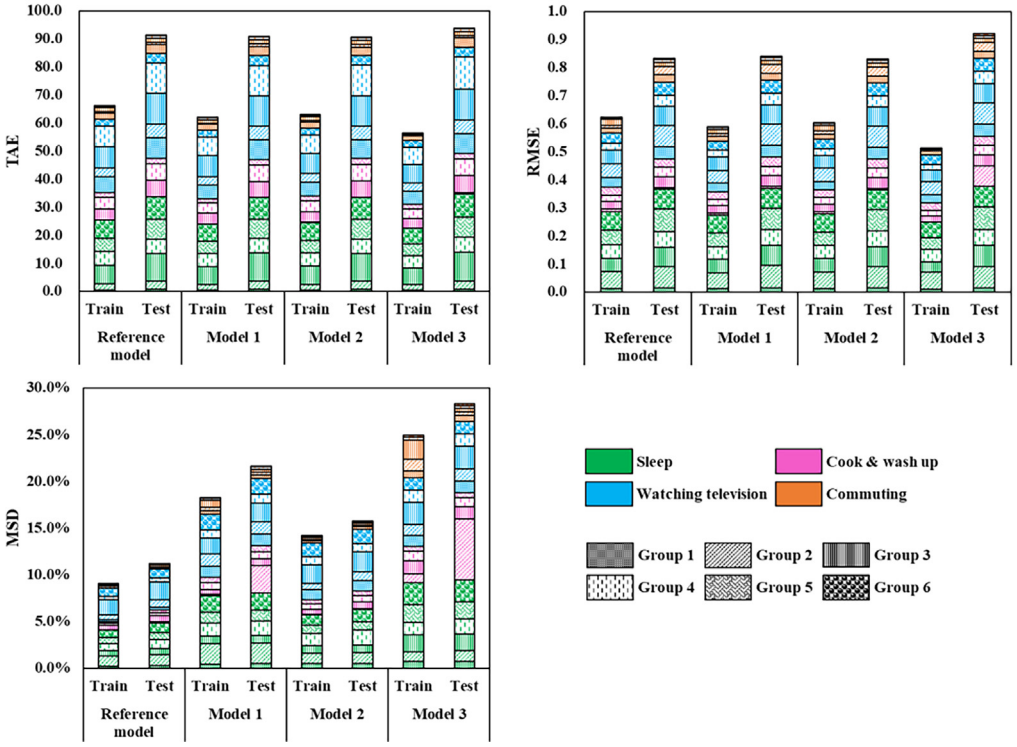

Fig. 12. Results of indicators at the state levels considering all the models in the training and test dataset.

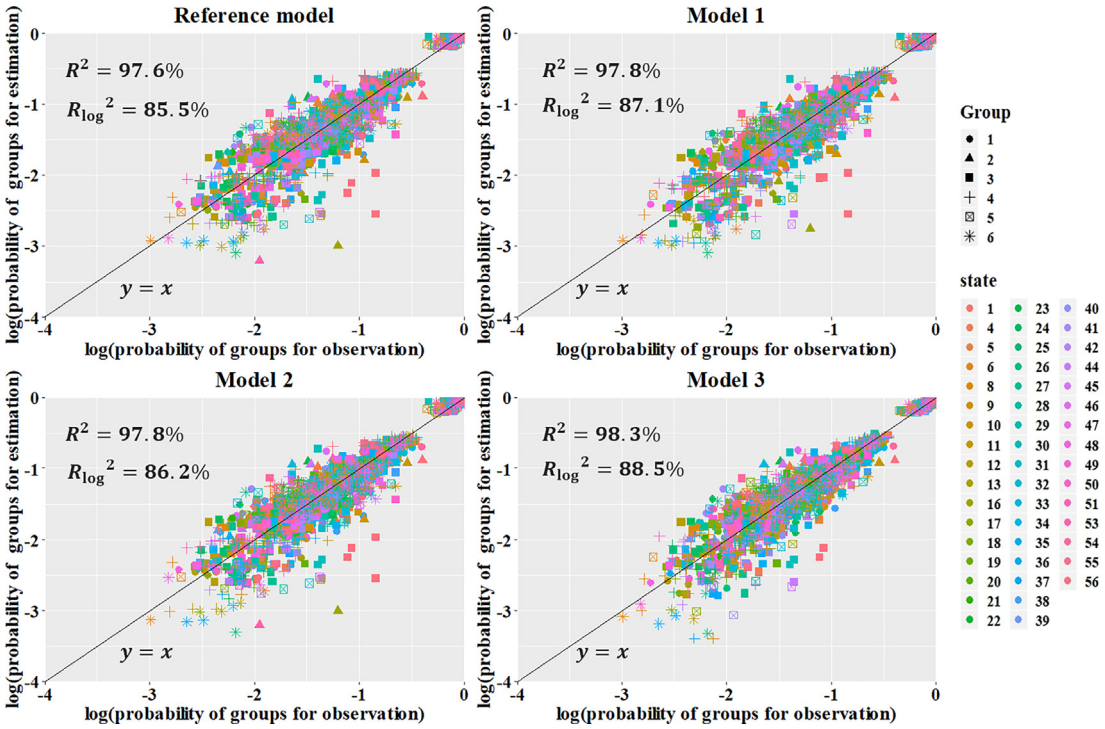

Fig. 13. Accuracy of the spatial logistic regression models at the state level. The horizontal axis shows the observation probabilities of the different combinations of the groups, states, and activities. The vertical axis shows the estimations. The black line is the reference line $y = x$. Logarithmic transformation was performed in the range $(-4, 0) \times (-4, 0)$.

sidering 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 relatively more accurate than those obtained from the variable-based approach.

Table 3 shows the comprehensive performance comparison through the statistical indicators of the two approaches for all models at the state level. According to Table 3, 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 inter-occupant diversity assessed by MSD, the variable-based approach showed a relatively better performance.

5. Discussion

5.1. Discussion of the results

This study demonstrated the existence of spatial variations in OBs and established a modeling method to consider these spatial variations in OBs. The established method is an extension of the existing modeling method (i.e., the logistic regression method combined with time use data sample segmentation in the pre-simulation process). We confirmed with women aged 30–59 in the U.S. for the four representative activities that the method contributes to better reproduction of spatial variation and enhancement of inter-occupant diversity in OB modeling.

The Moran's I tests in Section 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 Section 4.2 showed the SAR-based representation to be superior to the kriging-based representation, because the former accounts for the variation in other sociodemographic factors. Note that the representations deviated from the observations of each state, as shown in Fig. 8, particularly for those with small sample size. However, as the representations were within the 95 % confident intervals of the observations, using two representations contributed to avoiding the inclusion of the effect of sampling error in the following logistic regression modeling. If the location data required to develop a spatial representation is insufficient, spatial factors can be used to represent spatial variation for model development, as in the case of Model 1.

As discussed in Section 4.3, the developed spatial logistic regression models improved the inter-occupant 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 %). However, as shown in Figs. 10 and 12, the estimated results deviated significantly from the observed probabilities at the state level, mainly due to sampling error, as observed for the spatial representations. Note that the estimated probabilities were close to the estimation result of the kriging and SAR; TAE and RMSE were 89.5 % and 9.8 % for kriging and 30.2 % and 1.8 % for SAR, respectively. In addition, owing to the influence of sampling errors, the results for the test dataset showed significant differences compared to the training dataset. This result implies that the segmentation approach is disadvantageous as it involves more sampling errors. The variable-based approach examined in Group 7 was useful for increasing the number of samples for each location. As dis-

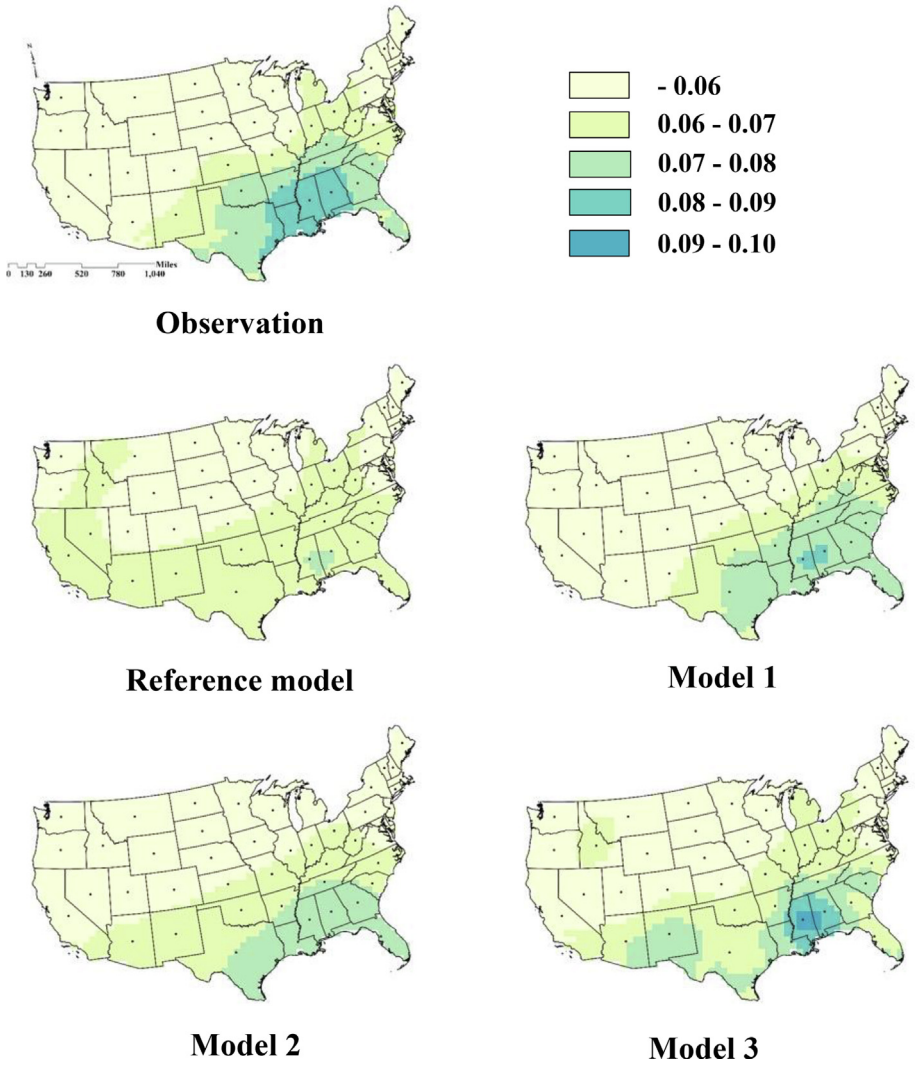

Fig. 14. Spatial probability of the women in Group 7 watching television at 13:00 based on observations and estimations.

Table 2
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 %

cussed in Section 4.4, the variable-based approach was effective as it approximately reflected the inter-occupant diversity, and the error was only marginally larger than the segmentation-based approach (the stacked TAE and RMSE increased by 1.3 and 0.1 %, respectively).

To overcome the sampling error issue, it is important to ensure that a sufficient number of samples is available for each study location. The 95 % confidence interval was calculated as $p \pm 1.96 * SE = p \pm 1.96 * \sqrt{\frac{p(1-p)}{n}}$, where p is the activity probability, SE is standard error, and n is the sample size. Fig. 16 shows the required number of samples for the corresponding width of the confidence intervals based on the calculation. As shown, to narrow the width of the confidence interval by 10 times, the required sample size needs to be increased by nearly 100 times. To obtain enough samples, it would be effective to 1) use a variable-based approach instead of a segmentation approach, 2) use multiple-year time use data, and 3) merge neighboring areas. The last method is important when high spatial resolution data are available because considering spatial variation at a detailed level reduces the sample size per location. In this case, using spatial representation methods is effec-

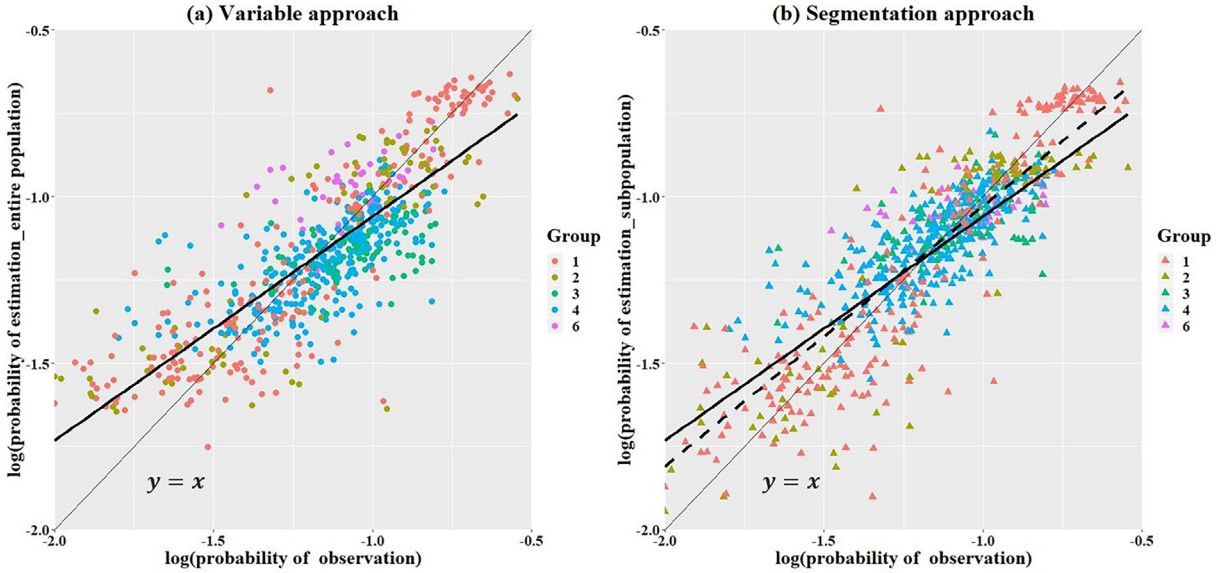

Fig. 15. Accuracy of Model 3 at the state level, considering two approaches (variables and segmentation). The different colors in the figure represent different groups. The circular and triangular shapes represent the entire population and the subpopulations, respectively. Logarithmic transformation was performed in the range of $(-2, -0.5) \times (-2, -0.5)$.

Table 3
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 %

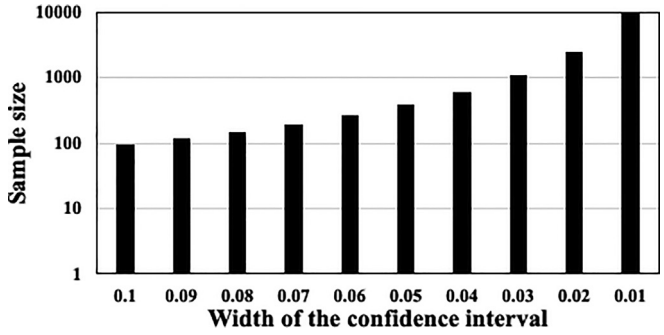

Fig. 16. Needed sample size for corresponding width of the confidence interval. p is considered as 0.5, representing the maximum value of $p(1-p)$.

tive in obtaining the spatial distribution of the modeling parameters of OBs.

5.2. Limitation and future work

As mentioned in Section 3 and dicussed in Section 5.1, a limited sample extracted from ATUS data representing women from states of the U.S. mainland for a limited number of activities and the low-resolution location data were used in this study. Therefore, the observations used to develop models contain non-negligible sampling errors. Thus, the developed spatial logistic regression models showed a large-scale error in the observed probabilities, whereas the developed models showed no significant improvement with the test dataset. However, further studies are required to address this issue.

Nevertheless, the developed modelling method can generate better results than traditional logistic regression methods, as revealed in Section 4.3. As time use data or equivalent datasets have been collected in many countries, the developed modelling method can be applied to different regions. For example, it applies to showing the differences in OBs between the areas in which lockdowns were implemented and those in which lockdowns were not implemented after the COVID-19 pandemic, thereby providing more useful references for relevant institutions. However, detailed information relevant to housing, households, and the environment should be supplemented by combining the data collected at the local level. Similarly, reliable new samples should be generated to enrich the sample size and represent spatial variation at the local level. In addition, the advancements in geographic information systems allow high-resolution location data to become more and more available. Thus, if the above conditions are satisfied, spatial representations can be generated with higher accuracy at the zip code or even household level. Therefore, subsequent spatial logistic regression methods can facilitate further improvements.

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 factors to insufficiently consider the spatial varia-

tion 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 consider 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. The developed modelling method is an extension of the exsiting logistic regression method which 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 at the state level from the ATUS data and low-resolution location data. Model performance may be improved with high resolution location data, and behavioral data with richer information and larger 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.

CRediT authorship contribution statement

Yuanmeng Li: Conceptualization, Methodology, Validation, Investigation, Data curation, Visualization, Writing – original draft. **Yohei Yamaguchi:** Conceptualization, Methodology, Supervision, Writing – review & editing. **Jacopo Torriti:** Conceptualization, Writing – review & editing. **Yoshiyuki Shimoda:** Writing – review & editing.

Data availability

Data will be made available on request.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by JSPS KAKENHI (grant number 20H02312) and JST SPRING (grant number JPMJSP2138); Japan Science and Technology Agency.

References

[1] A. Mastrucci, P. Pérez-López, E. Benetto, U. Leopold, I. Blanc, Global sensitivity analysis as a support for the generation of simplified building stock energy models, Energy Build. 149 (2017) 368–383, https://doi.org/10.1016/j.enbuild.2017.05.022.
[2] M. Osman, M. Ouf, A comprehensive review of time use surveys in modelling occupant presence and behavior: Data, methods, and applications, Build. Environ. 196 (2021), https://doi.org/10.1016/j.buildenv.2021.107785.
[3] D. Yan, T. Hong, B. Dong, A. Mahdavi, S. D'oca, I. Gaetani, X. Feng, IEA EBC Annex 66: definition and simulation of occupant behavior in buildings, Energy Build. 156 (2017) 258–270, https://doi.org/10.1016/j.enbuild.2017.09.084.
[4] W. O'Brien, H.B. Gunay, F. Tahmasebi, A. Mahdavi, A preliminary study of representing the inter-occupant diversity in occupant modelling, J. Build. Perform. Simul. 10 (5–6) (2017) 509–526, https://doi.org/10.1080/19401493.2016.1261943.
[5] D. Yan, W. O'brien, T. Hong, X. Feng, H.B. Gunay, F. Tahmasebi, A. Mahdavi, Occupant behavior modeling for building performance simulation: current state and future challenges, Energy Build. 107 (2015) 264–278, https://doi.org/10.1016/j.enbuild.2015.08.032.
[6] J. Li, Z. (Jerry) Yu, F. Haghighat, G. Zhang, Development and improvement of occupant behavior models towards realistic building performance simulation: a review, Sustain. Cities Soc. 50 (2019), https://doi.org/10.1016/j.scs.2019.101685.
[7] A. Druckman, T. Jackson, Household energy consumption in the UK: a highly geographically and socio-economically disaggregated model, Energy Policy 36 (8) (2008) 3177–3192, https://doi.org/10.1016/j.enpol.2008.03.021.
[8] S.H. Vega, E. van Leeuwen, N. van Twillert, Uptake of residential energy efficiency measures and renewable energy: Do spatial factors matter?, Energy Policy 160 (2022), https://doiorg/10.1016/j.enpol.2021.112659.
[9] A. Al-Mumin, O. Khattab, G. Sridhar, Occupants' behavior and activity patterns influencing the energy consumption in the Kuwaiti residences, Energy Build. 35 (6) (2003) 549–559, https://doi.org/10.1016/S0378-7788(02)00167-6.
[10] J. Torriti, Demand side management for the european supergrid: occupancy variances of european single-person households, Energy Policy 44 (2012) 199–206, https://doi.org/10.1016/j.enpol.2012.01.039.
[11] B. Jeong, J. Kim, R. de Dear, Creating household occupancy and energy behavioural profiles using national time use survey data, Energy Build. 252 (2021), https://doi.org/10.1016/j.enbuild.2021.111440.
[12] C.G.E. Ortiz-Ospina, M. Roser, Time Use, Our World Data, 2020.
[13] Y.S. Chiou, K.M. Carley, C.I. Davidson, M.P. Johnson, A high spatial resolution residential energy model based on American Time Use Survey data and the bootstrap sampling method, Energy Build. 43 (12) (2011) 3528–3538, https://doi.org/10.1016/j.enbuild.2011.09.020.
[14] A. Ibrahim, H. Ali, F. Abuhendi, S. Jaradat, Thermal seasonal variation and occupants' spatial behaviour in domestic spaces, Build. Res. Inf. 48 (4) (2020) 364–378, https://doi.org/10.1080/09613218.2019.1681928.
[15] V. Tabak, User Simulation of Space Utilisation: System for Office Building Usage Simulation, Technische Universiteit Eindhoven, Information Systems Built Environment, 2009.
[16] N. Mohammadi, J.E. Taylor, Urban energy flux: Spatiotemporal fluctuations of building energy consumption and human mobility-driven prediction, Appl. Energy 195 (2017) 810–818, https://doi.org/10.1016/j.apenergy.2017.03.044.
[17] J.W. Dziedzic, D. Yan, H. Sun, V. Novakovic, Building occupant transient agent-based model – Movement module, Appl. Energy 261 (7491) (2020), https://doi.org/10.1016/j.apenergy.2019.114417.
[18] K. Nassar, M. Elnahas, Occupant dynamics: Towards a new design performance measure, Archit. Sci. Rev. 50 (2) (2007) 100–105, https://doi.org/10.3763/asre.2007.5015.
[19] M. Kleinebrahm, J. Torriti, R. McKenna, A. Ardone, W. Fichtner, Using neural networks to model long-term dependencies in occupancy behavior, Energy Build. 240 (2021), https://doi.org/10.1016/j.enbuild.2021.110879.
[20] C. Wang, D. Yan, Y. Jiang, A novel approach for building occupancy simulation, Build. Simul. 4 (2) (2011) 149–167, https://doi.org/10.1007/s12273-011-0044-5.
[21] X. Feng, D. Yan, T. Hong, Simulation of occupancy in buildings, Energy Build. 87 (2015) 348–359, https://doi.org/10.1016/j.enbuild.2014.11.067.
[22] Y. Li, Y. Yamaguchi, Y. Shimoda, Impact of the pre-simulation process of occupant behaviour modelling for residential energy demand simulations, J. Build. Perform. Simul. 15 (3) (2022) 287–306, https://doi.org/10.1080/19401493.2021.2022759.
[23] L. Marín-Restrepo, M. Trebilcock, M. Gillott, Occupant action patterns regarding spatial and human factors in office environments, Energy Build. 214 (2020), https://doi.org/10.1016/j.enbuild.2020.109889.
[24] U. Wilke, F. Haldi, J. Scartezzini, D. Robinson, A bottom-up stochastic model to predict building occupants' time-dependent activities, Build. Environ. 60 (2013) 254–264, https://doi.org/10.1016/j.buildenv.2012.10.021.
[25] T. Okada, Y. Yamaguchi, Y. Shimoda, Data Preparation to Address Heterogeneity in Time Use Data Based Activity Modelling, Proceedings of Building Simulation 2019 16th Conference 16 (2020) 2356–2363. https://doi.org/10.26868/25222708.2019.211095.
[26] A. Rafiee, E. Dias, E. Koomen, Analysing the impact of spatial context on the heat consumption of individual households, Renew. Sustain. Energy Rev. 112 (2019) 461–470, https://doi.org/10.1016/j.rser.2019.05.033.
[27] N. Abbasabadi, M. Ashayeri, R. Azari, B. Stephens, M. Heidarinejad, An integrated data-driven framework for urban energy use modeling (UEUM), Appl. Energy 253 (2019), https://doi.org/10.1016/j.apenergy.2019.113550.
[28] N. Zhang, H. Huang, B. Su, X. Ma, Y. Li, A human behavior integrated hierarchical model of airborne disease transmission in a large city, Build. Environ. 127 (2018) 211–220, https://doi.org/10.1016/j.buildenv.2017.11.011.
[29] Y.-C. Wang, Examining landscape determinants of Opisthorchis viverrini transmission, Ecohealth 9 (3) (2012) 328–341, https://doi.org/10.1007/s10393-012-0789-z.
[30] G. Zhu, J. Liu, Q. Tan, B. Shi, Inferring the Spatio-temporal Patterns of Dengue Transmission from Surveillance Data in Guangzhou, China, PLoS Negl. Trop. Dis. 10 (4) (2016), https://doi.org/10.1371/journal.pntd.0004633.
[31] P. Monestiez, D. Courault, D. Allard, F. Ruget, Spatial interpolation of air temperature using environmental context: Application to a crop model, Environ. Ecol. Stat. 8 (4) (2001) 297–309, https://doi.org/10.1023/A:1012726317935.
[32] A. Degré, G.A. Tech, S.S. Passage, Different methods for spatial interpolation of rainfall data for operational hydrology and hydrological modeling at watershed scale : a review, Biotechnol. Agron. Soc. Environ. 17 (2013) 1–10.
[33] X. Xie et al., A review of urban air pollution monitoring and exposure assessment methods, ISPRS Int. J. Geo Inf. 6 (12) (2017) 1–21, https://doi.org/10.3390/ijgi6120389.
[34] D. Murakami, T. Yoshida, H. Seya, D.A. Griffith, Y. Yamagata, A Moran coefficient-based mixed effects approach to investigate spatially varying

relationships, Spatial, Statistics 19 (2017) 68–89, https://doi.org/10.1016/j.spasta.2016.12.001.
[35] C. Chasco, I. García, J. Vicéns, Modeling spatial variations in household disposable income with geographically weighted regression, Munich Personal RePEc Archive 50 (2007) 31.
[36] O. Berke, Estimation and prediction in the spatial linear model, Water Air Soil Pollut. 110 (3–4) (1999) 215–237, https://doi.org/10.1023/a:1005035509922.
[37] O. Berke, Modified median polish kriging and its application to the Wolfcamp-Aquifer data, Environmetrics 12 (8) (2001) 731–748, https://doi.org/10.1002/env.495.
[38] E.A. Varouchakis, Median polish kriging and sequential gaussian simulation for the spatial analysis of source rock data, J. Mar. Sci. Eng. 9 (7) (2021), https://doi.org/10.3390/jmse9070717.
[39] C. Xie, B. Huang, C. Claramunt, 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 (2000).
[40] C.J. Paciorek, Computational techniques for spatial logistic regression with large data sets, Comput. Statist. Data Anal. 51 (8) (2007) 3631–3653, https://doi.org/10.1016/j.csda.2006.11.008.
[41] L.C. Sayer, Gender, time and inequality: trends in women's and men's paid work, unpaid work and free time, Soc. Forces 84 (1) (2005) 285–303, https://doi.org/10.1353/sof.2005.0126.
[42] M. Li, N. Tilahun, A comparative analysis of discretionary time allocation for social and non-social activities in the U.S. between 2003 and 2013, Transportation 47(2) (2020) 893–909. https://doi.org/10.1007/s11116-018-9924-1.
[43] J. Gentry, S. Commuri, S. Jun, Review of Literature on Gender in the Family, Acad. Mark. Sci. Rev. 1 (2003) 1–18.
[44] D. Anxo, L. Mencarini, A. Pailhé, A. Solaz, M.L. Tanturri, L. Flood, Feminist Economy 17 (3) (2011) 159–195, https://doi.org/10.1080/13545701.2011.582822.
[45] J. Torriti, R. Hanna, B. Anderson, G. Yeboah, A. Druckman, Peak residential electricity demand and social practices: Deriving flexibility and greenhouse gas intensities from time use and locational data, Indoor Built Environ. 24 (7) (2015) 891–912, https://doi.org/10.1177/1420326X15600776.
[46] X. Xu, C. Fei Chen, Energy efficiency and energy justice for U.S. low-income households: an analysis of multifaceted challenges and potential, Energy Policy 128 (2019) 763–774, https://doi.org/10.1016/j.enpol.2019.01.020.
[47] M. J. Lőrincz, J. L. Ramírez-Mendiola, and J. Torriti, "Impact of time-use behaviour on residential energy consumption in the United Kingdom," Energies, vol. 14, no. 19, 2021, doi: 10.3390/en14196286.
[48] O. Lantang, A. Tiba, A. Hajdu, and G. Terdik, "Convolutional Neural Network for Predicting the Spread of Cancer," 10th IEEE Int. Conf. Cogn. Infocommunications, CogInfoCom 2019 - Proc., pp. 175–180, 2019, doi: 10.1109/CogInfoCom47531.2019.9089939.
[49] M. Chakraborty, S. K. Biswas, and B. Purkayastha, "Rule Extraction from Neural Network Using Input Data Ranges Recursively," New Gener. Comput., vol. 37, no. 1, pp. 67–96, 2019, doi: 10.1007/s00354-018-0048-0.
[50] M. Zhou, J. Li, R. Basu, J. Ferreira, Creating spatially-detailed heterogeneous synthetic populations for agent-based microsimulation, Comput. Environ. Urban Syst. 91 (2022), https://doi.org/10.1016/j.compenvurbsys.2021.101717.
[51] "Sample size issue when fitting logistic regression models," in Applied Logistic Regression, John Wiley & Sons, Ltd, 2013, pp. 401–408.
[52] M. Hayn, V. Bertsch, W. Fichtner, Electricity load profiles in Europe: The importance of household segmentation, Energy Res. Soc. Sci. 3 (C) (2014) 30–45, https://doi.org/10.1016/j.erss.2014.07.002.
[53] P.A.P. Moran, The interpretation of statistical maps, J. R. Stat. Soc. Ser. B 10 (2) (1948) 243–251, https://doi.org/10.1111/j.2517-6161.1948.tb00012.x.
[54] N. Cressie, Spatial prediction and ordinary kriging, Math. Geol. 20 (4) (1988) 405–421, https://doi.org/10.1007/BF00892986.
[55] P. Paul, M.L. Pennell, S. Lemeshow, Standardizing the power of the Hosmer-Lemeshow goodness of fit test in large data sets, Stat. Med. 32 (1) (2013) 67–80, https://doi.org/10.1002/sim.5525.
[56] S.J. Raudys, A.K. Jain, Small sample size effects in statistical pattern recognition: recommendations for practitioners, IEEE Trans. Pattern Anal. Mach. Intell. 13 (3) (1991) 252–264, https://doi.org/10.1109/34.75512.