OCCUPANT BEHAVIOR PREDICTION MODEL BASED ON ENERGY CONSUMPTION USING MACHINE LEARNING APPROACHES

By

Yunjeong Mo

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Planning, Design and Construction—Doctor of Philosophy

2018

ABSTRACT

OCCUPANT BEHAVIOR PREDICTION MODEL BASED ON ENERGY CONSUMPTION USING MACHINE LEARNING APPROACHES

By

Yunjeong Mo

Building sectors use the largest amount of energy among all energy-consuming sectors, and the residential sector constitutes 39 percent of the electricity consumption in the United States, which is the highest consumption among the various electricity-consuming sectors. The goals of this research are to identify a relationship between energy consumption and occupant behavior in a detailed level while also considering building technology, and to build a behavior prediction model using machine learning approaches based on energy consumption data. This research consists of four main parts: (1) Part I provides a theoretical foundation for the rest of the research, and develops the Occupant Behavior Prediction Model and apply the model to the American Time Use Survey (ATUS) data, (2) Part II focuses on analyzing energy usage-related behaviors and activities with the ATUS data, (3) Part III analyzes building technologies, including appliances, and energy usage with the Residential Energy Consumption Survey (RECS) data, and (4) Part IV combines the findings from the previous parts and applies the Occupant Behavior Prediction Model to the sensor-measured dataset. This research will have an impact on residential occupant behavior by helping occupants better understand their own behaviors' effects on energy usage, and detect what changes would improve energy efficiency in their homes. The findings will be beneficial to energy-related industries, and energy research area. In addition, the Occupant Behavior Prediction Model has the potential to be further integrated with research in other fields.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

OVERVIEW OF THE RESEARCH

1.1. Introduction

Building sectors use the largest amount of energy among all energy-consuming sectors, and approximately more than 70 percent of electricity and 50 percent of natural gas is consumed by the building sector in the United States (Diao, Sun, Chen, & Chen, 2017). The residential sector constitutes 39 percent of the electricity consumption in the United States, which is the highest consumption among the various electricity-consuming sectors (Johnson, Starke, Abdelaziz, Jackson, & Tolbert, 2014).

Residential building energy consumption is affected by various factors, such as climate, physical properties of the building, building services and energy systems, appliances in the household, occupants' activities and behavior, and the interactions among them (Widén & Wäckelgård, 2010). As the quality of thermal properties improves and the technologies for energy efficient appliances become more advanced, the overall energy consumption associated with buildings' physical properties and appliances is decreasing. Despite the decreased energy consumption due to the development of these technologies and the stricter requirements regarding energy efficiency of buildings and appliances, overall building energy consumption has not decreased (Chen et al., 2015). This can be explained by the influence of occupant behavior and living style, which emphasizes the significant role of occupant behavior in residential energy savings.

Unlike commercial building occupants, residential occupants have a high degree of energy control. They can control heating, ventilation, and air conditioning (HVAC) systems, lighting and electronic devices, and kitchen and laundry appliances, which are the main consumers of energy in residential buildings (Li & Jiang, 2006). This suggests that residential energy consumption can be significantly reduced by changing the energy usage-related behaviors of the occupants.

Various models explaining occupant behavior have been developed to estimate residential energy consumption. Darby (2006) stated that energy consumption was reduced by up to 20 percent when improved energy feedback was provided to the occupants. Wood and Newborough (2003) reported energy savings of more than 10 percent by using more specific information strategies. Similarly, Ouyang and Hokao (2009) reported an average of 14 percent energy savings achieved solely by improving occupant behavior.

Compared to the climate or buildings' physical attributes, occupant behavior is more difficult to quantify and assess. Recent studies (Aksanli, Akyurek, & Rosing, 2016; Diao et al., 2017; Sanquist, Orr, Shui, & Bittner, 2012; Santin, Itard, & Visscher, 2009) have analyzed detailed usage data of each appliance in a household to measure occupant behavior, since the use of appliances is heavily influenced by occupants' behavioral patterns at varying times and days. However, limitations still exist in the previous studies, and a more rational and systematic classification method for occupant behaviors and building attributes, and a solid model explaining their relationships, are needed to improve energy strategies.

1.2. Problem Statement

Several studies have examined occupant behavior with regard to energy consumption in residential buildings. However, a more comprehensive and systematic study is still needed to solve the existing problems outlined below.

> ***Problem #1****: There is a lack of comprehensive understanding of occupant behavior with regard to building technology and energy consumption in the residential sector.*

One of the significant barriers to finding a measurable relationship between occupant behavior and energy consumption is the lack of a thorough understanding of occupant behaviors in residential buildings.

Traditionally, behavioral patterns have been classified based on occupants' socioeconomic factors, such as age, gender, marital status, number of children, employment status, and income level. However, this method has significant shortcomings, as socioeconomic factors cannot fully explain their energy consumption patterns. Even if occupants have similar characteristics, it does not guarantee similar behavioral patterns (Diao et al., 2017).

Occupant behavior is associated with more than just socioeconomic factors, and actual occupant behavior is determined by multifaceted variables. It is critical to comprehensively identify all the relevant occupant characteristics and the hierarchy of occupant behavior, along with other external factors such as building attributes and climate. Therefore, a systematic and thorough approach is

necessary to define and understand occupant behavior comprehensively, and furthermore, to predict the resulting energy consumption in a more consistent and accurate way (Chen et al., 2015).

***Problem #2**: There is an absence of systematic structures explaining the relationship between occupant behavior, building technology, and energy consumption.*

Several studies have made efforts to identify the influences of occupant behavior and building technology on building energy consumption. Researchers also measured the influence of occupant behavior on energy consumption through observations and surveys. Although it is obvious that occupant behavior influences building energy usage, previous studies have lacked thorough and clear methods to quantify the effects of occupant behavior. The main reason is that various factors have influences on energy consumption simultaneously, and the individual and interactive effects of these factors are not clearly identified yet (Chen et al., 2015).

Yu et al. (2011) identified the simultaneous influences of behavior, physical building attributes, and external environmental factors on building energy consumption. However, the existing methods could not isolate the sole influence of occupant behaviors by removing the effects of other factors.

Compared to physical building attributes, such as thermal environment and envelope of the building, occupant behavior is difficult to assess and measure. In addition, when occupant behavior is combined with energy consumption and building attributes, the quantitative assessment of occupant behavior becomes even more complicated. The absence of a systematic model explaining

the relationship between behavior, technology, and energy is another significant obstacle to assessing occupant behavior quantitatively.

Problem #3*: There lacks a model to explain and predict occupant behavior.*

Although the American Society of Heating, Refrigerating, and Air-Conditioning Engineers (ASHRAE) suggests a standardized occupancy schedule to assess building energy, occupant behavior patterns and time use could be different for each household due to the different occupants' lifestyles, preferences, and other factors.

Most of the existing occupant behavior models have been implemented using survey data. They concentrated on statistical analysis of occupants' sociodemographic characteristics to predict their energy consumption, which implies the actual user behaviors were mostly guessed (Aksanli et al., 2016; Diao et al., 2017). In order to apply this framework to real-world cases to predict and quantify occupant behavior, the model needs to be more directly based on the actual measured data, but there is a lack of such a detailed prediction model in existing studies.

In addition, occupant behavior, building technology, and energy consumption simultaneously interact with one another, and their data are usually recorded as different types, such as numerical, ordinal, categorical, or text types. The concurrent effects of multiple factors and various types of data are difficult for quantitative data analysis to process, and increase the complexity of distinguishing the effects of occupant behavior from other building-related factors. Thus, an occupant behavior prediction model is needed to solve these issues.

To address the problems of the existing studies, the following hypotheses are established:

- Occupant behavior, energy consumption, and building technology interact all together and their interaction can be explained more effectively by understanding the procedure of behavior formation.
- Occupant behavior can be predicted based on their energy consumption pattern.

1.3. Goals and Objectives

The goals of this research are to understand occupant behavior based on energy consumption while also considering building technology, and to build a behavior prediction model using machine learning approaches on energy consumption data. This model can potentially be used for efficient building operation and control strategies. Unlike previous studies, which focused on using occupant behavior to predict energy consumption, or changing occupant behavior with interventions or education, this study investigates the reverse prediction model: using energy consumption to predict occupant behavior.

In this research, human behavior can be narrowed down to building occupant behavior. Energy consumption and building technology information are used as inputs to predict occupant behavior as the output. The research aims to reduce the gap between energy consumption and occupant behavior, and to optimize technologies for occupant behavior (Figure 1-1).

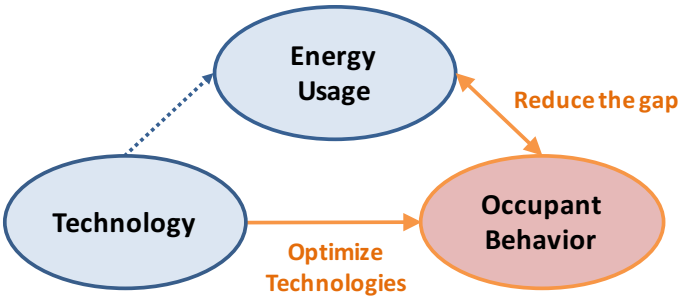


Figure 1-1. Research Goal Overview

Based on the problems stated earlier, the objectives of this research are to develop the prediction model as follows.

Objective 1: *Create a structured list of occupant behaviors, building technologies, and energy consumption based on comprehensive and refined definitions of each category.*

In order to achieve this objective, first, each main category (occupant behavior, building technology, energy usage) will be examined individually. The subcategories and elements under the main categories will be specified with various techniques, such as literature reviews and machine learning algorithms based on the interactions between energy consumption and other elements. The properties of occupant behavior will be assessed as a single activity level, quantitatively measured with its frequency per day, duration per day, and energy impact. Building technology will be defined based on the subcategories of (1) heating and cooling, (2) light and appliances, (3) ventilation, (4) water, (5) design and construction, and (6) insulation. In addition, the ideal time-interval (e.g. 5-minute interval, daily/weekly/monthly data interval) of time-series data will be examined for each critical element of occupant behavior and energy usage.

The comprehensive and inclusive elements of energy, building technologies, and behavior will be collected from existing study results and national public databases, such as the Residential Energy Consumption Survey (RECS)(EIA, 2018) and the American Time Use Survey (ATUS) (U.S.BLS, 2018) results. Among all of the elements collected, the ones critical to energy consumption will be identified and subsets of the highest impact elements will be selected. The refined final list will be used for the next step.

***Objective 2**: Establish an occupant behavior model that explains the systematic relationships between categories (occupant behavior, building technology, energy usage), and interactions between a more detailed level of features.*

The relationships between the categories and the interactions between individual elements will be evaluated. An occupant behavior model will be established using network-analysis and meta-analysis based on the relationships between categories. The model will be applied to a building simulation in order to evaluate the behavior patterns of the residential building's occupants, and the effectiveness of the model will be evaluated.

***Objective 3**: Explain and predict occupant behavior using machine learning algorithms.*

The model will be used to explain and make detailed predictions about occupant behavior, including activity, frequency, duration, and effect on energy consumption using machine learning algorithms. The complex relationships between building technology, energy usage, and user behavior will be simultaneously modeled by different subcategories using multi-task learning techniques. Different than most of the existing models, occupant activity data showing the

interactions between occupants and appliances are used, which have been measured in five-minute intervals using sensors. In this step, actual sensor-measured data will be used to train the model. The machine learning model will be evaluated by applying it to other datasets.

1.4. Research Structure

1.4.1. Research Design and Structure

Figure 1-2 explains the overall research structure: developing a behavior prediction model and validating the model using two different datasets. Additional analysis supports the second validation of the model. The reliability and validity of the research are achieved by the triangulation of dual validations using different data sources and types.

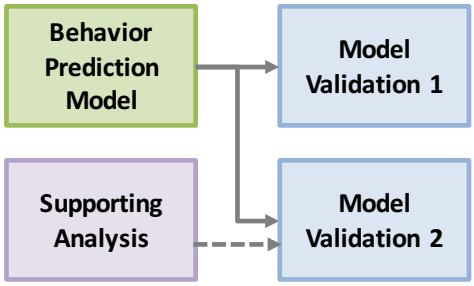


Figure 1-2. Research Structure Overview

Figure 1-3 summarizes the detailed research process, grouped into its major parts. Each part is independent, but all four parts are connected. First, the Occupant Behavior Prediction Model is developed and validated by applying it to the American Time Use Data (ATUS), the Residential Energy Consumption Survey (RECS), and sensor-measured data. Each part is explained as follows.

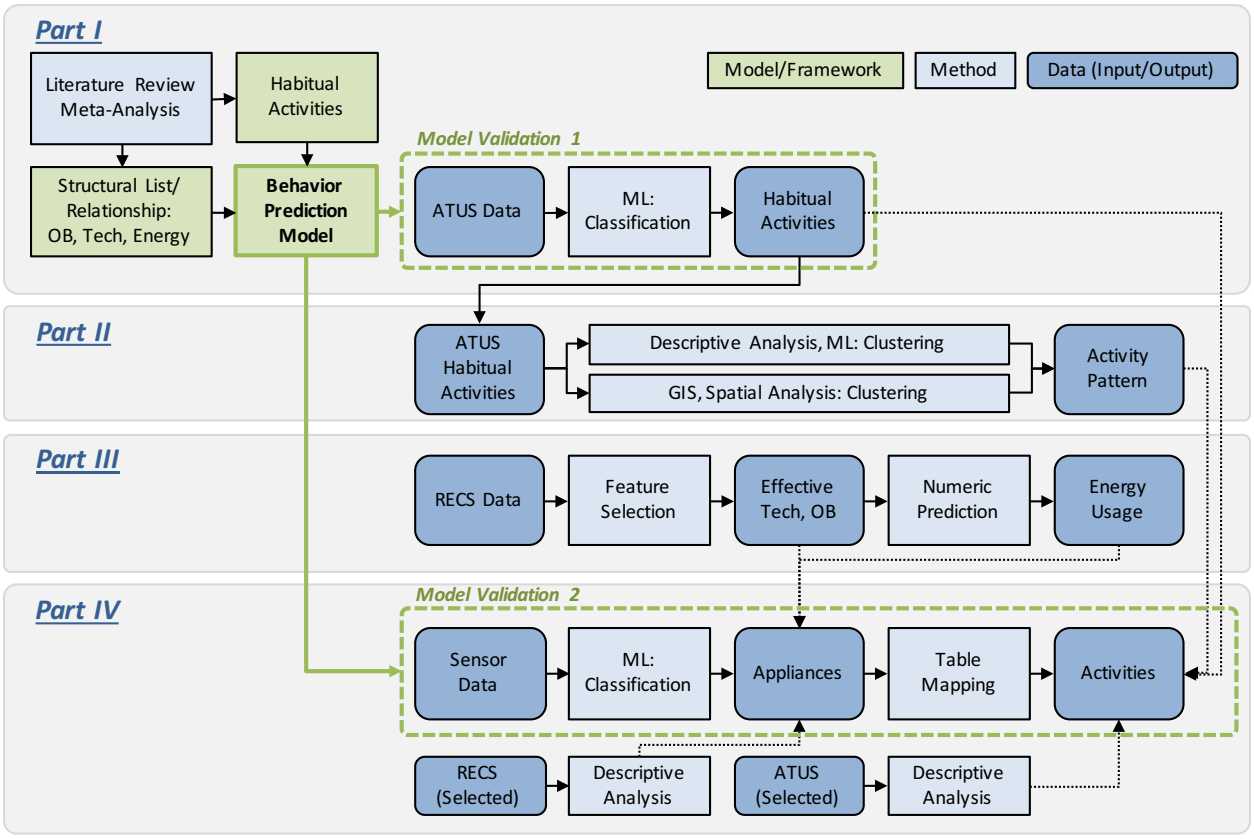


Figure 1-3. Research Design and Structure

Part I (Chapter 2): Objective 1, 2, 3

Part I aims to achieve Objectives 1, 2, and 3. It provides a theoretical foundation for the rest of the research, and develops the Occupant Behavior Prediction Model with the following steps. The ATUS data are used in Part I.

- Review existing literature/studies about occupant behavior and energy usage, and perform meta-analysis to combine the results of the selected studies.
- Derive structured lists and relationships between occupant behavior, building technology, and energy usage.

- Review existing literature/studies about habitual behaviors and activities from psychology, business, and building energy studies, then delineate the main characteristics of habitual behaviors and activities.
- Develop the Occupant Behavior Prediction Model based on the structured lists and the characteristics of habitual activities.
- Apply the model to the ATUS data using machine learning classification algorithms to predict energy usage–related activities and to define habitual/predictable activities (Model Validation 1).

Part II (Chapter 3): Objective 1, 3

Part II aims to achieve Objectives 1 and 3, and focuses on analyzing energy usage-related behaviors and activities with the steps below. The ATUS habitual energy usage–related activities defined in Part I are used as input data in Part II.

- Perform descriptive analysis and K-modes clustering to identify patterns in the energy usage–related activities.
- Perform spatial analysis (K-means clustering) and demonstrate the geographical differences of the activities using geographical information system (GIS).
- Detect the activity patterns by region, gender, day of the week, etc.

Part III (Chapter 4): Objective 1, 3

Part III aims to achieve Objectives 1 and 3, and focuses on analyzing building technologies, including appliances, and energy usage. The RECS data are used in Part III.

- Select features that have significant impacts on energy usage. The categories of the features include home appliances, building envelopes, demographic information of the respondents, occupant behavior, etc.
- Predict energy consumption with the selected features using machine learning algorithms to verify the features' predictive effectiveness.

Part IV (Chapter 5): Objective 2, 3

Part IV aims to achieve Objectives 2 and 3. It combines the findings from the previous parts and applies the Occupant Behavior Prediction Model to the sensor-measured dataset. The sensor-measured data are mainly used, and the ATUS and the RECS are also used to support further analysis in Part IV.

- Predict appliances with the features specified by the Occupant Behavior Prediction Model using machine learning numeric prediction algorithms on the sensor-measured data.
- Estimate the related activities using the appliance-activity mapping table defined in Part I.
- Support appliance information with the selected features and energy consumption from Part III, and additional descriptive analysis of the RECS.
- Support activity information with the habitual activities from Part I, activity patterns from Part II, and additional descriptive analysis of the ATUS.

1.4.2. Main Datasets

In this research, various types of empirical data are collected as follows.

- 1. Energy Consumption Data
    - Install sensors (electricity) in participants' residential buildings

- Weekly data downloads to save more granular data (5-minute intervals, .csv file type)

- 2. Building Technology Data
  - Technical data (year built, building type, size, materials, energy certification, green building technology, etc.) is collected by survey or site visit
  - Weather data is collected from weather station websites during the measurement period
- 3. User Behavior Data
  - Major occupant behaviors are recorded as a form of appliance-level energy data measured by the sensors
  - They are also captured by analyzing patterns in the aggregated energy consumption data using nonintrusive load monitoring (NILM)
- 4. Residential Energy Consumption Survey (RECS) Data
  - Download from the RECS website
  - Latent variables represent qualities that are not directly measured, but rather inferred from the observed covariation among a set of variables (Piedmont, 2014). Latent variables are examined among those three types of datasets, and the RECS data are also analyzed to identify other potential latent variables.
- 5. American Time Use Survey (ATUS) Data
  - Download from the ATUS website
  - Includes respondents' time use data of each activity done in a day
  - Occupant behavior activities are derived from the ATUS datasets.

1.4.3. Main Methodology: Machine Learning

In this research, Machine Learning (ML) approaches are mainly used for data analysis, which is novel in occupant behavior studies. Other quantitative methods can be also considered depending on the characteristics of the input datasets and the format of the expected outcome. This study uses multiple large datasets with more than 270 features and/or more than 70000 instances. Therefore, ML methods are selected since ML involves searching a very large space of possible hypotheses to find one that best fits the observed data and any prior knowledge held by the learner (Mitchell, 1997).

ML is concerned with answering questions such as the following (Bishop, 2006; Mitchell, 1997):

- What algorithms exist for learning general target functions from specific training examples? In what settings will particular algorithms converge to the desired function, given sufficient training data? Which algorithms perform best for which types of problems and representations?
- How much training data is sufficient?
- When and how can prior knowledge held by the learner guide the process of generalizing from examples?
- What is the best strategy for choosing a useful next training experience?
- What is the best way to reduce the learning task to one or more function approximation problems?
- How can the learner automatically alter its representation to improve its ability to represent and learn the target function?

As described above, generalization performance and model complexity regarding testing/training data are fundamental concerns in ML, and they will be further examined in the following sections.

Generalization Performance

Generalization is the ability to correctly categorize new examples that differ from those used for training. In practical applications, the variability of the input vectors will be such that the training data can comprise only a tiny fraction of all possible input vectors, so the ability to generalize and make accurate predictions for new data is a central goal in ML (Bishop, 2006).

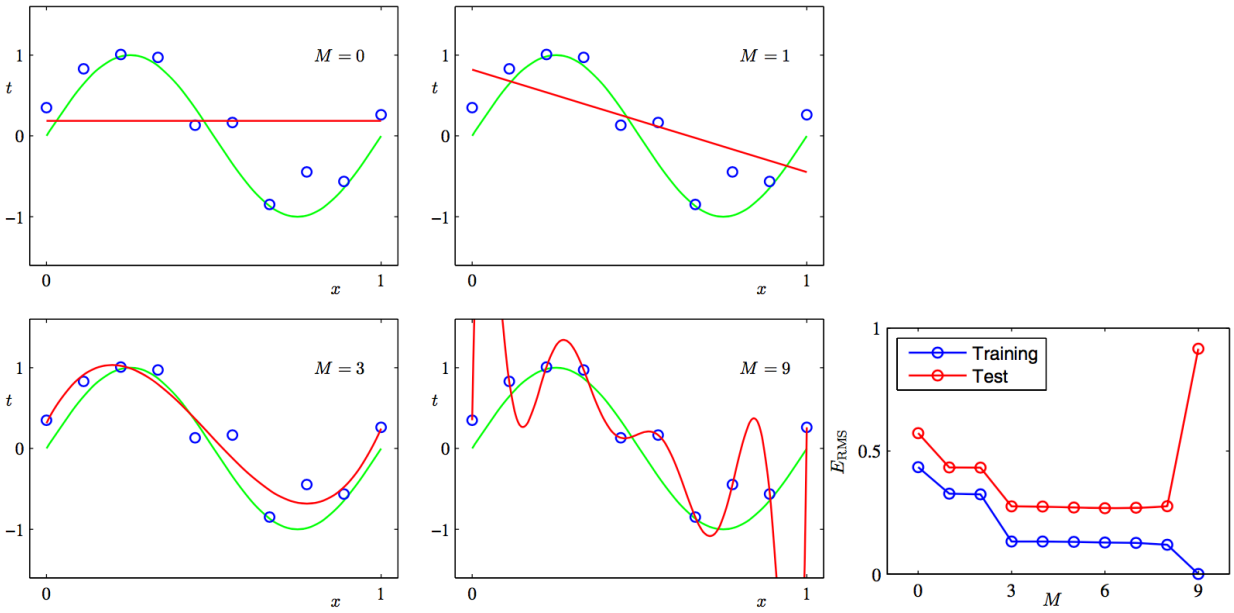


Figure 1-4. Plots of Polynomials Having Various Orders (Bishop, 2006)

Figure 1-4 illustrates plots of polynomials with various orders and the root-mean-squared-error (RMSE) of training and test sets for each order M. As the order increases (signifying more complex models), the training set error goes to zero (when M = 9). However, this model has an over-fitting problem, and the test set error value becomes very large (Bishop, 2006). The model's

generalization performance can be improved with some strategies, and two examples are explained below.

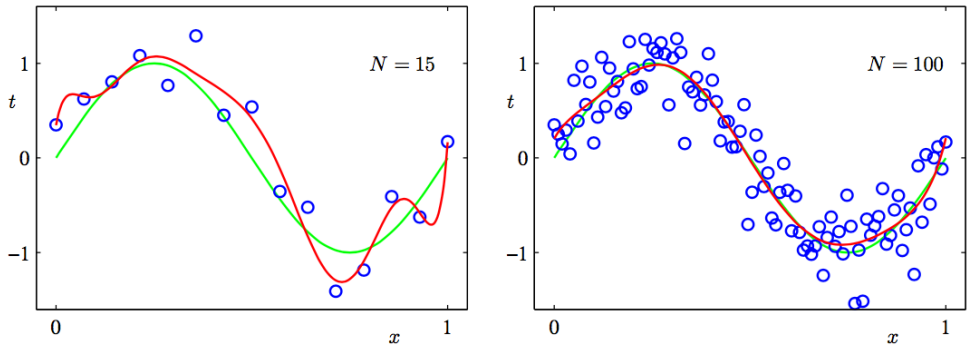


Figure 1-5. Reducing Over-Fitting with More Data (Bishop, 2006)

As seen in Figure 1-5, for a given model complexity, the over-fitting problem becomes less severe as the size of the dataset increases. In other words, the larger the dataset, the more complex (more flexible) the model we can afford to fit to the data.

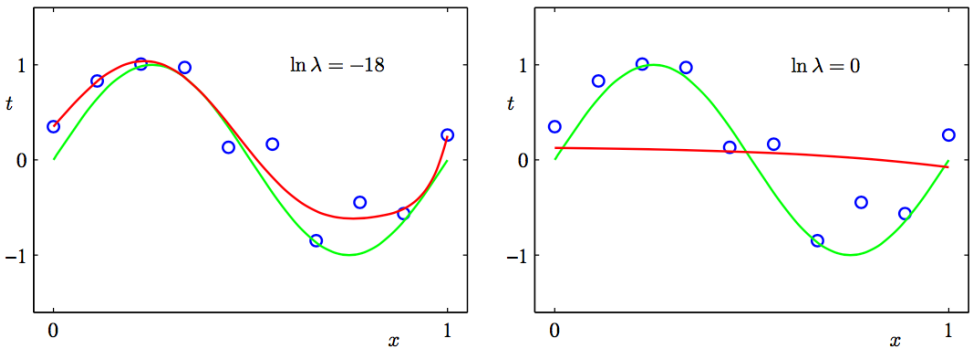


Figure 1-6. Reducing Over-Fitting with Regularization (Bishop, 2006)

However, it is not always possible to have enough data, and we should consider how we can apply the model to datasets of limited size where we may still wish to use relatively complex and flexible models. One technique that is often used to control the over-fitting phenomenon in such cases is that of regularization, which involves adding penalty terms to the error function. Figure 1-6 shows the results of fitting the polynomial of order M = 9 to the same dataset as before, but now using

the regularized error function. For a value of $\ln \lambda = -18$, the over-fitting has been suppressed and the model obtains a much closer representation of the underlying function. However, if we use too large a value for λ, then we again obtain a poor fit, as shown for $\ln \lambda = 0$ (Bishop, 2006).

Model Complexity

If we were trying to solve a practical application using this approach of minimizing an error function, we would have to find a way to determine a suitable value for the model complexity. In the previous example of polynomial curve fitting using least squares, we saw that there was an optimal order of polynomial that gave the best generalization. The order of the polynomial controls the number of free parameters in the model and thereby governs the model's complexity. With regularized least squares, the regularization coefficient λ also controls the effective complexity of the model, whereas for more complex models, such as mixture distributions or neural networks, there may be multiple parameters governing complexity. In practical applications, we need to determine the values of such parameters, and the principal objective in doing so is usually to achieve the best predictive performance on new data. As well as finding the appropriate values for complexity parameters within a given model, we may also wish to consider a range of different types of model in order to find the best one for our particular application (Bishop, 2006).

ML Model Selection

The phenomenon of model complexity and over-fitting can be considered with bias-variance tradeoff. Bias is the extent to which the average prediction over all datasets differs from the desired regression function, and variance is the extent to which the solutions for individual datasets vary around their average (Zhou, 2016). In ML models, low bias and low variance are the most desirable,

and high bias and high variance are the least desirable. However, bias and variance have a trade-off relationship in model complexity as shown in Figure 1-7. Thus, it is important to find the optimal model complexity to minimize the expected loss.

$$Expected\ Loss = (Bias)2 + Variance + Noise$$

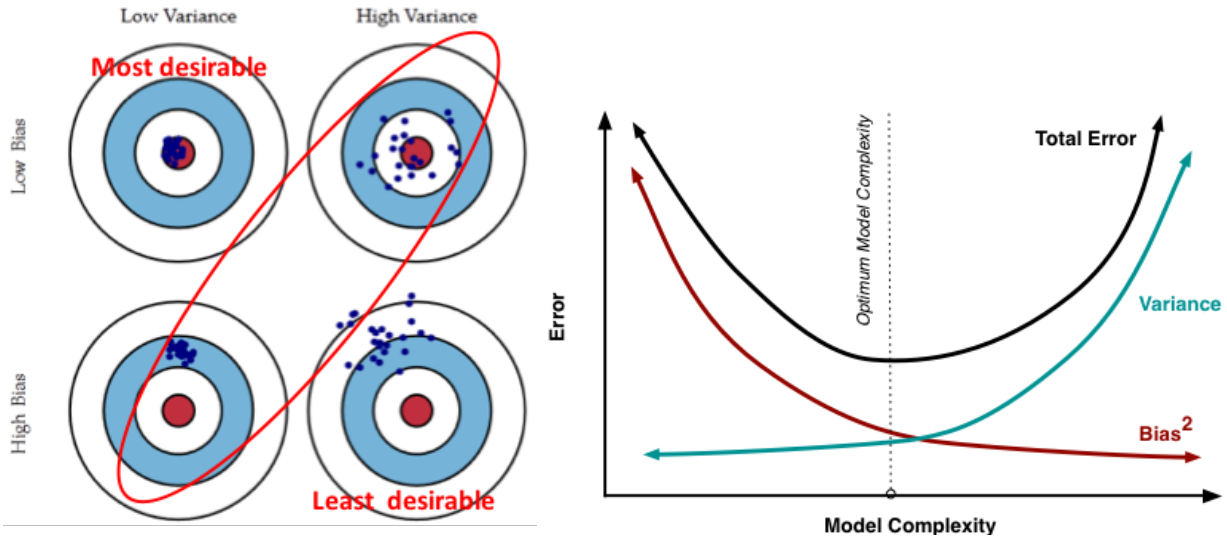


Figure 1-7. Bias-Variance Tradeoff (Dubrawski, 2015)

In a given relationship, flexible models with strong approximators (high degree polynomials) have low bias and high variance, and rigid models with weak approximators (low degree polynomials) have high bias and low variance (Zhou, 2016). This is further explained in the next example.

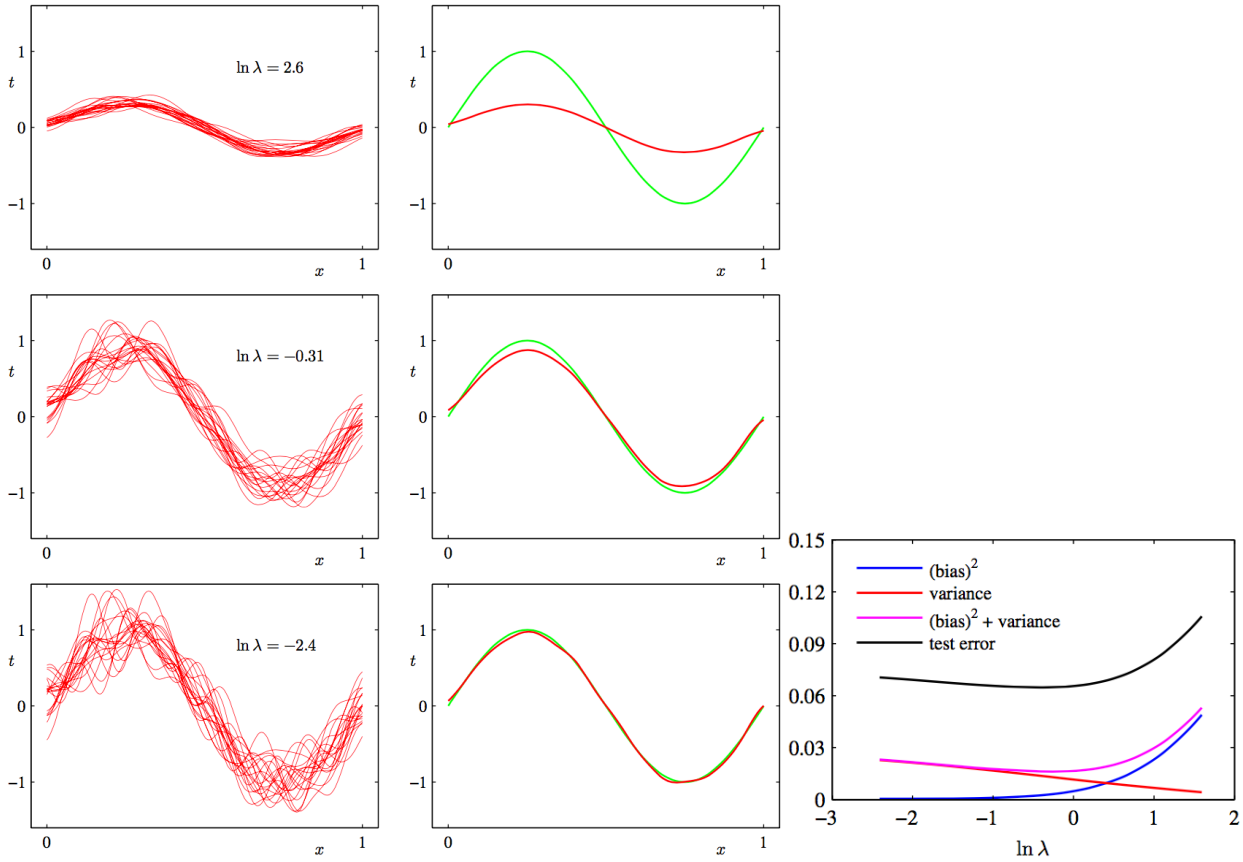


Figure 1-8. Dependence of Bias and Variance on Model Complexity (Bishop, 2006)

Figure 1-8 illustrates the dependence of bias and variance on model complexity, governed by a regularization parameter λ. The left column shows the result of fitting the model to the datasets for various values of ln λ (for clarity, only 20 of the 100 fits are shown). The center column shows the corresponding average of the 100 fits (red) along with the sinusoidal function from which the datasets were generated (green). The right graph is the plot of squared bias and variance, together with their sum, corresponding to the results shown on the left side. Also, the average test set error for a test dataset size of 1000 points is shown. In this example, the minimum value of $(\text{bias})^2$ + variance occurs around $\ln \lambda = -0.31$, which is close to the value that gives the minimum error on the test data (Bishop, 2006).

Although the bias-variance decomposition may provide some interesting insights into the model complexity issue from a frequentist perspective, it has its limitations. Bias-variance decomposition is based on averages with respect to ensembles of datasets, whereas in practice, we have only a single observed dataset. If we had a large number of independent training sets of a given size, we would be better off combining them into a single large training set, which of course would reduce the level of over-fitting for a given model complexity (Bishop, 2006).

In the larger picture, in order to choose the best ML models and to design a good learning system, we should consider (1) training experience, (2) target function, (3) representation of the target function, and 4) function approximation algorithm. The type of training experience available can have a significant impact on the success or failure of the learner (Mitchell, 1997). There are three key attributes to a good training experience:

- Whether the training experience provides direct or indirect feedback regarding the choices made by the performance system.
- The degree to which the learner controls the sequence of training examples.

How well the training experience represents the distribution of examples over which the final system performance must be measured.

1.5. Research Scope and Assumptions

In this research, the building type is confined to residential buildings, and the boundary of occupant behavior for the machine learning (ML) model is limited to energy consumption–related behaviors. Due to limitations of the measurements, the ML model contains occupant behavior and energy consumption data regarding electricity and gas, but does not include water-related data.

1.6. Definition of Occupancy

Depending on the purpose of the research, researchers define "occupant behavior" from different perspectives. Thus, the definition and scope of occupant behavior need to be defined at the early stages of research. In this research, occupant behavior is limited to only energy use within a built environment, especially in residential buildings. Existing studies generally divide the effects of occupant behavior into two categories: (1) simple occupancy effects on building energy consumption, and (2) occupants' actions/activities influencing energy consumption (Yu et al., 2011).

Chen et al (2015) defined behavior as discernable actions or reactions of a person in response to external or internal stimuli, or to adapt to external conditions such as weather or indoor air quality. In built environments, the impact of behavior on building energy consumption is closely related to building elements, such as windows and curtains, and appliances controlled by the occupants. Thus, the operation of building elements and appliances indicate occupant behavior (Chen et al., 2015).

Santin (2011) defined behavior to include all activities of occupants in the residential building. In particular, they defined "use" as the direct interaction between an occupant and an action to accomplish a certain goal. Occupant behavior was specified further as the use of residential space, building systems, and other services in the house that can affect energy consumption, including space and water heating. In many cases, an occupant's psychological factors, including attitudes and motivations, leading to a specific action are explained separately (Chen et al., 2015).

1.7. Relationship between Energy, Technology, and Behavior

Energy, technology, and behavior are the three main concepts in residential energy studies. In most of the current research, the influences of technology and behavior on energy consumption have been studied separately. However, recent studies have introduced a novel way to explore the relationship of the main concepts, and this research will accept that new point of view.

Zhao et al. (2017) explained these concepts as follows. (1) Home energy consumption is measured and recorded by utility companies. They combine monthly consumption data, distribution, transmission, taxes, and service charges to produce a monthly energy bill that is sent to the occupant for their previous month's service. (2) "Green building technology" refers to the collection of advanced technologies and products for building design and construction that reduce overall energy use and carbon emissions. (3) Occupant behavior and its position as part of the overall development process, and specifically with building systems, is critical for understanding residential energy consumption. Occupant actions impacting energy use can be divided into three categories: time-related usage, environment-related mode, and quantitatively described behavior.

Existing studies suggest that the efficiency and efficacy of building technology, such as heating and cooling systems, have a considerable impact on residential energy consumption. Literature also asserts that some resident behaviors considerably affect home energy use. Most of the existing studies investigated the effects of either occupant behaviors or technologies. However, Zhao et al. (2017) identified a new point of view on energy efficiency in residential buildings. Unlike other earlier studies, which isolated the effects of technology or behavior on energy consumption, their

study investigated the interaction between building technology and occupant behavior and their joint impact on energy use (Figure 1-9).

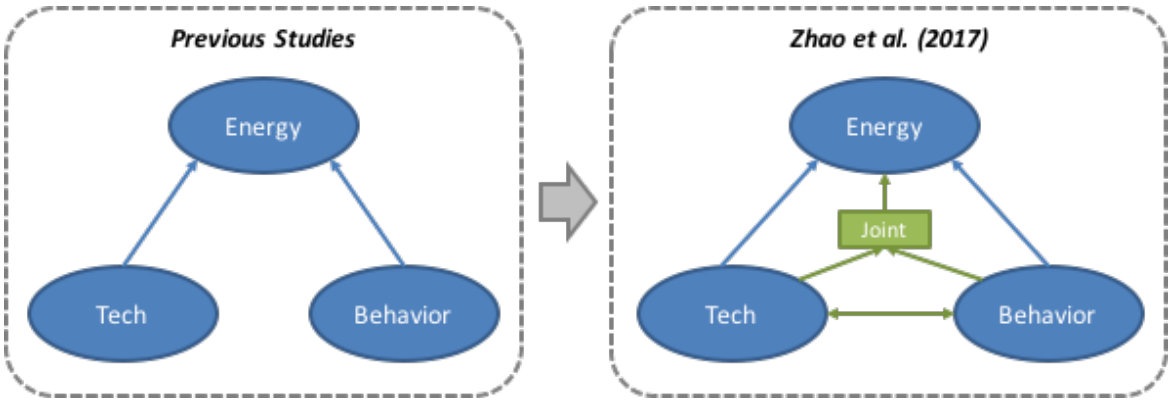


Figure 1-9. Differences of Models

Occupant behavior and building technology are two indispensable factors for enabling energy efficiency. In addition, the effects of either behavior or technology depend on the other's specific values. The effects of one level of technology vary for different occupants, and vice versa. It is obvious that a higher level of green building technology will lead to less energy use. However, Zhao et al. (2017) argued that when considering the interaction with occupant behavior, the most advanced technologies might not necessarily be the optimal option for all occupants. They asserted that the identified interaction effects are mutual rather than one-way, and thus implied that behavior can impact the technology's performance, and that performance can influence occupant behavior in kind (Zhao et al., 2017).

1.8. Summary

In this chapter, problems of the existing studies were examined and the goals and objectives of this research were defined based on the hypotheses to address the problems. Then, the research design and structure was explained. The term "occupancy" was defined for this research, and the

relationship between the three main data categories (occupant behavior, building technology, and energy usage) were discussed. More details will be studied in the following chapters.

CHAPTER 2

OCCUPANT BEHAVIOR PREDICTION MODEL ON ENERGY CONSUMPTION IN RESIDENTIAL BUILDINGS

Abstract

Occupant behavior consists of multifaceted variables and thus a systematic approach is required to comprehensively understand occupant behavior. This research aims to define a structure of relationship between energy consumption, building technology, and occupant behavior, using the Occupant Behavior Prediction Model. The model can predict and explain occupant energy usage-related activities. This model can also identify the predictability and habitual characteristics of each activity. A machine learning approach is used to develop the model, and datasets from the American Time Use Survey (ATUS) are used to verify the model. The results show that the energy use activities with higher predictive performances are more stable and habitual compared to the ones with lower predictive performances. Occupants' habitual behaviors are difficult to change, but they are more predictable. The prediction accuracy achieved by this model for these habitual activities reached as high as 99%. For example, the accuracy was 99% when predicting washing and grooming activity, and 82% for watching TV. Such findings imply that the building systems and control strategies need to be adjusted to accommodate habitual energy use behaviors, rather than changing the behaviors. In addition, educational interventions seem more effective on the less habitual behaviors, which often change.

2.1. Introduction

Residential building energy consumption is affected by climate, physical properties of the building, building services and energy systems, appliances in the household, occupant behavior, and the interactions among them (Widén & Wäckelgård, 2010). As the building technologies grow more advanced, the energy consumption in residential buildings becomes more influenced by occupant behavior and living style, which emphasizes the need to understand occupant behavior and the relationship between occupant behavior and energy consumption.

Occupant behaviors have been often studied based on socioeconomic factors, such as age, gender, marital status, number of children, employment status, and income level. However, this method

has significant shortcomings, in that socioeconomic factors cannot fully explain occupants' energy consumption patterns. Even if occupants have similar socioeconomic characteristics, these similar characteristics do not guarantee similar behaviors. When an analysis only considers socioeconomic factors, the result will provide limited information (Diao et al., 2017).

Occupant behavior is associated with more than socioeconomic factors, and occupant behavior can be caused by a variety of factors. It is critical to comprehensively identify not just occupant-specific characteristics like socioeconomic status and behavior hierarchy, but also external factors such as building attributes and climate. Therefore, a model is necessary to define and understand occupant behavior comprehensively (Chen et al., 2015).

Many studies have examined the relationship between occupant behavior and energy consumption in residential buildings. However, a more comprehensive study is still needed to solve several existing problems, which are as follows: first, there is a lack of comprehensive understanding of occupant behavior regarding building technology and energy consumption in residential sectors, and second, there is an absence of systematic models able to predict the behaviors and the habitual properties of activities related to residential energy usage. In order to solve these problems, this research aims to define a model of relationships between energy consumption, building technology, and energy usage-related behavior, then uses that model to explain occupant behavior. This model is applied to predict occupants' behavior and to identify how predictable and habitual each activity is. "Habitual behavior" denotes a behavior influenced by habits. This new model integrates the concept of habitual behavior and reduces the gap between energy consumption and occupant behavior. The outline of this chapter is summarized in Figure 2-1.

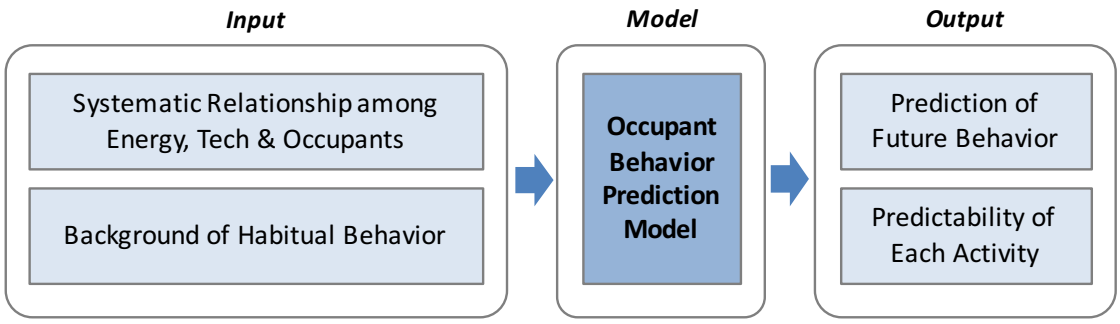


Figure 2-1. Chapter Outline

The structure of this chapter is as follows. First, the background section explains the categories of behavior, introduces the stability and habitual characteristic of behavior, and analyzes the relationship between energy, building technology, and occupant behavior. Then, a behavior prediction model is defined based on the concept of habitual behavior, and the main components of the model are specified. This model is applied to a case study with a machine learning approach. In the case study, an overview of the dataset is explained, and the methodology and results follow. Finally, the implications and limitations of this study are discussed.

2.2. Theoretical Background

2.2.1. Habitual Occupant Behavior

Behavioral routines and lifestyles are critical for energy saving because they have significant influences on daily energy use, but they are difficult to affect, changing gradually over time or not at all. Most people want to maintain their existing behavioral routines, lifestyles, and habits. Therefore, changing attitudes is easier than changing behaviors, and many studies report that building occupants' attitudes have changed to be more energy-conscious, but they are unlikely to change their behaviors to match (Lutzenhiser, 1993).

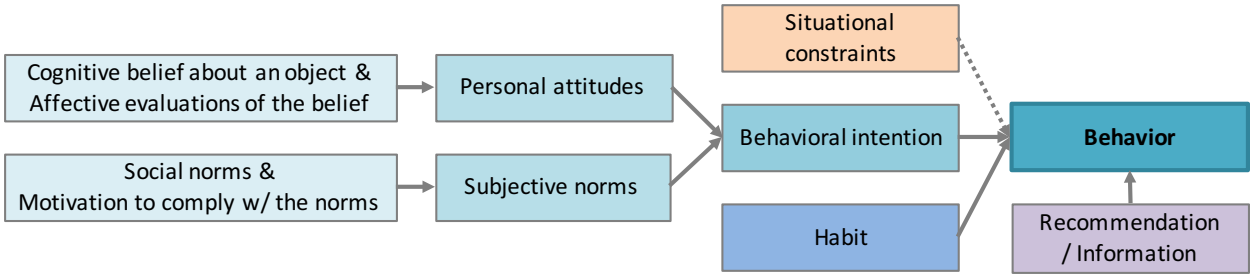


Figure 2-2. Formation of Occupant Behavior

Energy usage-related behavior is determined by behavioral intention and influenced by habit and situational constraints (Figure 2-2). Van Raaij and Verhallen (1983) explain that behavioral intention determines behavior. Behavioral intention is the subjective probability that a person will perform a behavior, and it is created by personal attitudes and subjective norms, if there are no unanticipated situational constraints. Personal attitudes about an object are constituted of cognitive beliefs about an object and affective evaluations of the beliefs. Subjective norms are determined by social norms and the motivation to abide by the norms.

Energy-related personal attitudes include concerns about energy price, environment, building energy efficiency, health, and personal comfort (Van Raaij & Verhallen, 1983). As discussed before, these attitudes influence behavior by affecting behavioral intention. Van Raaij and Verhallen (1983) explained that people try to be consistent in their personal attitudes and behaviors, and if we change behaviors to be more energy-saving, people may develop energy-conscious attitudes. However, energy-conscious attitudes do not always cause energy-saving behavior. Also, certain behaviors may be directly changed through recommendations, prompts, and information (i.e. rewards, information about energy costs) without changing attitude first (Lutzenhiser, 1993). Attitudes may develop good behavioral intentions, but when the subjective norms are weak, the behavioral intention cannot be fully influenced.

Situational constraints may also hinder behavioral intentions from realizing actual behaviors. Thus, a desired behavior can be achieved when a person has positive personal attitudes and subjective norms without situational constraints (Van Raaij & Verhallen, 1983). Additionally, repeated past behaviors form habits, which affect future behavior. This means that not only changes in personal attitudes or subjective norms but also changes in earlier behavior may cause a desired behavior (Van Raaij & Verhallen, 1983). In this chapter, "habitual behavior" refers to behavior influenced by habits.

In sum, behavioral intention and habit lead to behavior when situational constraints do not exist. Danner et al. (2008) studied the role of habit and intention in the prediction of people's future behavior. They suggest that the frequency and stability of the context of past behavior mediates the role of intention. Intention has more influence on future behavior when habits are weak with low frequency or unstable context, while it has less influence when habits are strong with high frequency and stable context. Similarly, Triandis (1979) suggested a model explaining the interaction between habit and intention in the prediction of future behavior: when a habit is stronger, the relationship between intention and behavior becomes weaker.

Energy consumption and energy usage-related behavior are highly patterned. Daily and weekly energy consumption patterns within a household—such as appliance usage, hot water usage, and thermostat settings—are quite stable over time, but energy consumption patterns of households are often different from one another (Lutzenhiser, 1993). Some energy consumption occurs under conscious control, while others are associated with habitual or unconscious activities (e.g. habitual water usage patterns or keeping the lights on). The micro-behavioral research explained that

significant differences in energy consumption can be derived from patterned behavior, including conscious vs. habitual activities, and routine vs. extraordinary activities. The differences are influenced by the interactions of buildings, equipment, and actors (Lutzenhiser, 1993). Residential energy consumption can be classified with regard to occupant behavior, building, and events as follows (Bernard, McBride, Desmond, & Collings, 1988; Lutzenhiser, 1993):

- ***Habitual consumption:*** It is caused by a routine of conscious and unconscious management.
- ***Structural consumption:*** It happens when the building is unoccupied.
- ***Daily variation consumption:*** It results from unusual events such as vacations, parties, holidays, visitors, sick children, or broken windows.

2.2.2. Energy, Building Technologies and Occupants

In the previous subsection, behavior is examined regarding internal factors that influence the formation of behavior, especially focusing on habitual occupant behavior. In this subsection, occupant behavior is explained with other external factors, including energy factors and building technologies. In order to understand energy usage-related behavior in residential buildings, overall factors affecting energy, physical building properties, and occupants should be examined and the relationship between energy, building technology, and occupant behavior should be understood (Figure 2-3).

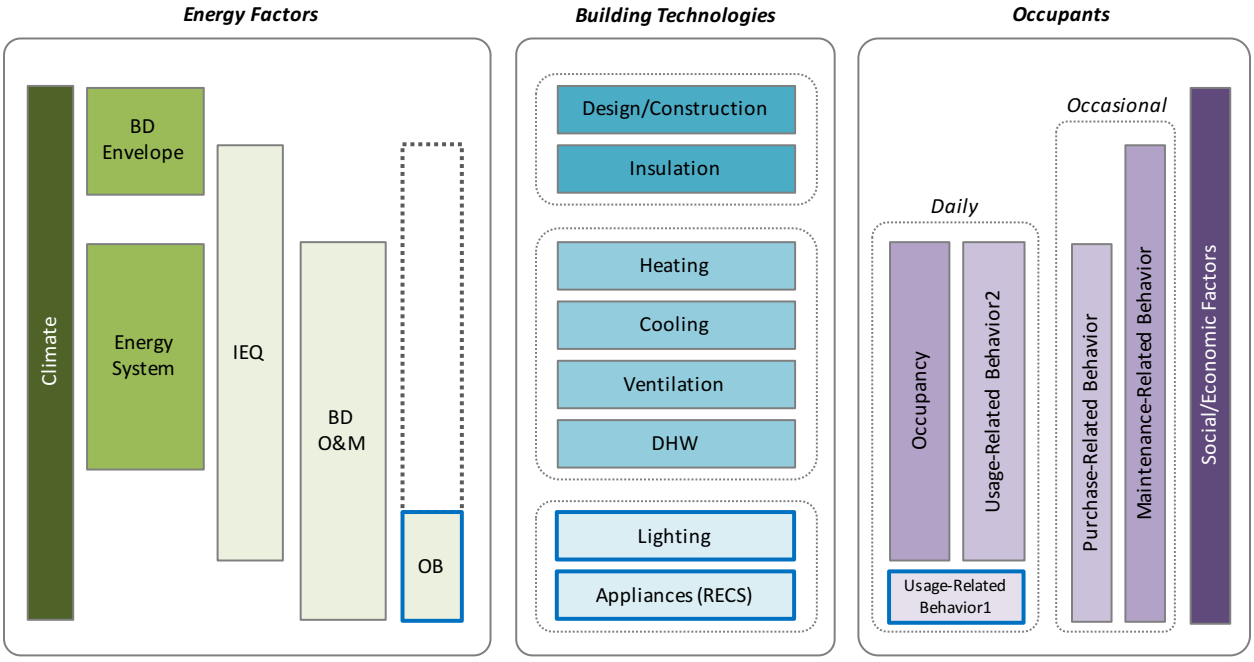


Figure 2-3. Subcategories of Energy-Tech-Occupants

Energy Factors

Building energy consumption is mainly influenced by six factors (Hong, Taylor-Lange, D'Oca, Yan, & Corgnati, 2016; Yoshino, Hong, & Nord, 2017; Yu et al., 2011): (1) climate, (2) building envelope, (3) building services and energy systems, (4) building operation and maintenance, (5) indoor environmental quality (IEQ) provided, and (6) occupant activities and behaviors. The former three are external factors and the latter three are behavior-related factors.

- ***Climate:*** The climate of the region and weather, such as outdoor air temperature, solar radiation, wind velocity, etc.
- ***Building envelope:*** The physical characteristics of the building, including orientation, building type, shape, area, insulation, windows, materials, etc.
- ***Building services and energy system:*** This includes building services and physical characteristics of energy systems, such as space cooling/heating, hot water supply, etc.

- ***Building operation and maintenance:*** This includes building operation hours, week/weekend usage schedule, etc. In residential buildings, the usage pattern of HVAC (heating, ventilation, air-conditioning), lighting, and appliances are included in this category.
- ***Indoor environmental quality:*** This includes indoor air quality, thermal and visual comfort, occupants' satisfaction with indoor conditions, etc.
- ***Occupant activities and behavior:*** This includes user-related characteristics, social and economic factors, occupants' activities in the building, and energy usage-related behaviors.

Building Technologies

Zhao et al. (2017) defined the main categories of current green building technology based on IECC 2009 (ICC, 2009): (1) Design/Construction, (2) Heating/Cooling, (3) Hot Water, (4) Ventilation, (5) Insulation, and (6) Lighting/Appliances. Each category is further specified for residential houses as follows.

- ***Design and Construction****:* The physical condition of the building. Main parameters affecting energy efficiency are the size of the house, number of bedrooms, house type, and foundation type.
- ***Heating and Cooling****:* The main energy consumption in residential buildings. Important parameters are heat pump fuel, heating seasonal performance factor, and seasonal energy efficiency ratio (SEER).
- ***Water****:* Domestic hot water consumes a significant amount of fuel, and main parameters are water heater type, water heater energy factor, and water heater tank size. Also, the amount of water usage is related to weather and occupants' behavior.

- ***Ventilation**:* Ventilation is another critical category of HVAC (Heating, Ventilation, and Air-Conditioning) systems. Important factors include duct leakage, ventilation system type, and ventilation system air flow.
- ***Insulation**:* Insulation is highly correlated with energy consumption for heating and cooling. Main factors are R-value, U-value, solar heat gain coefficient (SHGC), and infiltration rate.
- ***Lights and Appliances**:* Energy efficient light bulbs and appliances contribute energy saving in the residential sector. Main factors are the energy consumption of interior lighting, exterior lighting, refrigerators, dishwashers, ranges and ovens, clothes dryers, and ceiling fans.

Occupants

Yu et al. (2011) described occupants of buildings as their (1) user-related characteristics, (2) social and economic factors, and (3) occupants' activities in the building and behavior about energy usage.

- ***User-related characteristics:*** This includes number of occupants, occupancy (user presence in a building), etc.
- ***Social and economic factors:*** This includes age, gender, job, degree of education, energy cost, etc.
- ***Occupant behavior and activities:*** This includes what occupants do in the buildings, energy use behavior, and activities regarding temperature settings, appliance purchases, energy usage, etc. According to the American Psychological Association (APA) Dictionary of Psychology, behavior is "*an organism's activities in response to external or internal stimuli, including objectively observable activities, introspectively observable activities*

(see covert behavior), and nonconscious processes" (APA, 2018). In this chapter, "activities" refers simply to "*objectively observable activities*".

Occupants' activities and behavior can be further specified. Van Raaij and Verhallen (1983) categorized energy usage-related behaviors as (1) purchase, (2) maintenance, and (3) usage-related behaviors.

- ***Purchase-related behavior:*** The process of purchasing Heating, Ventilating, Air-Conditioning (HVAC) equipment, household appliances, and energy-using products. It includes the consideration of the energy attribute of the appliances regarding energy efficiency in daily use.
- ***Maintenance-related behavior:*** The behavior to maintain HVAC system and appliances, including repairs, home improvements, and servicing.
- ***Usage-related behavior:*** The daily energy consumption of household appliances (usage-related behavior 1 in Figure 2-3), lighting, and HVAC systems in the home (usage-related behavior 2 in Figure 2-3) regarding frequency, duration, and intensity of the energy use. It includes the energy-conscious behavior of setting the set-point temperature of thermostats, using ventilation systems. This usage-related behavior is more directly related to habits and behavioral patterns, which are generally more difficult to change.

All of the factors about energy, building technology, and occupants are illustrated in Figure 2-3. In this figure, the heights of the bars indicate the relationships between the factors. For example, occupant behavior (OB) in energy factors are mainly related to lighting and appliances in building technologies, but in the long term, it can be related to insulation, heating, cooling, ventilation, or

domestic hot water (DHW) in building technologies. Occupant factors are more detailed in the right part of the figure. The latter part of this research will focus more on the highlighted parts of the figure, lighting and appliances of building technologies, and occupants' usage-related behavior in daily life.

2.3. Behavior Prediction Model

2.3.1. Occupant Behavior Prediction Model

The Occupant Behavior Prediction Model aims to predict occupant behavior through energy consumption data. In addition, this model can identify habitual and non-habitual behaviors, which can potentially be used for efficient building operation/control strategies, interventions/education, and so on. Unlike previous models that focused on predicting energy consumption by occupant behavior, or changing occupant behavior through intervention or education, this model investigates the reverse: predicting occupant behavior based on energy consumption.

The Occupant Behavior Prediction Model incorporates the function of habit on the formation of behavior, which is innovative in residential energy and occupant behavior studies. Existing studies (Ouellette & Wood, 1998; W. Wood, Tam, & Witt, 2005) suggested that the strength of a habit should be measured by reflecting its frequency and stability of its context. They estimated the strength of habits by multiplying a measure of past behavior frequency with a measure of context stability. This provided a habit scale, where a higher score indicates a strong habit with high frequency in a stable context, and lower score indicates a weak or nonexistent habit with low frequency in an unstable context. Given that the contexts remain relatively stable, past choice of behavior can have more influence on later choice of behavior (C.-F. Chen & Chao, 2011). Wood

et al. (2002) defined habits as behaviors that are performed repeatedly in stable contexts, because context stability is important for automatic responding.

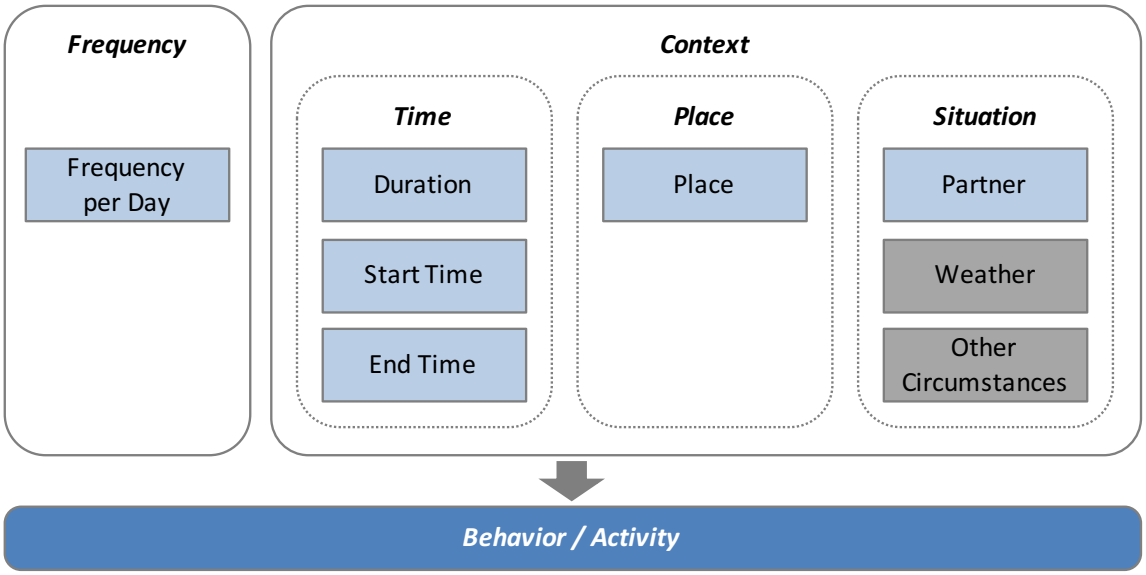


Figure 2-4. Occupant Behavior/Activity Prediction Model

The components of this Occupant Behavior/Activity Prediction Model are extracted from those habitual behavior studies and used to measure the strength of habit in occupant behavior. Behaviors and activities are explained with the following main components (Figure 2-4).

- ***Frequency:*** Number of times a single activity s performed per day
- ***Context:*** Context is broken down into Time, Place, and Situation
    - ***Time***
        - ***Duration:*** Total minutes of an activity, from the start time to the end time
        - ***Start Time:*** Start time (HH:MM) of an activity
        - ***End Time:*** End time (HH:MM) of an activity
    - ***Place (Where):*** Physical location where an activity is performed
    - ***Situation***

- ***Partner (Who):*** Person/people with whom an activity is performed
- ***Weather:*** Weather conditions when an activity is performed
- ***Other Circumstances:*** Other circumstances affecting an activity

In this study, *Frequency, Duration, Start Time, End Time, Place (Where), Partner (Who)* are mainly used as input features of machine learning algorithms, which are then used predict occupants' behaviors and activities and identify the predictability of each activity.

2.3.2. Main Components of the Model

In order to predict a person's behavior, we must first determine how predictable and habitual that behavior is. This section will examine the main component used to predict behavior.

Habits and intentions jointly predict future actions, and strong habits are difficult to change with intentions. New intentions must be sufficiently strong to override stable habits. Continuous control is required until the new behavior is more strongly settled than existing habits. If the new behavior is not as well established as the existing habits, because the behavior is new and not performed frequently enough, or because the context of the behavior is unstable or difficult, behavior is more like to be influenced by intentions, conscious and controlled processes (Ouellette & Wood, 1998).

The relationship between the existing habits, the new behavior, and intentions implies that education or intervention on behavior intend to influence intentions, and by doing so, change the behavior. However, education or intervention might be less effective on strong habits. Thus, after identifying which habits are strong or weak, researchers and stakeholders can set more effective

strategies to change behavior by focusing on the weak habits, which have more potential to be easily changed. In contrast, a different approach is required to deal with strong habits. Energy control systems need to understand the patterns behind occupants' strong habits and set effective control strategies following those patterns, rather than trying to change the behavior directly.

Effective interventions to change weak habits tend to involve stimulus control (i.e., limitation of exposure to stimulus cues), and response substitution (i.e., linkage of a competing response to the cues). In addition, effective interventions to change intentional action tend to give new information that changes the value of behavioral outcomes. (W. Wood et al., 2005).

Habits are constructed when one behavior is frequently and consistently repeated for the same purpose in similar contexts (Danner et al., 2008; Ouellette & Wood, 1998). Habits are signs of the cognitive and motivational changes caused by repeated behavior. With repetition, the practical action is associated with the times, locations, and other features of that context, and these associations form habitual actions which are automatically triggered by those features (W. Wood et al., 2002; W. Wood et al., 2005).

Frequency

Frequency of past behavior plays a significant role in the prediction of future behavior, over and above intention (Ajzen, 1991; Ouellette & Wood, 1998), which means that those behaviors are performed without much thought and deliberation (Danner et al., 2008). The impact of the frequency of past behavior on future behavior emphasizes how heavily behavior is influenced by habit (Danner et al., 2008).

Context: Time, Place, Situation

Although frequency plays a significant role in forming habits and predicting future behavior, it is not the sole factor needed to form habits. Another important factor is the consistency of the behavior (Danner et al., 2008; Ouellette & Wood, 1998; W. Wood et al., 2005). The consistency denotes the stability of the context in which the behavior has happened in the past. The stability of the context contributes to habit formation based on the assumption that people tend to be sensitive to changes in a given context. The context includes place, time, and situation. The time is the time of a day, the place is the physical location, and the situation includes circumstances such as other people and weather (Danner et al., 2008). Kahneman et al. (2004) explained that the situation more focused on interaction partners. They asked structured questions about respondents' daily activities: what they were doing (activities), when they started and ended (time), where they happened (place), and whom they were with (interaction partner). A context is considered stable when the time, place, and situation (partner) in which the behavior is performed are always similar (Danner et al., 2008).

Aarts et al. (1997) explained that habits are supposed to be developed when a behavior is frequently performed at the same time, in the same place, in the same situation. If a behavior is performed very frequently, but it is always performed in different contexts (time, place, situation/partner), the behavior will be more dependent on intentions and will not be established as habit. Similarly, if a behavior is always executed in the same context, but it only occurs occasionally, it will again be more determined by intentions rather than from being a stable habit (Danner et al., 2008).

2.4. Case Study: Extract Attributes from the ATUS to Fit the Model

2.4.1. Overview of the ATUS Data

The American Time Use Survey (ATUS) is an annual national survey conducted by the U.S. Bureau of Labor Statistics (U.S. BLS) (Kahneman et al., 2004). The U.S. BLS conducts the national survey on how the population allocates time in their daily lives. The ATUS assesses what (activity), where (place), and with whom (partner) a nationally representative sample of Americans spends their time in a regular day. The survey has been annually conducted since 2003, and it contains detailed daily activities from more than 10,000 respondents per year (U.S.BLS, 2018). The diary of the activities starts from 4 AM to 4 AM of the next day.

In this study, the ATUS 2015 data are used to examine energy usage-related behavior, focusing on habitual consumption among habitual, structural, and daily variation consumptions (see the end of subsection 2.2.1). The survey results are recorded in the following seven basic data files (U.S.BLS, 2018).

- ***Respondent file:*** Contains data about respondents, including their workforce status and earnings.
- ***Roster file:*** Contains data about household members and non-household children of the respondents, including age and sex.
- ***Activity file:*** Contains data about how the respondents spent a day, including activity codes, locations, and start/end times.
- ***Activity summary file:*** Contains data about the total time each respondent spent on each activity during the day.
- ***Who file:*** Contains data about who was with the respondent during each activity.

- ***Eldercare roster file:*** Contains data about elderly people whom the respondents take care of, including duration of care, age, and sex.
- ***Current population survey (CPS) file:*** Contains data about all individual household members who were selected to take part in the survey. These data were collected 2-5 months ahead of the actual ATUS interview.

Table 2-1. ATUS 1st Tier Activities

Code	Activity
01	Personal care
02	Household activities
03	Caring for and helping household members
04	Caring for and helping non-household members
05	Work and work related activities
06	Education
07	Consumer purchases
08	Professional and personal care services
09	Household services
10	Government services and civic obligations
11	Eating and drinking
12	Socializing, relaxing, and leisure
13	Sports, exercise, and recreation
14	Religious and spiritual activities
15	Volunteer activities
16	Telephone calls
18	Traveling
50	Data codes

The main data for this research are extracted from the activity file, and other supporting information is extracted from the who, respondent, roster, and CPS files. The activities are defined in three tiers: the first tier has 18 overall categories of activities (Table 2-1), the second tier has more detailed 110 subcategories under the first tier, and the third tier has the most detailed 465 categories under the first and second tiers.

2.4.2. Reclassification of Activities

Most of the existing studies using the ATUS data analyzed the activities in the 1st tier level (Aksanli et al., 2016; Diao et al., 2017). However, the 1st tier activity categories are too broad to explain energy usage-related behaviors. In order to understand residential energy behaviors more accurately, this study uses the 3rd tier categories.

Table 2-2. Energy Usage-Related Activities (3rd Tier)

New Code	3rd Tier Code	Activity
AA01	010201	Washing, dressing, and grooming oneself
BB01	020101	Interior cleaning
BB02	020102	Laundry
BB03	020201	Food and drink preparation
BB04	020203	Kitchen and food clean-up
BB05	020303	Heating and cooling
BB06	020501	Lawn, garden, and houseplant care
	020502	Ponds, pools, and hot tubs
BB07	020601	Care for animals and pets (not veterinary care)
BB08	020701	Vehicle repair and maintenance (by self)
CD01	030101	Physical care for household children
	040101	Physical care for non-household children
CD02	030401	Physical care for household adults
	030501	Helping household adults
	040401	Physical care for non-household adults
EF01	050101	Work, main job
	050102	Work, other job(s)
	060301	Research/homework for class for degree, certification, or licensure
LL01	120303	Television and movies (not religious)
	120304	Television (religious)
LL02	120305	Listening to the radio
	120306	Listening to/playing music (not radio)
LL03	020904	Household & personal e-mail and messages
	050401	Job search activities
	120307	Playing games
	120308	Computer use for leisure (exc. Games)
	150101	Computer use

Since the ATUS data record all of the respondents' activities on the diary day, the dataset contains both energy usage-related and non-energy-usage-related activities. Among the 465 activities, 27 activities with the potential to use electricity, gas, or water were selected by examining their

descriptions. The selected activities were re-grouped based on their similarity, the energy types and appliances that they could possibly use. Table 2-2 shows the new codes for the modified groups of activities, the original 3rd tier activity codes from the ATUS, and the descriptions of the activities. The 3rd tier code shows the hierarchy of the activities: the first 2 digits indicate the 1st tier activity groups, the middle 2 digits indicate the 2nd tier activity groups, and the last 2 digits indicate the 3rd tier activity groups. Table 2-3 explains the new code of activities and the energy types and appliances for the activities.

Table 2-3. New Activity Code, Energy, Appliances

Code	Activity	Energy	Appliances (Electricity and Gas)
AA01	Washing, dressing, and grooming	E,W,G	Lighting, Shower, Hair dryer, Shaving
BB01	Interior cleaning	E	Lighting, Vacuum
BB02	Laundry	E,W,G	Lighting, Washer, Dryer
BB03	Food and drink preparation	E,W,G	Lighting, Oven, Stove, Toaster, Blender, Coffee machine, Cooker, etc.
BB04	Kitchen and food clean-up	E,W	Lighting, Dish washer
BB05	Heating and cooling	E,G	Lighting, HVAC
BB06	Gardening, ponds, pools, and hot tubs	W,G,E	Lighting
BB07	Care for animals and pets	E,W	Lighting
BB08	Vehicle repair and maintenance	E	Lighting, Repair tools
CD01	Physical care for children	E,W	Lighting
CD02	Physical care for/helping adults	E,W	Lighting
EF01	Work for job(s)/research/homework	E	Lighting, Computer
LL01	Television	E	Lighting, TV
LL02	Listening to/playing radio or music	E	Lighting, Computer, Music player, Radio
LL03	General computer use	E	Lighting, Computer

** *E: Electricity, W: Water, G: Gas*

In the ATUS, *Heating and cooling* (new code BB05, 3rd tier code 020303) activity does not mean operating HVAC systems or setting set-point temperature of a thermostat, but means "collecting/chopping woods, lighting fireplace, shoveling coal, filling heater with fuel, installing fireplace etc." (U.S.BLS, 2018), which is less usual in households. The specific meaning of *Heating and cooling* activity of the ATUS should be considered in the latter part of data analysis in this chapter.

2.4.3. Variables (Attributes)

Based on the Occupant Behavior Prediction Model defined in Section 2.3.2, *Frequency* and Context (*Time, Place, Partner*) variables were extracted from the ATUS files as follows.

- ***Activity:*** New group of activities
- ***Frequency:*** Number of times the activity was recorded during the day
- ***Start Time:*** Start time of the activity in minutes
- ***End Time:*** End time of the activity in minutes
- ***Duration:*** Minutes spent doing the activity from start time to end time
- ***Place:*** Place where the activity was performed
- ***Partner:*** Partner with whom the activity was performed

Table 2-4. Partner Code

Code	Partner	ATUS Code	Detailed Partner
1	Alone	18	Alone
		19	Alone
2	Household	20	Spouse
		21	Unmarried partner
		22	Own household child
		23	Grandchild
		24	Parent
		25	Brother/sister
		26	Other related person
		27	Foster child
		28	Housemate/roommate
		29	Roomer/boarder
3	Non-Household	30	Other nonrelative
	(Friends, Acquaintances)	40	Own non-household child < 18
		51	Parents (not living in household)
		52	Other non-household family members < 18
		53	Other non-household family members 18 and older (including parents-in-law)
		54	Friends
		56	Neighbors/acquaintances
		57	Other non-household children < 18
		58	Other non-household adults 18 and older
4	Work-Related	59	Boss or manager
		60	People whom I supervise
		61	Co-workers
		62	Customers

Activity is a dependent variable and others are independent variables. An activity has 15 unique values, as explained in Table 2-3. *Frequency, Start Time, End Time*, and *Duration* are numeric variables and *Partner* and *Place* are categorical variables, explained in Table 2-4 and Table 2-5. The ATUS defined the Partner with 25 categories, but this was simplified to Alone, Household, Non-Household, and Work-Related people in this research.

Table 2-5. Place Code

Code	Place	Code	Place
1	Respondent's home or yard	14	Walking
2	Respondent's workplace	15	Bus
3	Someone else's home	16	Subway/train
4	Restaurant or bar	17	Bicycle
5	Place of worship	18	Boat/ferry
6	Grocery store	19	Taxi/limousine service
7	Other store/mall	20	Airplane
8	School	21	Other mode of transportation
9	Outdoors away from home	30	Bank
10	Library	31	Gym/health club
11	Other place	32	Post Office
12	Car, truck, or motorcycle (driver)	89	Unspecified place
13	Car, truck, or motorcycle (passenger)	99	Unspecified mode of transportation

2.5. Case Study: ML Classification Process

This research used a machine learning (ML) approach to understand energy usage-related behavior based on the behavior prediction model using the ATUS data. The model is used to predict energy usage-related activities and to identify the predictability and the habitual characteristic of each activity. For the data analysis, various packages in Python and R are used. The process is explained as follows (Figure 2-5).

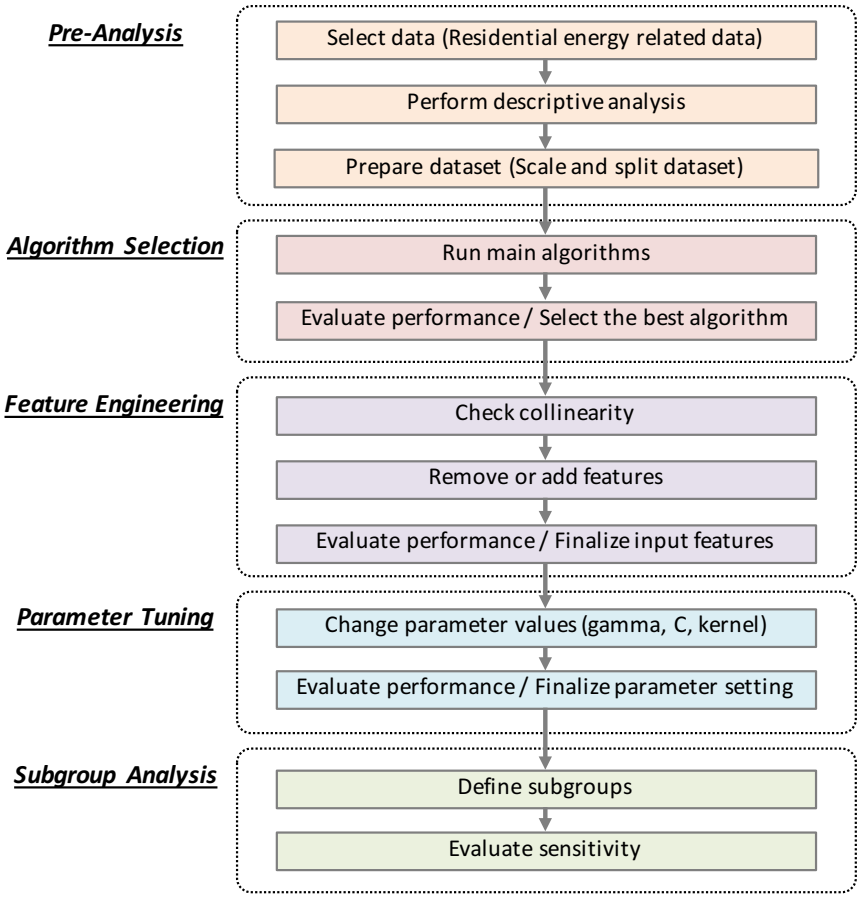


Figure 2-5. Data Analysis Process

2.5.1. Pre-Analysis

The goal of the pre-analysis is to understand the characteristics of the overall dataset and the data distribution of each variable. This step includes the following tasks: (1) select data based on the given conditions, (2) perform s descriptive data analysis on the main variables, and (3) split datasets for machine learning processes.

2.5.2. Algorithm Selection

Machine learning algorithms show different performances depending on the characteristics of the given dataset. Thus, multiple machine learning algorithms are compared in this step and the

algorithm with the best performance is selected for further improvement. Algorithms that are frequently used for classification include Naïve Bayes (NB), Logistic Regression (LR), K-Nearest Neighbor (KNN), Decision Tree (DT), and Support Vector Machine (SVM) (Mayfield & Rose, 2010; Shermis & Burstein, 2013).

NB is a probabilistic classifier that is built on the Bayes' theorem that assumes independence between attributes. NB is often employed as a baseline algorithm due to its easy and fast implementation. NB performs well with plenty of fairly weak predictors and efficiently extends to classification tasks with multiple class values (Mayfield, Adamson, & Rose, 2014).

LR is a conditional probability model that builds a linear model by reducing incorrect probability values based on a transformed target variable. LR generates accurate probability estimates by maximizing the probability of the training data (Witten, Frank, Hall, & Pal, 2016).

KNN is an instance-based classification. Nearest-neighbor classification compares each new instance with existing ones using a distance metric, and a class is assigned to the new instance using the closest existing instance. K neighbors use more than one nearest neighbor for the categorical class or the distance-weighted average for the numeric class. KNN is simple and often works efficiently, and each attribute has the same effect on the decision. However, it is easily influenced by noisy data (Witten et al., 2016).

DT is a divide-and-conquer approach that compares the value of some attribute with a constant and divides the data at a node. Nodes in a decision tree test a particular attribute, and the test

compares an attribute value with a constant. A DT constructs the comparisons recursively. First, it selects an attribute, places it at the root node, and creates a branch for each possible value. Then it splits the dataset into subsets and repeats the process recursively for each branch until all instances at a node have the same class value (Witten et al., 2016).

SVM is based on the maximum-margin hyperplane, an algorithm used to find a special type of linear model (Witten et al., 2016). SVM adapts linear models to investigate nonlinear class boundaries with a focus on marginal instances.

Two different versions of the dataset are compared: one where the numeric variables are used without standardization, and a second where the numeric variables are standardized to a mean of 0 and a standard deviation of 1. The performance of the algorithms is evaluated with Accuracy, Precision, Recall, and F1-score.

Accuracy, Precision, Recall, and F1-score are used to evaluate the predictive performance of the algorithms. Accuracy is the percentage of correct predictions, or the ratio of true predictions to the total number of instances. Precision and Recall are the indexes of relevance. Precision is the ratio of correct positive predictions to all positive predictions. A low precision implies a large number of false positives. Recall is the ratio of correct positive predictions to the sum of correct positive predictions and wrong negative predictions. A low recall implies a large number of false negatives. F1-score is the harmonic mean of precision and recall. Their functions are described in Figure 2-6 and Equations 1 through 4.

		Predicted	
		Positive	Negative
Actual	Positive	True Positive	False Negative
	Negative	False Positive	True Negative

Figure 2-6. Confusion Matrix

$$Accuracy = \frac{True\ Positive + True\ Negative}{Total} \quad \text{(Equation 1)}$$

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \quad \text{(Equation 2)}$$

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \quad \text{(Equation 3)}$$

$$F1\ Score = \frac{2\ Precision \times Recall}{Precision + Recall} \quad \text{(Equation 4)}$$

2.5.3. Feature Engineering

The goal of feature engineering is to identify the best combination of features to achieve a higher performance. First, problems among the existing features are examined, such as collinearity or noisy features. In order to diagnose the collinearity of the variables, correlations between variables are checked. Then, additional features are considered. All of the possible combinations of additional features are applied to the model, and the performance is evaluated. Finally, the highest-performing combination of features is selected for the next step of analysis.

2.5.4. Parameter Tuning

The goal of parameter tuning is to further refine the performance of the selected algorithm. Most machine learning algorithms' performance varies by parameter setting. The performance of SVM,

in particular, is heavily influenced by its parameters, such as kernel, C, gamma, etc. Several different values of the parameters are tested and the values with the best performance are selected.

2.5.5. Subgroup Analysis

Subgroup analysis finds patterns in a subset of the dataset, and it is useful to assess whether different types of subsets respond differently to the model (Lagakos, 2006). In this study, subgroups are defined by (1) quantile of feature values, (2) predictive performance of each activity, and (3) number of instances of activities.

- ***(1) Quantile:*** Each feature has outliers, and subgroup analysis is performed to evaluate the influence of outliers and the various range of quantiles. For each feature, quantiles are calculated and subgroups are set with different instances having certain ranges of middle values, such as 95%, 90%, 80%, and 50% of middle ranges.
- ***(2) Performance:*** Activities with low performance might include noisy data, which have a negative impact on the overall model performance. In this subgroup analysis, the model's performance is compared between high- and low-performance activities.
- ***(3) Number of Instances:*** Activities with a high number of instances are more common among respondents and activities with a lower number of instances are less common. In order to examine the influence of an activity's frequency on the model's performance, subgroups are defined based on the number of activity instances.

2.6. Case Study: Result

2.6.1. Pre-Analysis

2.6.1.1. Features (Variables) and Instances

Based on the Occupant Behavior Prediction Model, *Frequency, Duration, Start Time, End Time,* and *Partner* variables are initially selected. The *Place* variable is excluded in the baseline algorithm selection, since it only contains the values of Home (1) and Not Collected (-1).

Originally, the activity file from the 2015 ATUS contained 214,429 activities from 10,905 respondents. For this study, only energy usage-related activities were selected, so 76,980 activities from 10,849 respondents remained. Since this study focuses on residential energy behaviors, those activities were narrowed down to only include the ones that happened in the respondent's home or yard. The ATUS does not collect the location and partner information for certain types of activities, such as sleeping and grooming, due to privacy concerns. Therefore, it was assumed that those activities happened alone at home (Diao et al., 2017). This left 67,115 activities from 10,772 respondents, which this study used for the analysis. 70 percent of the whole dataset was set as the training set and the remaining 30 percent was set as the testing set.

2.6.1.2. Descriptive Analysis

Descriptive analysis helps researchers understand the overall properties of a dataset and provides ways to help analyze the given data more efficiently. In this section, the distribution of categorical data values and the range of numeric data values are examined.

Table 2-6 summarizes the data distribution of a dependent variable, *Activity*, and a categorical variable, *Partner*. *Watching television* (LL01), *Washing, dressing, and grooming* (AA01), and *Food and drink preparation* (BB03) have the highest numbers, which means they are the most common and frequent energy usage-related activities in daily life. *Physical care for/helping adults* (CD02), *Vehicle repair and maintenance* (BB08), and *Heating and cooling* (BB05) have the lowest numbers among the energy usage-related activities in the ATUS data. In the ATUS, heating and cooling activity includes preparation of fuels (such as collecting/chopping/stacking wood, shoveling coals, or filling a heater with fuel), and installing and maintaining heating and cooling systems (such as installing a fireplace or window air-conditioning unit, or changing a furnace filter), which are less common activities in households with gas or electricity-based heating and cooling systems. This may be the reason why the number of heating and cooling activities is very low in this dataset.

Table 2-6. Distribution of Partner for Each Activity

Code	Total	Not Collected(-1)		Alone(1)		Household(2)		Non-Household(3)		Work-Related(4)	
AA01	15266	15266	(100%)	0	(0%)	0	(0%)	0	(0%)	0	(0%)
BB01	3535	3	(0%)	2486	(70%)	975	(28%)	71	(2%)	0	(0%)
BB02	2952	0	(0%)	2419	(82%)	483	(16%)	50	(2%)	0	(0%)
BB03	9986	7	(0%)	6191	(62%)	3450	(35%)	336	(3%)	2	(0%)
BB04	3360	1	(0%)	2231	(66%)	1035	(31%)	93	(3%)	0	(0%)
BB05	111	0	(0%)	85	(77%)	23	(21%)	3	(3%)	0	(0%)
BB06	1345	0	(0%)	1016	(76%)	290	(22%)	39	(3%)	0	(0%)
BB07	2167	2	(0%)	1858	(86%)	279	(13%)	26	(1%)	2	(0%)
BB08	228	0	(0%)	158	(69%)	55	(24%)	15	(7%)	0	(0%)
CD01	5063	0	(0%)	87	(2%)	4819	(95%)	157	(3%)	0	(0%)
CD02	253	0	(0%)	11	(4%)	225	(89%)	17	(7%)	0	(0%)
EF01	2422	0	(0%)	1844	(76%)	459	(19%)	61	(3%)	58	(2%)
LL01	16334	7	(0%)	8750	(54%)	6940	(42%)	637	(4%)	0	(0%)
LL02	371	0	(0%)	295	(80%)	63	(17%)	13	(4%)	0	(0%)
LL03	3722	1	(0%)	2714	(73%)	891	(24%)	116	(3%)	0	(0%)

Due to privacy issues, the ATUS does not collect *Partner* information for some activities, including *Washing, dressing, and grooming* (AA01). Except for *Caring for children and adults* (CD01, CD02), most of the activities are done by oneself. *Care for animals and pets* (BB07),

Laundry (BB02), and *Listening to/playing radio or music* (BB02) are the activities most likely to be performed alone. Following *Caring for children and adults* (CD01, CD02), the next most likely activities to be performed with household members are *Watching television* (LL01), *Food and drink preparation* (BB03), *Kitchen and food clean-up* (BB04), and *Interior cleaning* (BB01). This shows that people tend to watch television, have meals, and do household chores with family members. *Physical care for/helping adults* (CD02) and *Vehicle repair and maintenance* (BB08) are relatively likely to be performed with non-household members, which implies that these activities need more help from other experts. The only activity that was likely to be performed with work-related people (2%) was *Work for job(s)/research/homework* (EF01).

Since only the activities that were performed at home are selected, the *Place* variable has only "home," except for *Washing, dressing, and grooming* (AA01), when location information was not collected due to privacy concerns.

Figure 2-7 shows the range of *Frequency* values for each activity. For most of the activities, the number of an activity performed in a day is between 1-3 times. *Physical care for children* (CD01) has the highest value of *Frequency* and it also has the highest-value outlier. *Physical care for/helping adults* (CD02) and *Work for job(s)/research/homework* (EF01) show higher values than other activities. This implies that caring for others (especially children) happens more frequently because they (children or other adults) need help often. Also, when respondents report their *Work for job(s)/research/homework* (EF01) activity at home, the number of instances of the working activity in a day is relatively high.

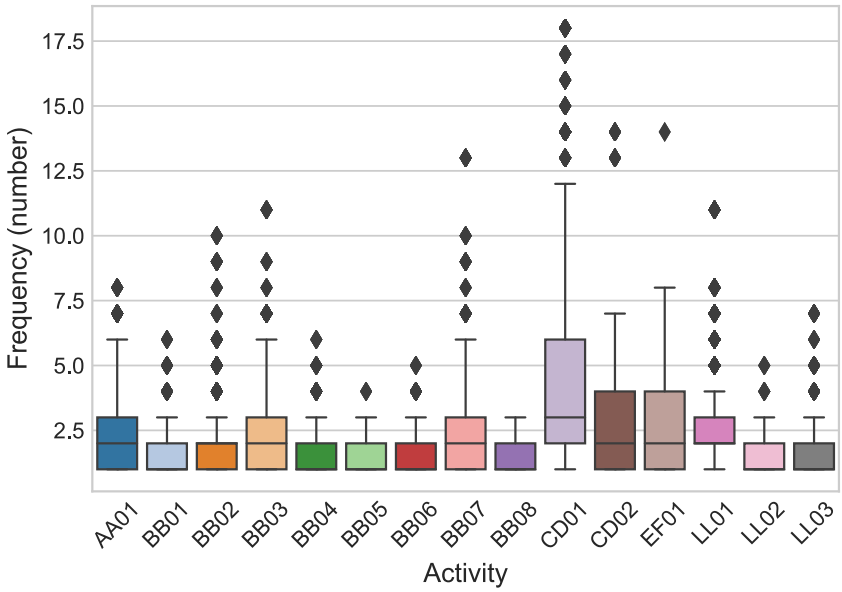


Figure 2-7. Range of Frequency for Each Activity

Figure 2-8 illustrates the range of *Duration* per one instance of an activity for each activity type. The ranges of *Duration* values vary by activity, which suggests that the *Duration* variable can have distinctive power to predict activities. The respondents spend longer times *Watching television* (LL01) and *Working for job(s)/research/homework* (EF01), and spend shorter times on *Care for animals and pets* (BB07), *Care for children and adults* (CD01, CD02), *Kitchen and food clean-up* (BB04), and *Heating and cooling* (BB05) in a single instance of the activity. *Watching television* has the highest-value outlier (1400 minutes). When explaining with *Frequency* together, the data show that if the respondents watch television (LL01), most of them watch television between 2-3 times a day, and spend approximately 60-144 minutes each time. If they work at home for jobs/research/homework (EF01), most of them work between 1-4 times a day and spend about 30-150 minutes each time. The activity of watching television shows has a relatively low frequency per day, but once people start watching TV, they spend longer times on it compared to other activities. Similarly, if the respondents take care of children or adults (CD01, CD02), most of them take care of children 2-6 times and adults 1-4 times a day, and spend 10-30 minutes each time. The

respondents with children to take care of do physical care for children (CD01) most often among the given activities, but they spend relatively short amounts of time on each instance.

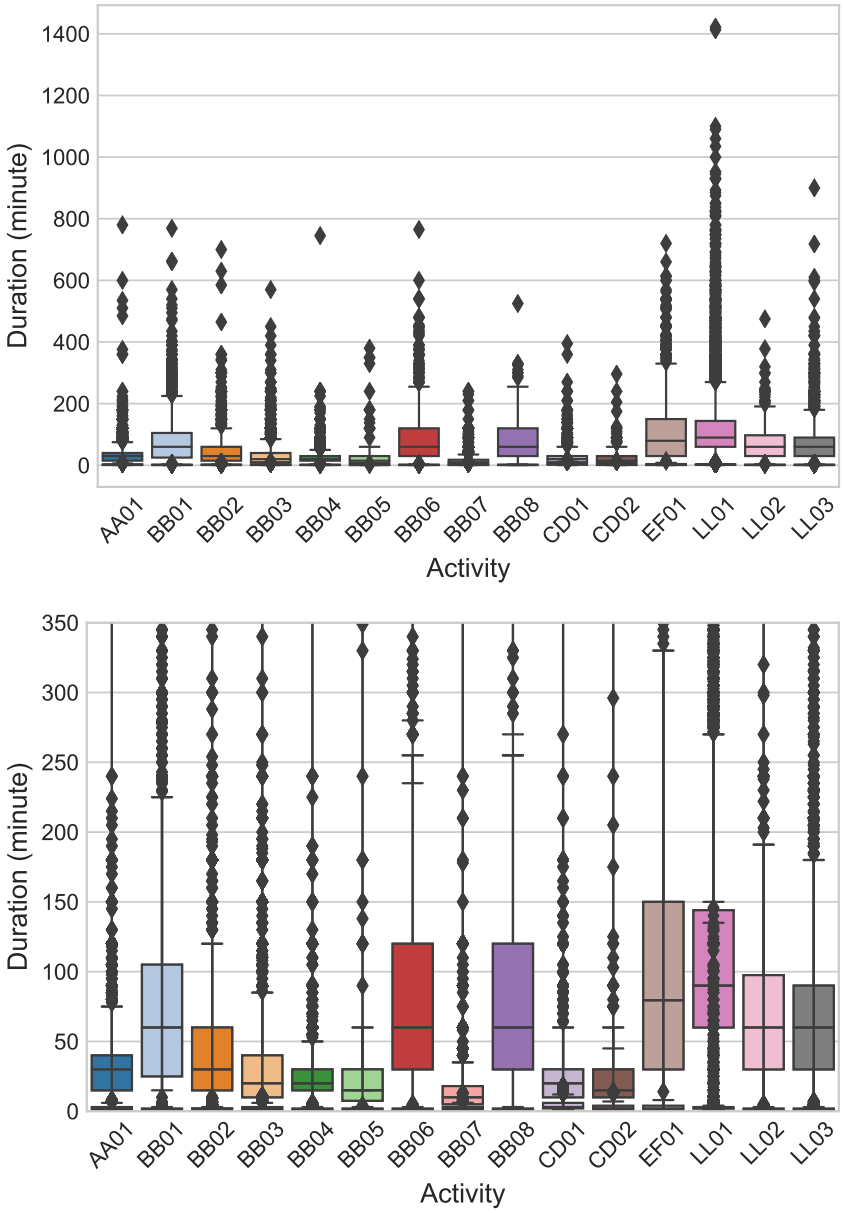


Figure 2-8. Range of Duration for Each Activity (a: top, b: bottom)

Figure 2-9 and Figure 2-10 illustrate the range of *Start Time* and *End Time* for each activity, and the values are varied by activity. Most of the respondents start *Washing, dressing, and grooming* (AA01), *Care for animals and pets* (BB07), *Physical care for children and adults* (CD01, CD02),

and *Food and drink preparation* (BB03) earlier than other activities (before 8 AM) in the morning. In contrast, most of them start watching television (LL01) and listening to/playing radio or music (LL02) in the afternoon (after 12 PM), and the end times of these activities are later than that of other activities.

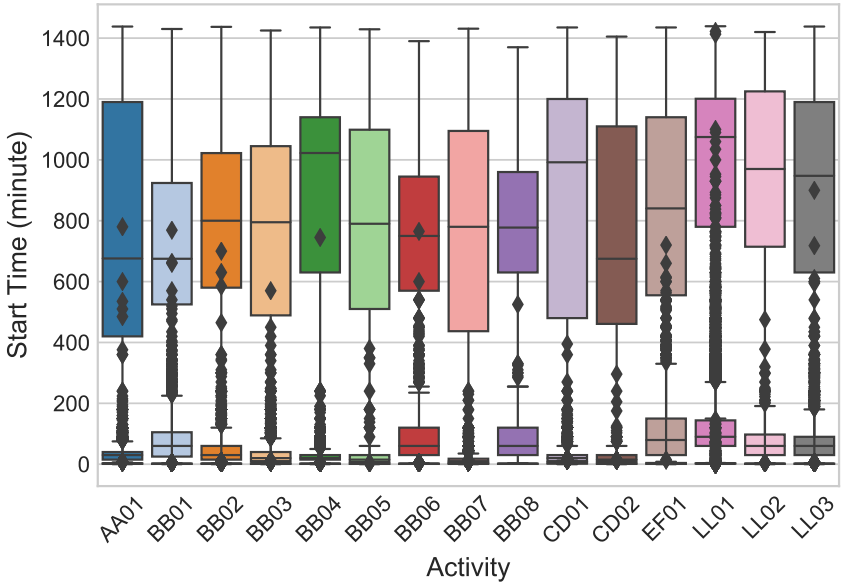


Figure 2-9. Range of Start Time for Each Activity

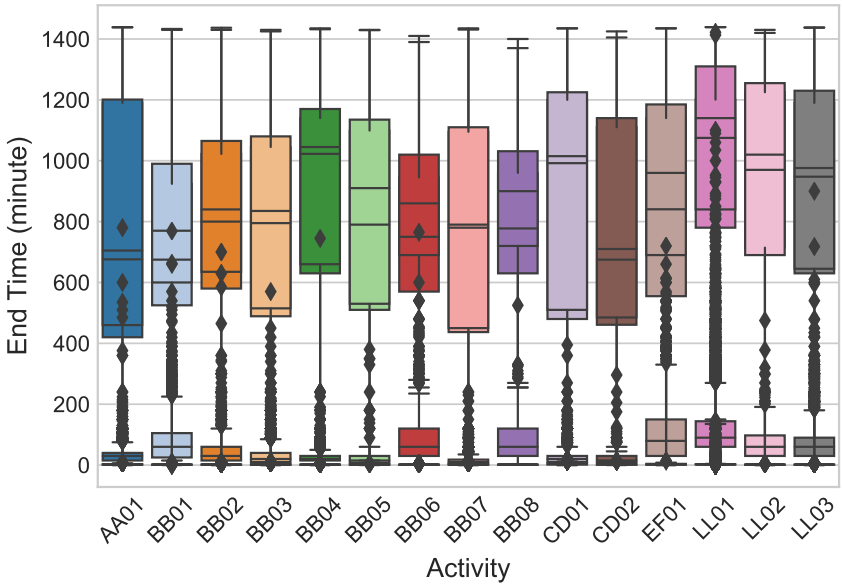


Figure 2-10. Range of End Time for Each Activity

2.6.2. Algorithm Selection

Table 2-7 lists the results of the algorithm selection for the prediction of energy usage-related activities. Firstly, Naïve Bayes (NB), Logistic Regression (LR), K-Nearest Neighbor (KNN), Decision Tree (DT), and Support Vector Machine (SVM) were tested with the original data. Then, the same algorithms were run with standardized data. When using the original non-standardized features, LR showed the best performance (Accuracy 0.57). However, SVM improved significantly with standardized features: Accuracy improved from 0.53 to 0.61, which is better than LR performed with non-standardized features.

Table 2-7. Performance of Different Algorithms

Algorithm	No-Standardization				Standardization			
	Accuracy	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score
NB	0.56	0.53	0.56	0.53	0.56	0.53	0.56	0.53
LR	0.57	0.46	0.57	0.49	0.57	0.46	0.57	0.49
KNN	0.51	0.47	0.51	0.48	0.57	0.55	0.57	0.55
DT	0.53	0.53	0.53	0.53	0.53	0.53	0.53	0.53
SVM	0.53	0.48	0.53	0.47	0.61	0.58	0.61	0.55

Among the features, *Partner* is a categorical variable, and *Frequency, Duration, Start Time,* and *End Time* are numeric variables. As examined in the previous descriptive analysis, the value ranges are very different among the numeric variables, which can affect the performance of the machine learning algorithms. For example, many algorithms (such as the radial basis function (RBF) kernel of SVM) assume that the all input variables/features have means of 0 and variances in the same order of magnitude. Thus, if one feature has much larger variance than the others, it might have too heavy an influence on the objective function and weaken the estimating power from other features as expected (Scikit-Learn, 2017). This explains the big performance improvement of SVM with standardized features because the RBF kernel is used in this run. In the following steps, SVM with standardized features is further developed to improve its predictive performance.

2.6.3. Feature Engineering

In the baseline algorithm selection, *Frequency, Duration, Start Time, End Time,* and *Partner* features were used. Since collinearity can cause biased estimates or insignificant estimates that were considered to be important (Belsley, Kuh, & Welsch, 2005), collinearity among the existing features are tested before adding more features, which potentially improves the model's performance.

In order to diagnose the collinearity, Pearson correlations between features are calculated as summarized in Table 2-8. *Start Time* and *End Time* have high correlation, 0.86. Although the baseline algorithm selection did not include the *Place* variable, it is included when checking collinearity, and the result indicates very high correlation (0.9) between the *Partner* and *Place* variables.

Table 2-8. Pearson Correlations between Features

	Frequency	Duration	Start Time	End Time	Partner	Place
Frequency	1.00	-0.08	-0.02	-0.02	0.09	0.04
Duration	-0.08	1.00	0.02	0.11	0.19	0.22
Start Time	-0.02	0.02	1.00	0.86	0.14	0.11
End Time	-0.02	0.11	**0.86**	1.00	0.16	0.13
Partner	0.09	0.19	0.14	0.16	1.00	0.90
Place	0.04	0.22	0.11	0.13	**0.90**	1.00

** All correlations are significant at the 0.01 level (2-tailed)*

Johnson et al. (2014) studied the probability of transitioning from a current activity to the next activity using a Markov Chain behavior model, which shows that the previous activity is related to the next activity. Thus, previous activities are also considered additional features in this section.

Based on the correlations between features and past studies about the influence of previous activities on the next activity, input features are adjusted with the following options: (1) baseline

features are *Frequency, Duration, Start Time, End Time*, and *Partner*, (2) exclude *End Time* from baseline, (3) add *Place* to baseline, (4) add previous activity *(A-1)* to baseline, (5) add the 2 steps previous activity *(A-2)* to baseline, (6) add *Place* and *A-1* to baseline, (7) add *Place* and *A-2* to baseline, (8) add *A-1* and *A-2* to baseline, and (9) add *Place, A-1,* and *A-2*to baseline. The performances are compared in Table 2-9.

Table 2-9. Performance of Additional Features

Features	Accuracy	Precision	Recall	F1-score
(1) Baseline	0.61	0.58	0.61	0.55
(2) – End Time	0.61	0.57	0.61	0.55
(3) + Place	0.61	0.58	0.61	0.55
(4) + A-1	0.64	0.61	0.64	0.60
(5) + A-2	0.62	0.60	0.62	0.58
(6) + Place, A-1	0.64	0.60	0.64	0.59
(7) + Place, A-2	0.62	0.60	0.62	0.58
(8) + A-1, A-2	0.59	0.58	0.59	0.55
(9) + Place, A-1, A-2	0.60	0.58	0.60	0.56

Despite excluding the *End Time* feature (option 2), the performance kept almost same with the baseline features. When adding the *Place* feature (option 3), it did not improve the performance as was expected given the previous collinearity test. However, when adding the previous activity, *A-1,* as a feature (option 4), the performance improved. Adding the 2 steps previous activity, *A-2,* (option 5) slightly improved performance compared to the baseline. Adding only *A-1* (option 4) showed the highest performance among all other options of new features. Some options, such as adding *A-1* and *A-2* together (option 8) and adding *Place, A-1,* and *A-2* all together (option 9) showed even lower performance than the baseline. Based on this feature engineering, the *Frequency, Duration, Start Time, Partner,* and *A-1* features are used for the next step.

2.6.4. Parameter Tuning

The performance of SVM is sensitive to parameter settings. Based on the previous studies about parameter tunings (Dong, Cao, & Lee, 2005; Friedrichs & Igel, 2005; C.-L. Huang & Wang, 2006), the gamma and C values of the RBF kernel are tested in this section. The performance is compared by changing the values of gamma and C. In the previous steps, the default parameters of support vector classification (SVC) from Python Scikit-Learn package were used with the RBF kernel, C as 1, and gamma as auto, which is automatically calculated as 1/number of features. Since the SVM had five parameters, gamma was set as 0.2.

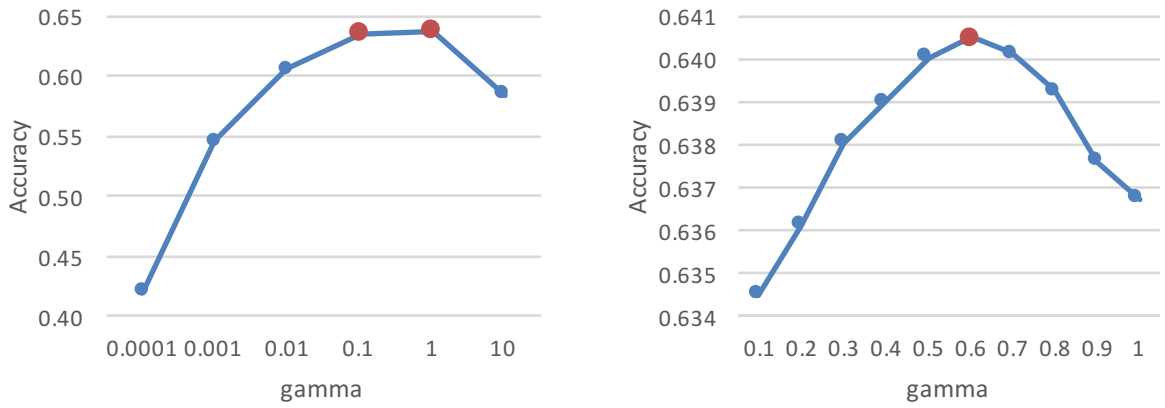


Figure 2-11. Accuracy by Different gamma Values (a: left, b: right)

First, C is fixed as 1 (default value), and only gamma values are changed as 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, and 10^{1}. As shown in Figure 2-11a, the accuracy is highest (0.6368) when gamma is 1 (10^{0}), which is slightly higher than the accuracy (0.6361) with the default gamma value (0.2). In Figure 2-11b, gamma values are more finely tested between 0.1 and 1, and the accuracy (0.6405) is highest when gamma is 0.6.

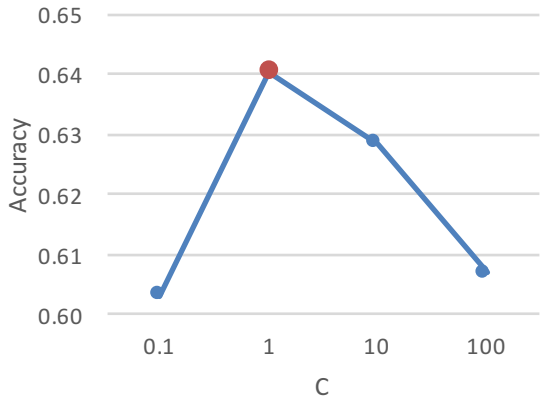
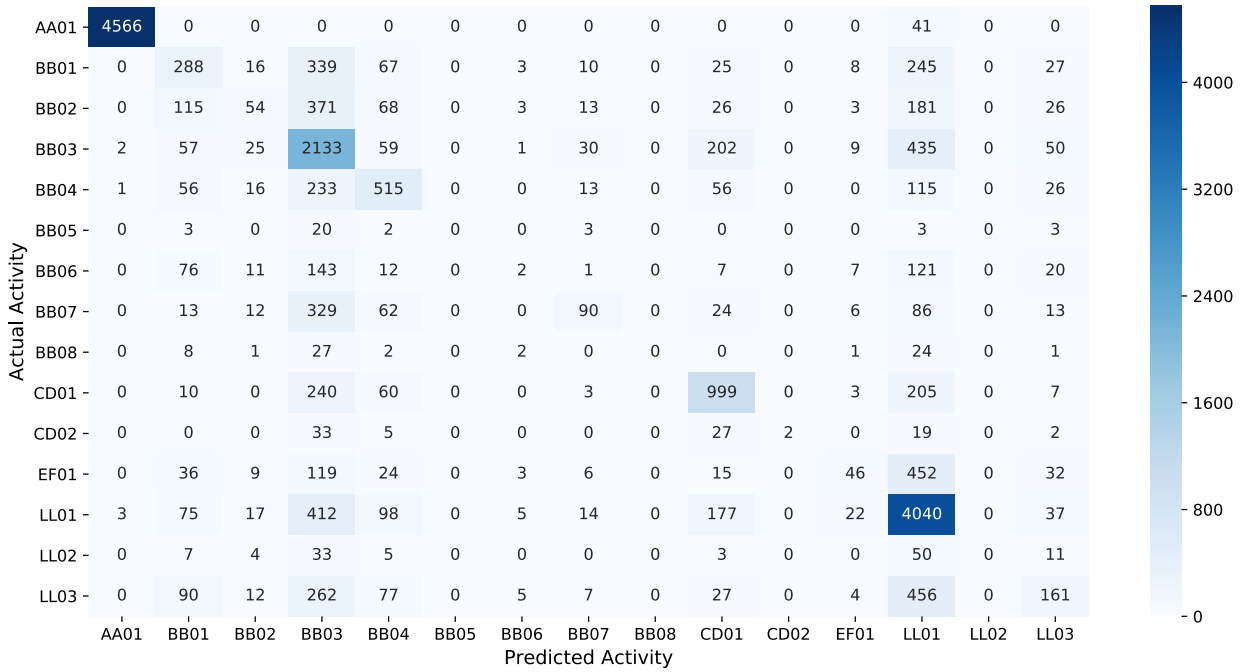


Figure 2-12. Accuracy by Different C Values

Next, gamma value is fixed as 0.6, and C values are changed to 10^{-1}, 10^{0}, 10^{1}, and 10^{2}. As shown in Figure 2-12, the accuracy is highest (0.6405) when C value is 1 (10^{0}). With C value 1, and gamma value 0.6, the sigmoid, linear, and polynomial kernels are tried, but the sigmoid kernel shows very low accuracy (0.2434), and linear and polynomial kernels are much more computationally expensive than the RBF kernel. Therefore, the RBF kernel with C value 1 and gamma value 0.6 is used for the final SVM model. The final performance has Accuracy 0.6405 which is slightly higher than the accuracy of the baseline SVM (0.6382), and Precision 0.61, Recall 0.64, and F1-score 0.60, which are almost the same as the baseline. Since the default parameter settings were already compatible with the given dataset, the final parameter tuning result showed similar performance to the baseline.

Figure 2-13 and Figure 2-14 display the confusion matrix of actual activity and predicted activity. In the matrix, cells along the right downward diagonal line represent the correct predictions, while cells out of the diagonal line represent incorrect predictions. Actual activities are on the Y-axis and predicted activities are on the X-axis. The sum of each row in Figure 2-13 is the total number of each activity in the testing set, and similarly, the sum of each row in Figure 2-14 is 1 (100%)

for each activity. For example, 2133 instances (71%) of BB03 are correctly predicted as BB03, but 435 instances (14%) of BB03 are incorrectly predicted as LL01.



** *Numbers are from the testing set, which is 30% of the whole dataset*

Figure 2-13. Confusion Matrix: Comparison of Actual vs. Predicted Numbers

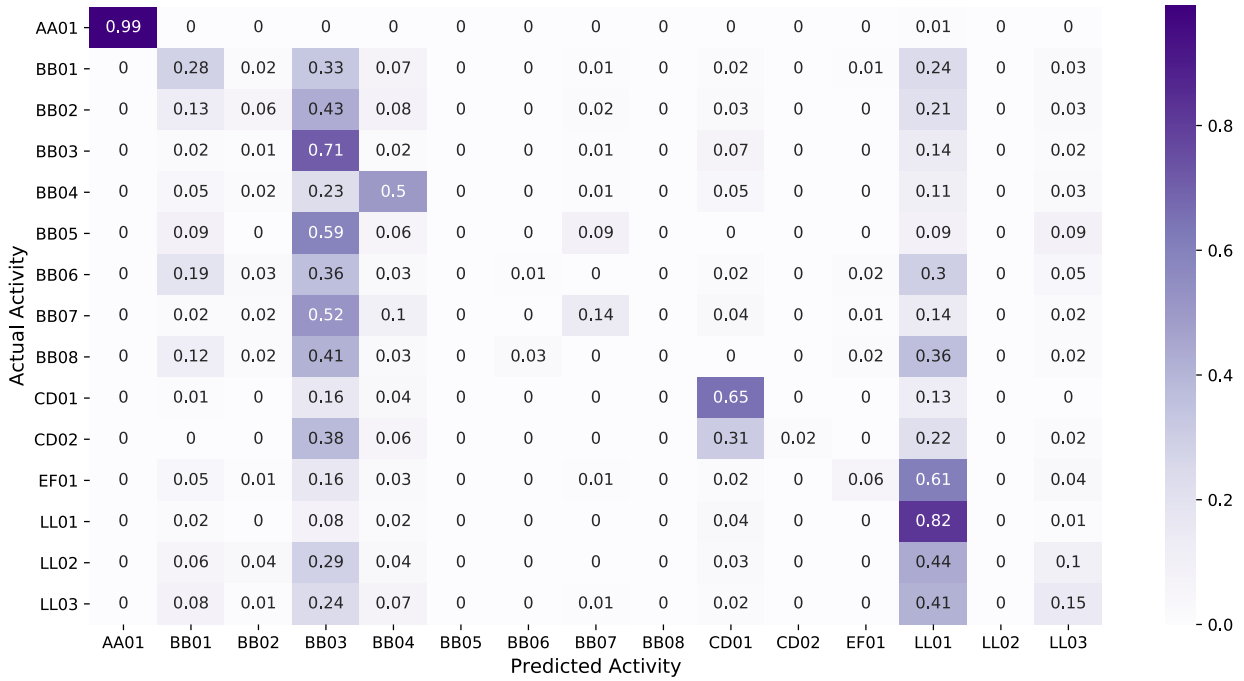


Figure 2-14. Confusion Matrix: Comparison of Actual vs. Predicted Accuracy

Confusion matrices can be used to identify problematic cells, where the ML algorithm has less power to distinguish the differences between classes. Table 2-10 summarizes some problematic cells with error numbers higher than 400 or error rates higher than 50%. Actual EF01 (*Work for job(s)/research/homework*) being predicted as LL01 (*Watching Television*) is the most problematic since both its error number is 452 (higher than 400) and its error rate is 61% (higher than 50%).

Table 2-10. Problematic Cells in Confusion Matrix

Actual	Predicted	Error Number	Error Rate
BB05	BB03	20	0.59
BB07	BB03	329	0.52
LL01	BB03	412	0.08
BB03	LL01	435	0.14
EF01	**LL01**	**452**	**0.61**
LL03	LL01	456	0.41

Table 2-11 compares the mean values of the numeric attributes and the ratio of the categorical attributes of EF01 and LL01. The mean values of *Frequency, Duration, Start Time,* and *End Time* are similar, and the mode value of *Partner* is *Alone (1),* although the ratios of the categories are different. It can be inferred that the main factors of the ML algorithm are the numeric attributes and the algorithm has less power of classification when most of the attribute values are similar. In order to improve overall accuracy by lowering the error numbers and error rates of the problematic cells, additional feature engineering can be considered, such as adding additional features or variable transformations, in the future research.

Table 2-11. Descriptive Analysis for Problematic Cells

Code	Mean				Ratio
	Frequency	Duration	Start Time	End Time	Partner
EF01	2.63	109.67	846.36	909.65	76:19:3
LL01	2.59	114.23	975.05	1024.48	54:42:4

** *Partner Ratio is Alone(1) : Household(2) : Non-Household(3)*

Based on the confusion matrix, the performance of the model for each activity is calculated as summarized in Table 2-12. *Washing, dressing, and grooming* (AA01) shows the highest accuracy

(0.99), which means this model predicts 99% of AA01 activity correctly. The model predicts *Watching television* (LL01), *Physical care for children* (CD01), and *Food and drink preparation* (BB03) with higher performance. However, the model incorrectly predicts *Heating and cooling* (BB05), *Vehicle repair and maintenance* (BB08), and *Listening to/playing radio or music* (LL02). The number of instances of an activity is also relevant to the model's predictive performance for each activity, since the model can be trained better with more data.

Table 2-12. Predictive Performance of Each Activity

Code	Accuracy	Precision	Recall	F1-score	Count	Description
AA01	0.99	1.00	0.99	0.99	4607	Washing, dressing, and grooming oneself
LL01	0.82	0.62	0.82	0.71	4900	Watching TV
CD01	0.65	0.63	0.65	0.64	1527	Physical care for children
BB03	0.71	0.45	0.71	0.55	3003	Food and drink preparation
BB04	0.50	0.49	0.50	0.49	1031	Kitchen and food clean-up
BB01	0.28	0.35	0.28	0.31	1028	Interior cleaning
BB07	0.14	0.47	0.14	0.22	635	Care for animals and pets
LL03	0.15	0.39	0.15	0.21	1101	General computer use
EF01	0.06	0.42	0.06	0.11	742	Work for job(s)/research/homework
BB02	0.06	0.31	0.06	0.10	860	Laundry
CD02	0.02	1.00	0.02	0.04	88	Physical care for/helping adults
BB06	0.01	0.08	0.01	0.01	400	Gardening, ponds, pools, and hot tubs
BB05	0.00	0.00	0.00	0.00	34	Heating and cooling
BB08	0.00	0.00	0.00	0.00	66	Vehicle repair and maintenance (by self)
LL02	0.00	0.00	0.00	0.00	113	Listening to/playing radio or music

*** Counts are from the testing set, which is 30% of the whole dataset*

2.6.5. Subgroup Analysis

To further examine the sensitivity of the model, its performance is tested with different subgroups, which are defined as follows.

- ***Quantile Subgroups:*** Quantiles of 2.5%, 5%, 10%, 25%, 75%, 90%, 95%, and 97.5% are calculated and subgroups are set with instances having certain ranges of middle values.
  - **95%:** Instances with all features between the 2.5% and 97.5% range are selected.
  - **90%:** Instances with all features between the 5% and 95% range are selected.
  - **80%:** Instances with all features between the 10% and 90% range are selected.

- **50%:** Instances with all features between the 25% and 75% range are selected.

- ***Performance Subgroups:*** Subgroups are set with activities showing high performance and activities showing low performance.
  - **High:** Activities with high performance (accuracy above 0.49) are included in this group (AA01, LL01, BB03, CD01, BB04).
  - **Low:** More activities with lower performance are added into the High group (High + LL03, BB07, BB02, EF01, CD02). Activities with almost 0 Accuracy are excluded (BB06, BB05, BB08, LL02).
- ***Number of Instances Subgroups:*** Subgroups are set with activities having high numbers of instances and activities having lower numbers of instances.
  - **Major:** Activities with high numbers of instances are included in this group (LL01, AA01, BB03, CD01, LL03, BB01, BB04, BB02).
  - **Minor:** Additional activities with lower number of instances are included in the Major group (Major + EF01, BB07, BB06). LL02, CD02, BB08, and BB05, which represent less than 1% of the total activity instances, are excluded.

Table 2-13 summarizes the results of the subgroup analysis.

- ***Quantile Subgroups:*** The overall accuracy of SVM after the parameter tuning with the whole dataset was 0.6405, and the performance of all quantile subgroups was lower than the overall performance. The result indicates that the model loses its distinctive power when the value ranges of the instances are reduced.
- ***Performance Subgroups:*** The accuracy of the high performance subgroup reaches 0.8282, which means the model can predict correctly more than 82% of the time for the activities

of washing, dressing, and grooming (AA01), watching television (LL01), physical care for children (CD01), and food and drink preparation (BB03). The accuracy of the low performance subgroup also shows higher accuracy (0.6585) than the overall accuracy (0.6405). The result shows that when the quality of the data is improved, the model reaches higher performance.

- ***Number of Instances Subgroups:*** The accuracy of the major subgroup shows 0.7047 and the minor subgroup shows 0.6438. Both are higher than the overall accuracy. These subgroups excluded activities with a very few number of instances, and the result shows that the performance of the model can be improved with more training data for each class. Also, the result suggests that the data quality of the activities with few instances might not be good in this dataset.

Table 2-13. Performance of Subgroups

Criteria	Group	Accuracy	Precision	Recall	F1-score	Count
Quantile	95%	0.6245	0.59	0.62	0.59	59040
	90%	0.6117	0.57	0.61	0.57	52715
	80%	0.5918	0.55	0.59	0.55	39677
	50%	0.5414	0.49	0.54	0.49	16061
Performance	**High**	**0.8282**	**0.83**	**0.83**	**0.83**	50009
	Low	**0.6585**	**0.63**	**0.66**	**0.62**	65060
# Instances	**Major**	**0.7047**	**0.68**	**0.70**	**0.68**	60218
	Minor	**0.6438**	**0.62**	**0.64**	**0.61**	66152

** *Counts are from the whole dataset*

The results of the subgroup analysis demonstrate that the performance of the model can be even more improved depending on the quality of the data. Its performance with the whole dataset is lower than its performance with only the subgroups with better data quality, since the whole dataset includes outliers and generally poorer quality of data. Depending on the data quality, the model could reach 83% accuracy, and it can be further improved with other datasets of better quality.

2.7. Discussion

The Occupant Behavior Prediction Model can predict occupant behavior with overall 64% accuracy for the ATUS dataset, and its accuracy can reach up to 83% for a subgroup of habitual activities. Notably, the model shows 99% accuracy for predicting washing, dressing, and grooming activity and 82% accuracy for predicting watching television activity. The multi-class classification problems are challenging, and achieving high accuracy in these compared to binary classification problems is difficult (Farid, Zhang, Rahman, Hossain, & Strachan, 2014; Guyon & Elisseeff, 2003). The Occupant Behavior Prediction Model is applied for multi-class classification with 15 classes (15 activities). The result demonstrates high performance pertaining to multi-class classification, especially considering that the probability of correct predictions with simple statistical calculation is 6.7%.

This model can identify more-habitual activities and less-habitual activities based on the prediction performance of each activity. The model was tested on the ATUS data to predict activities of the general occupants from nationally representative samples. From the results, people tend to wash, dress, and groom (AA01) as more predictable routines, and watch television in a predictable pattern. They take care of children (CD01) frequently when the children are in need of their care and help. Food and drink preparation (BB03) and kitchen and food clean up (BB04) are habitual and predictive behaviors. Interior cleaning (BB01), laundry (BB02), care for adults (CD02) or pets (BB07), general computer use (LL03), and working at home (EF01) are less predictive, meaning less habitual behavior. Heating fuel preparation (BB05), vehicle maintenance (BB08), and listening to radio/music or playing music (LL02) are very difficult to predict, and therefore they are non-habitual behaviors.

There exist some limitations to this study. The ATUS collects diary data for only one specific day from a respondent and does not ensure that it is a typical day for the respondent. Although this shortcoming is compensated for by the large number of samples collected, another study using occupants' daily records of multiple days is suggested to identify more precise and specific patterns of occupants' behavior. Also, while the ATUS records one activity at a time, multiple activities can happen concurrently in reality. For example, people may do laundry while watching television. Thus, the complexity of the activities should be considered when applying this model to another dataset.

The Occupant Behavior Prediction Model innovatively incorporated the concept of habit to predict occupant behaviors and identify habitual/non-habitual activities, while previous studies about occupant behaviors have tended to focus more on socioeconomic attributes to predict energy consumption. This novel approach explores the past habitual characteristics of the households, predicts their future behaviors, and identifies their habitual behaviors. Habitual behaviors are more difficult to change, but they are easier to predict. For these activities and behaviors, energy systems need to find efficient control strategies that are suitable for these behaviors rather than trying to change the behaviors. In contrast, less habitual behaviors, which are difficult to predict, might be easier to change, and education or intervention might be more effective on these activities. The result can be used to develop more improved occupant schedules and to set specific energy control strategies. Also, the results can be used to develop effective intervention or education for residential occupants. This model will be further applied to examine the geographical patterns of activities (horizontal analysis), and the timely patterns of activities (vertical analysis) in the following chapters.

CHAPTER 3

DAILY BEHAVIOR PATTERN AND FACTORS AFFECTING OCCUPANT BEHAVIOR IN RESIDENTIAL BUILDINGS

Abstract

Residential occupants have a high degree of energy control, unlike commercial building occupants, which implies that residential energy consumption is significantly influenced by the energy usage-related behaviors of the occupants. This study aims to strategically identify the daily routines of habitual behaviors and activities in residential buildings with diverse methods including clustering, comparative analysis, and Geographical Information System (GIS) using the American Time Use Survey (ATUS) data. The patterns of occupant energy usage-related activities are identified using K-modes clustering, and the activities are compared by different perspectives including region, day of the week, gender, and job status. The main energy usage-related activities are analyzed with GIS at the state level. The findings include (1) day of the week, gender, job status affect the similarities and differences in energy usage-related activities, (2) watching TV is one of the most common activities in cluster analysis, and it happens between 18:30 and 21:30. The results can be used to provide more realistic information regarding energy and behavior to the occupants in residential buildings, and it can be applied to new energy and behavior strategies and policies for residential building energy plans.

3.1. Introduction

Residential occupants have significant influences on and control over energy consumption. Residential building energy consumption is affected by climate, physical properties of the building, building services and energy systems, appliances in the household, occupant activities and behavior, and the interactions among them (Widén & Wäckelgård, 2010). As the quality of thermal properties is improved and the technology for energy efficient appliances grows more advanced, the energy consumption associated with buildings' physical properties and appliances is decreasing. For example, the U.S. Department of Energy (DOE) reports that most recently built houses are 14 percent more energy efficient than houses built 30 years ago, and 40 percent more energy efficient than houses built 60 years ago (U.S.DOE, 2015; Zhao et al., 2017). Also, building design standards and requirements are becoming stricter with regard to energy efficiency of

buildings and appliances. However, overall building energy consumption has not decreased (Chen et al., 2015). This energy consumption can be explained by the influence of occupant behavior and living style, and it emphasizes the role of occupant behavior in residential energy savings.

Residential energy consumption can be significantly reduced by changing the energy usage-related behaviors of the occupants. Unlike commercial building occupants, residential occupants have a high degree of energy control. They can control heating, ventilation, and air conditioning (HVAC) systems, lighting and electronic devices, and kitchen and laundry appliances, are the main causes of energy consumption in residential buildings (Li & Jiang, 2006).

In recent years, the relationship between occupant behavior and residential energy consumption has been studied more actively. However, the existing studies have been less focused on the occupants' habitual activities and daily routines.

The goal of this study is to identify the habitual daily routines of occupant behavior in different groups of occupants, and find out the factors that influence the similarity and differences of energy usage-related activities. To achieve this goal and to solve existing problems, this study (1) defines an occupant behavior prediction model that includes the concept of habit to help understand occupant behavior in a more realistic way by considering past behavior patterns, (2) uses detailed levels of activities for occupant behavior and activity analysis, (3) analyzes the pattern of occupant habitual energy usage-related behavior using the U.S. national behavior data separated out by diverse context, such as by region, day of the week, gender, and job status, and (4) uses GIS to identify if geographical location affects the characteristics of an activity. The identified habitual

daily routine of occupant behavior and the factors affecting energy usage-related activities can be used to set more realistic occupant schedules for energy control strategies or energy simulation in residential buildings.

The structure of this paper is as follows. First, the background section explains the Occupant Behavior Prediction Model, the use of the American Time Use Survey (ATUS) dataset which records participants' daily activity logs in a national survey, and the use of Geographic Information System (GIS) as a research method to geographically explain and analyze datasets. Then, the methodologies for clustering analysis, comparative analysis, and GIS techniques are described, and finally, the results are explained and discussed.

3.2. Background

3.2.1. Occupant Behavior Prediction Model

Behavior refers to occupants' activities, including introspectively observable activities, objectively observable activities, and non-conscious processes responding to internal or external stimuli (APA, 2018). In this research, "behaviors" refers to broader activities as explained, and "activities" refers to a narrower range of objectively observable activities. In addition, "habitual behavior" refers to a behavior influenced by habits.

The Occupant Behavior Prediction Model aims to predict occupant behaviors and identify habitual activities. The model incorporates the concept of habit with the components of activity frequency and context. Context includes time (start time, end time, duration), place, and situation (partner,

weather, other circumstances). These components are derived from habitual behavior studies to measure the strength of habit in occupant behavior.

The predicted occupant activities and the identified habitual and non-habitual activities can be used for efficient building operation and control strategies, more effective interventions, or education on occupant energy usage-related behaviors. In this chapter, this model will be used to identify the habitual daily routines of occupants and the factors affecting their energy usage-related activities.

3.2.2. Use of the ATUS in Occupant Behavior Studies

The American Time Use Survey (ATUS) is a survey conducted by U.S. Bureau of Labor Statistics every year. The purpose of the survey is to record the activities, locations, and demographic information of the respondents in a regular day from 4 AM to 4 AM of the next day (Diao et al., 2017). The ATUS provides (1) population measurement and (2) participant measurement. The population measurement provides the average time of an activity for a particular population. The participant measurement estimates the average time spent on an activity per day (Diao et al., 2017). While the time use surveys conducted in other countries, such as the United Kingdom and Sweden, require respondents to record their activity with 5- or 10-minute intervals, the ATUS asks participants to report the start and end times of an activity. The activities in the ATUS data are in a hierarchical tree structure with 3 tiers. The 1st tier consists of overall categories of activities, the 2nd tier consists of intermediate categories of activities, and the 3rd tier contains the most detailed activities.

The ATUS data have been used in many behavioral studies since the ATUS records detailed daily diaries for each respondent with activities, times, places, partners, and so on. In addition, the ATUS provides the respondent's socioeconomic information which supports behavior data analysis. Some examples of past analyses are as follows.

Johnson et al. (2014) presented a statistical model for the behavior of residential occupants with the ATUS data. They specified ten simplified activities from the 1st and 2nd tiers of activities, which correspond to the major energy-consuming appliances in a household. Then, they developed time-based statistical models by different occupant types (working male occupant, working female occupant, non-working male occupant, non-working female occupant, and child occupant) using the Markov chain method. The models were applied to energy simulations to show how a residential occupant affects major energy consumption during a day.

Diao et al. (2017) identified and classified occupant behavior with energy consumption outcomes. They used 8-17 activities from the 1st tier and 2nd tier ATUS activities. They derived occupant behavior patterns using K-modes clustering, and extracted occupant features from the 2009 ATUS data. These were combined with demographic-based probability neural networks (PNN) to identify 10 behavior patterns.

Aksanli et al. (2016) developed a residential energy modeling framework based on human activities to estimate the energy consumption in residential buildings. They used seven simplified activities: sleeping, personal grooming, cooking, cleaning, entertainment, working at home, and going to work, which are derived from the 1st tier activities in the ATUS. They extracted action-

and activity-related parameters from the ATUS, such as duration of each activity by different user group based on the demographic information of the occupants, such as age, gender, job status, and number of household members. These parameters were applied to a probabilistic model to capture the time-series characteristics of occupant behavior.

While most of the current studies used the 1st tier or 2nd tier activities, this study uses the 3rd tier activity list to provide more detailed and realistic behavioral analysis. Especially, all of the original 3rd tier activities are included to identify the habitual daily routines of occupants, and the main energy usage-related activities are directly selected from the 3rd tier activities in this study.

3.2.3. Use of GIS in Building/Construction Studies

Geographic Information System (GIS) has been used often in research. GIS is beneficial as a useful cognitive tool to analyze and gather spatial data with its visual interface, which can help experts from other areas understand the data easily (Fonseca & Schlueter, 2015). It allows the researchers and other stakeholders to quickly identify data patterns and outliers (Kolter & Ferreira Jr, 2011). GIS is used not only as a tool to display data on the map, but also as a method to analyze data by geographical location.

Recently, building and other construction fields have been actively employing GIS as a part of their research methods as well, since GIS can capture, store, analyze, manage, and present spatial or geographic data including not only the energy or construction related data but also the location (i.e. address, city, state) and physical properties of buildings (i.e. size, height) (Ma & Cheng, 2016). Also, city-wide GIS databases have been available in many regions of the world and accessible to

the general public (Reinhart & Davila, 2016). GIS has more potential to combine with 3D models of buildings, energy simulations, and real-time databases at a large geographical scale. Some examples are as follows.

Fonseca and Schlueter (2015) developed an integrated model for building energy consumption patterns in city districts. They used spatial analysis, dynamic building energy modeling, and energy mapping combined with GIS. The model focused on determination of the spatiotemporal variability of energy services in existing and future buildings in commercial, residential, and industrial sectors. It provided detailed assessments of potential energy efficiency measures in the city district scale.

Howard et al (2012) developed a model to estimate the energy end-use intensity (EUI) in the building sector for space heating, cooling, domestic hot water, and appliances in New York City. They assumed that energy consumption primarily depended on building functions (residential, office, educational, etc.) and not on the physical characteristics of a building (construction type, age of building, etc.). The end-use ratios were obtained from the Microdata of the Residential Energy Consumption Survey (RECS) and the Commercial Building Energy Consumption Survey (CBECS), and they estimated the energy consumption. The modeled energy usage and the percent differences between the measured and predicted consumption were calibrated by ZIP code level and displayed on a map using GIS.

Heiple and Sailor (2008) presented a technique to estimate hourly and seasonal energy consumption profiles in buildings at detailed spatial scales, tax lots, or parcels. They combined GIS framework and annual building energy simulations for city-specific prototypes. They applied

the method to Houston, TX, and the result can estimate sensible and latent wastes of heat emissions related to building energy consumption.

Kolter and Ferreira (2011) suggested a system to model end use energy consumption in residential and commercial buildings in Cambridge, MA with a data-driven approach. They combined monthly electricity and gas bill data, tax assessor records, and the GIS database containing polygonal outlines and estimated roof heights for buildings and parcels in the city. They predicted energy distributions using both parametric and non-parametric methods, and provided a system that visualized each building's energy consumption in the city using GIS.

Most of the existing building and construction studies using GIS have focused more on building energy consumption, physical building properties, or demographic information of occupants. However, this study combines GIS with the Occupant Behavior Model, and uses spatial analysis (the grouping analysis with K-means clustering) to explain the similarities and differences in energy usage-related behaviors of residential occupants by state.

3.3. Methodology

The behaviors and activity patterns in the ATUS data will be analyzed in the following parts.

- **Part 1**
  - ***Data Selection:*** In this part, all of the ATUS activity data are used. All recorded activities from all places are included – neither limited to energy usage-related activities nor limited to the activities at home.

- ***Instance:*** One row of the dataset is one respondent's daily activities in one-minute intervals during 24 hours from 4:00 AM to next day 4:00 AM.
- ***Expected Outcome:*** Identify different groups of occupants based on their daily behavioral patterns using clustering analysis.

- **Part 2**
  - ***Data Selection:*** In this part, only energy usage-related activities at home are analyzed. Also, the data format is different from the format used in Part 1.
  - ***Instance:*** One row of the dataset is one energy usage-related activity with its respondent code, region, day of the week, gender, job status, and properties of the activity including Frequency, Duration, Start Time, End Time, and Partner.
  - ***Expected Outcome:*** Identify similarities and differences in energy usage-related activities by the clusters in Part 1, region, day of the week, gender, and job status using comparative analysis.

- **Part 3**
  - ***Data Selection:*** Among the energy usage-related activities in Part 2, the five most habitual activities are selected.
  - ***Instance:*** One row of the dataset is similar to an instance in Part 2. Instances in this part include state information.
  - ***Expected Outcome:*** Identify geographical similarities and differences between selected energy usage-related activities using GIS.

In all parts, the original ordinal HH:MM format of *Start Time* and *End Time* is converted to a numeric minute format. For example, 13:10 is converted to 790 minutes. The detailed process is further explained in the following subsections.

3.3.1. Clustering of Occupant Daily Activities by Time

Clustering methods are used to identify distinctive groups of occupants based on their daily activities. The ATUS data is pre-processed for the clustering analysis, and K-modes clustering is selected based on the data type. For clustering of occupants, national level occupant data is used without segmentation by state.

3.3.1.1. Data Preparation

The original ATUS data recorded activities of an occupant as a form of sequential list with activity names (codes), places, partners, start times, and end times. In this chapter, all activities (not only limited to energy usage-related activities) from all places (not only limited to activities at home) are included. For the clustering analysis, the data are re-organized with the features (columns) of one-minute interval timestamps and the instances (rows) of occupants. It standardizes the data format of occupant activities by the same timestamps and helps the clustering algorithm identify the pattern of activities more clearly. The number of features are 1440 (1440 minutes per day), and the number of instances are 10772 (10772 respondents). The sample inputs are described in Table 3-1. All of the 465 original 3rd tier activities in the ATUS data are included in this clustering analysis with the initial 3rd tier activity codes in a numeric format. For example, sleeping is coded as 010101, grooming as 010299, cleaning as 020101, working as 050101, watching TV as 120303, and cooking as 150201.

Table 3-1. Sample Inputs for Clustering Analysis

	t1	t2	t3	…	t1438	t1439	t1440
Person 1	010101	010101	010299	…	010101	010101	010101
Person 2	150201	150201	150201	…	020101	120303	120303
…							
Person 10772	010299	050101	050101	…	010299	010101	010101

3.3.1.2. K-modes Clustering

Clustering divides a set of instances into a number of groups (clusters) so that instances in the same cluster are similar to each other and different from those in other clusters. K-means clustering is one of the most common clustering algorithms for numeric values. However, another approach is necessary for categorical data. K-modes clustering employs a simple matching dissimilarity measure which is suitable for categorical data. It uses the modes of the clusters instead of means, and updates modes in the clustering to minimize the cost function using a frequency-based method (Z. Huang, 1998). K-modes clustering uses a function minimizing cluster distance as follows (Diao et al., 2017).

$$D(X, C) = \sum_{k=1}^{K} \sum_{X_i \in C_k} d(X_i, C_k)$$

Where *D (X, C)* is the sum of within-cluster distance. $X = \{X_1, X_2, X_3, \ldots, X_n\}$ is the dataset with n instances with a vector of categorical attributes $\{A_1, A_2, A_3, \ldots, A_m\}$, and X_i is the ith instance of X. $C = \{C_1, C_2, C_3, \ldots, C_k\}$ is the centers of K different clusters, and C_k is the center of the kth cluster. *d (x, c)* is the distance function for calculating the distance between two categorical vectors.

As described in Table 3-1, the type of inputs are categorical data, and K-modes clustering is selected for the clustering of the occupant activities.

To determine a suitable number of clusters, Bayesian inference criterion (BIC) is selected. BIC is explained as follows (Jain, 2010; Kodinariya & Makwana, 2013).

$$BIC = -2 \times \ln(likelihood) + \ln(N) \times k$$

Where k is the degrees of freedom calculated as the rank of variance–covariance matrix of the parameters and N is the number of independent terms in the likelihood (Kodinariya & Makwana, 2013). The number of clusters is selected when the BIC value is the minimum in the BIC plot (Ramsey et al., 2008).

3.3.2. Comparative Analysis for Energy Usage-Related Activities

Energy usage-related activities are selected from all of the 3rd tier activities in the ATUS. The selected activities are reorganized or re-grouped, and new codes are assigned to the energy usage-related activities to prevent confusion with the original 3rd tier activities (Table 3-2). Table 3-3 explains the new codes for activities, energy types, and appliances associated with the activities.

To compare the energy usage-related activities in the ATUS data, a comparative analysis is performed. As defined by the Occupant Behavior Prediction Model (Mo, 2018)(Chapter 2), the properties of the activities are identified as *Frequency, Duration, Start Time, End Time*, and *Partner*. The activities are compared by several conditions as follows.

- ***Region:*** The activities are compared by regions, which are specified in Table 3-4, to see the geographical/regional differences.
- ***Day of the Week:*** The activities during weekdays (0) and weekends (1) are compared to find if occupants show different patterns depending on the day of the week.

- ***Gender:*** The activities are compared by male (1) and female (2) occupants.
- ***Job Status:*** The activities are compared by the occupants who have a job (Yes, 1), and who have no job (No, 0).

Table 3-2. Energy Usage-Related Activities (3rd Tier) (=Table 2-2)

New Code	3rd Tier Code	Activity
AA01	010201	Washing, dressing, and grooming oneself
BB01	020101	Interior cleaning
BB02	020102	Laundry
BB03	020201	Food and drink preparation
BB04	020203	Kitchen and food clean-up
BB05	020303	Heating and cooling
BB06	020501	Lawn, garden, and houseplant care
	020502	Ponds, pools, and hot tubs
BB07	020601	Care for animals and pets (not veterinary care)
BB08	020701	Vehicle repair and maintenance (by self)
CD01	030101	Physical care for household children
	040101	Physical care for non-household children
CD02	030401	Physical care for household adults
	030501	Helping household adults
	040401	Physical care for non-household adults
EF01	050101	Work, main job
	050102	Work, other job(s)
	060301	Research/homework for class for degree, certification, or licensure
LL01	120303	Television and movies (not religious)
	120304	Television (religious)
LL02	120305	Listening to the radio
	120306	Listening to/playing music (not radio)
LL03	020904	Household & personal e-mail and messages
	050401	Job search activities
	120307	Playing games
	120308	Computer use for leisure (exc. Games)
	150101	Computer use

Table 3-3. Activities and Associated Energy and Appliances (=Table 2-3)

Code	Activity	Energy	Appliances (Electricity and Gas)
AA01	Washing, dressing, and grooming	E,W,G	Lighting, Shower, Hair dryer, Shaving
BB01	Interior cleaning	E	Lighting, Vacuum
BB02	Laundry	E,W,G	Lighting, Washer, Dryer
BB03	Food and drink preparation	E,W,G	Lighting, Oven, Stove, Toaster, Blender, Coffee machine, Cooker, etc.
BB04	Kitchen and food clean-up	E,W	Lighting, Dish washer
BB05	Heating and cooling	E,G	Lighting, HVAC
BB06	Gardening, ponds, pools, and hot tubs	W,G,E	Lighting
BB07	Care for animals and pets	E,W	Lighting
BB08	Vehicle repair and maintenance	E	Lighting, Repair tools
CD01	Physical care for children	E,W	Lighting
CD02	Physical care for/helping adults	E,W	Lighting
EF01	Work for job(s)/research/homework	E	Lighting, Computer
LL01	Television	E	Lighting, TV
LL02	Listening to/playing radio or music	E	Lighting, Computer, Music player, Radio
LL03	General computer use	E	Lighting, Computer

Table 3-4. Census Regions

Region (Code)	States
Northeast (1)	CT, MA, ME, NH, NJ, NY, PA, RI, VT
Midwest (2)	IA, IL, IN, KS, MI, MN, MO, ND, NE, OH, SD, WI
South (3)	AL, AR, DC, DE, FL, GA, KY, LA, MD, MS, NC, OK, SC, TN, TX, VA, WV
West (4)	AK, AZ, CA, CO, HI, ID, MT, NM, NV, OR, UT, WA, WY

Mean and Mode

The mean values of the given conditions are compared for the numeric variables including *Frequency, Duration, Start Time,* and *End Time*, and the mode values are compared for the categorical variable, *Partner*.

Coefficient of Variation (CV)

The coefficient of variation (CV) is a standardized measurement of dispersion. It measures the extent of variability of a variable with numbers by eliminating the unit of measurement. CV is defined as follows (Abdi, 2010; Lovie, 2005).

$$CV = \frac{Standard\ Deviation}{Mean}$$

The CV can be used to measure the similarity of the variable values in a given condition: when the CV is smaller, the values are more similar to each other. Also, it can be used to compare distributions of the variables with different units. In this section, CV is used to evaluate the similarity of an activity in a given condition, region, day of the week, gender, and job status.

t-test and Analysis of Variance (ANOVA)

The *t-test* is a statistical hypothesis test which can be used to decide if the two group of datasets are significantly different from each other. The Analysis of Variance (*ANOVA*) provides a statistical test which generalizes the *t-test* to more than two groups, and it can be used to evaluate the statistical significance of differences among three or more group means. Among the compared conditions in this study, the independent samples *t-test* is used to determine the group differences for day of the week (weekday or weekend), gender (male or female), and job status (Yes, No) conditions, which each have two groups, and the *ANOVA* is used for region, which has more than two groups.

3.3.3. GIS Analysis for Habitual Energy Usage-Related Activities

Geographical visualization helps us understand the results of data analysis more easily and clearly, and geographical analysis considers the geographical distribution of the data. In this study, ArcGIS 10.1 by ESRI is used to compare and analyze the habitual activities by state, and to identify if geographical location affects the characteristics of the activity.

Habitual energy usage-related activities are further analyzed to compare their characteristics by different geographical locations (in this section, state level). The components of the Occupant Behavior Prediction Model are integrated into the GIS analysis. First, the mean or mode values of each activity are compared by state using GIS map visualization. Then, geographical similarities and differences of activities are identified using GIS Grouping analysis. In the previous study (Mo, 2018) (Chapter 2), the predictability of each energy usage-related activity was evaluated and the activities with higher predictability were regarded as more habitual activities. The top five most predictable and habitual activities in Table 3-5 are selected for further analysis. Their *Frequency, Duration, Start Time, End Time,* and *Partner* values are compared by state using GIS.

Table 3-5. Main Habitual Energy Usage-Related Activities

Code	Description
AA01	Washing, dressing, and grooming oneself
LL01	Watching TV
CD01	Physical care for children
BB03	Food and drink preparation
BB04	Kitchen and food clean-up

3.3.3.1. GIS Visualization: Comparison of Activities by States

The mean values of the numeric variables (*Frequency, Duration per act, Sum duration of an activity per day, Start Time,* and *End Time*), and the mode value of the categorical variable (*Partner*) of each activity are calculated for each state. Among the numeric variables, *Duration per act* is the duration of single occurrence of an activity, and *Sum duration of an activity per day* is the total (sum) duration of multiple occurrences of an activity by one person in a day.

The mean values and the mode value by state are sorted and classified with the quantile method on the map with different colors. Quantile assigns the same number of data to each class, and the resulting map is suitable to explain the order or sequential comparison for linearly distributed data

(ESRI, 2018). In this study, five quantiles are used to display the data on the map. For example, to compare the total time spent on watching TV in a day across 50 states, the average values of the states are ordered, and each quantile has 20% of the data (10 states).

3.3.3.2. GIS Grouping Analysis: Grouping of Activities with K-means Clustering

The pattern of each activity was grouped by similar states using the Grouping Analysis of ArcGIS. The Grouping Analysis is a part of the Spatial Analysis in ArcGIS, and it uses K-means clustering. Since most of the features (*Frequency, Duration per act, Duration per day, Start Time, End Time*) are numeric data, and only *Partner* feature is categorical, K-means clustering is applicable for clustering activity data by states. For the clustering with mixed feature types, K-prototype clustering is applicable as well, but currently ArcGIS only provides K-means clustering algorithm for Grouping Analysis.

To determine a suitable number of K, the pseudo F-statistic is computed. The pseudo F-statistic is the ratio of between-cluster variance to within-cluster variance, which is explained as follows (Caliński & Harabasz, 1974; Wilkinson, Engelman, Corter, & Coward, 2004).

$$pseudo\ F = \frac{GSS\ /\ (K-1)}{WSS\ /\ (N-K)}$$

where K is the number of clusters at any step in the hierarchical clustering, and N is the number of instances, GSS is the between-group sum of squares, and WSS is the within-group sum of squares. Large values of pseudo F denote cohesive and separated clusters. Especially, peaks in the pseudo F statistic indicate greater cluster separation. For the Grouping Analysis in GIS, pseudo F

static is run first, then the largest pseudo F value is used as the number of clusters in K-means clustering.

The result of the Grouping Analysis is displayed on a parallel box plot. The value ranges of the input features, *Frequency, Duration* (*Sum duration of an activity per day*), *Start Time, End Time* and *Partner*, are standardized with z-transform to remove the unexpected weight effect from different variances of the features. Z-transform is explained as follows (Witten et al., 2016).

$$z = \frac{x - \mu}{\sigma}$$

where x is the actual value, μ is the mean of the feature, and σ is the standard deviation of the feature.

3.4. Result

3.4.1. Clustering of Occupant Daily Activities by Time

Clustering of occupant daily activities are performed to identify the habitual daily routine of different groups of occupants. The occupant groups are distinguished by an unsupervised machine learning method, clustering, and a typical daily routine of activities for each group is detected by the centroid values of the group. The typical daily activities by time show when energy usage-related activities are performed during a day by the group of occupants. It helps to estimate the usage of appliances, lighting, heating and cooling systems which are associated with the activities.

The occupants are clustered from all of the national level ATUS data without segmentation by States. Before running K-modes clustering, the suitable number of clusters is examined using

Bayesian inference criterion (BIC). The K values are ranged from 2 to 10, and the BIC value is lowest when K is 6 as described in Figure 3-1.

Figure 3-1. BIC for Number of K

Then, K-modes clustering is run with K value of 6, and Table 3-6 summarizes the result. Cluster 1 (26%) and Cluster 2 (33%) have more occupants than the rest of the clusters (6-15%) and Cluster 3 has the lowest percentage of occupants (6%).

Table 3-6. Number of Occupants by Cluster

Cluster	Count	Percent
1	2761	26%
2	3510	33%
3	661	6%
4	1108	10%
5	1651	15%
6	1081	10%

Table 3-7 summarizes the distribution of the occupant cluster data categorized by Region (1: Northeast, 2: Midwest, 3: South, 4: West), Day (Weekday, Weekend), Gender (Male, Female) and Job Status (Yes: have one or more jobs, No: have no job). Yellow cells indicate the dominant subcategories which take approximately more than 60% of instances in the given category. They explain the main characteristics of the cluster. For example, in Cluster 1, Weekday (WD) is the

main subcategory in Day category, and Having job (Yes) is the main subcategory in Job category. It means that Cluster 1 more represents the activity pattern of the occupants who have a job during weekdays.

Table 3-7. Distribution of Data by Cluster

Cluster			C1	C2	C3	C4	C5	C6	All
Percent	Region	R1	15%	16%	16%	18%	16%	17%	16%
		R2	25%	24%	23%	25%	22%	26%	24%
		R3	38%	41%	39%	36%	37%	32%	38%
		R4	22%	19%	21%	21%	26%	25%	22%
	Day	WD	83%	38%	49%	29%	44%	39%	50%
		WE	17%	62%	51%	71%	56%	61%	50%
	Sex	M	51%	46%	51%	34%	30%	44%	44%
		F	49%	54%	49%	66%	70%	56%	56%
	Job	Yes	93%	41%	85%	53%	45%	50%	60%
		No	7%	59%	15%	47%	55%	50%	40%
Number	Region	R1	421	567	109	198	258	185	1738
		R2	685	832	151	273	356	278	2575
		R3	1037	1456	260	402	610	348	4113
		R4	618	655	141	235	427	270	2346
	Day	WD	2294	1345	322	316	730	422	5429
		WE	467	2165	339	792	921	659	5343
	Sex	M	1397	1605	335	381	492	480	4690
		F	1364	1905	326	727	1159	601	6082
	Job	Yes	2563	1451	559	583	746	543	6445
		No	198	2059	102	525	905	538	4327
	Total		2761	3510	661	1108	1651	1081	10772

*** All column: For "Percent" part, the ratio of the instance numbers of each subcategory to the category in percentage (e.g. Region1 takes 16% of all Regions). For "Number" part, the instance number of each subcategory.*
*** WD: Weekday, WE: Weekend, M: Male, F: Female*

The centroid is the center value of each cluster, and it represents the typical characteristics of each cluster. The centroid is stabilized after the several iterations of the K-modes clustering process, and the final values of the centroid can be simpler than the other actual instances in the cluster. In this occupant activity dataset, the centroid values (the activities and their start and end time) represent the typical daily schedule of the occupant group, and the actual instances close to this centroid in the cluster might have more diverse activities by time than the centroid. The start time and end time of the habitual energy usage-related activities help to estimate residential energy

consumption during the time by associating the relevant appliances. Also, it provides the daily routine of the habitual energy usage-related activities which can be used for energy control strategies, occupant intervention or education. The centroid values of the clusters are plotted in Figure 3-2, and their characteristics are explained in the following tables (from Table 3-8 to Table 3-13). Among the activities in the tables, the energy usage-related activities (specified in Table 3-2) are marked as bold fonts, and the most habitual activities (specified in Table 3-5) are underlined.

- ***Occupant Cluster 1***: This group takes 26% of all occupants (Table 3-6). Based on the data distribution of the clusters (Table 3-7), this schedule is more from weekday diaries (83%) than weekend diaries (17%). This group consists of similar ratio of male and female occupants (51%:49%), and most of the occupants in this group have a job (93%). They wake up early in the morning around 6:45 and work till early evening. After having dinner and watching TV, they go to bed around 21:30. In this cluster, *Work, main job* and *Television and movies* are the energy usage-related activities, which infers that residential energy consumption might be changed during the time of these activities. In this dataset, *Work, main job* can happen either at home or any other places, such as an office. If they work at home, residential energy consumption might be higher, but if they work at their office (or any other places), residential energy consumption might be lower during the time. *Television and movies* is one of the most habitual energy usage-related activities at home, and it means that the occupants in this group tend to watch TV around from 18:27 to 21:32 (Table 3-8).

Table 3-8. Centroid of Occupant Cluster 1

Start	End	Code	Description
4:00	6:44	010101	Sleeping
6:45	**17:59**	**050101**	**Work, main job**
18:00	18:26	110101	Eating and drinking
18:27	**21:32**	**120303**	**Television and movies (not religious)**
21:33	3:59	010101	Sleeping

- ***Occupant Cluster 2***: This group takes 33% of all occupants. This schedule is more from weekend diaries (62%) than weekday diaries (38%). This group consists of similar ratio of male and female occupants (46%:54%), and more occupants in this group have no job (59%). They wake up late in the morning around 10:00, and watch TV all day long until late night around 22:30, then go to bed. In this cluster, *Television and movies* is the habitual energy usage-related activity. It infers that the occupants in this group tend to watch TV around from 10:00 to 22:29, and residential energy consumption might be higher during the time of this activity (Table 3-9). It is also possible that this simple schedule is due to the respondents' simplified answers about their daily activities.

Table 3-9. Centroid of Occupant Cluster 2

Start	End	Code	Description
4:00	9:59	010101	Sleeping
10:00	**22:29**	**120303**	**Television and movies (not religious)**
22:30	3:59	010101	Sleeping

- ***Occupant Cluster 3***: This group takes 6% of all occupants. This schedule is similarly from weekday diaries (49%) and weekend diaries (51%). This group consists of similar ratio of male and female occupants (51%:49%), and most of the occupants in this group have a job (85%). In this cluster, *Work, main job* is the energy usage-related activity, which infers that residential energy consumption might be higher or lower during this time of the activity depending on where they work (Table 3-10).

Table 3-10. Centroid of Occupant Cluster 3

Start	End	Code	Description
4:00	11:59	010101	Sleeping
12:00	**23:29**	**050101**	**Work, main job**
23:30	3:59	010101	Sleeping

- ***Occupant Cluster 4***: This group takes 10% of all occupants. This schedule is more from weekend diaries (71%) than weekday diaries (29%). This group consists of more female occupants (66%) than male occupants (34%), and the ratio of occupants who have a job or not is similar (53%:47%). They socialize with others during the day time and watch TV at night. In this cluster, *Television and movies* is the habitual energy usage-related activity. It infers that the occupants in this group tend to watch TV around from 21:00 to 21:59, and residential energy consumption might be higher during the time of this activity (Table 3-11).

Table 3-11. Centroid of Occupant Cluster 4

Start	End	Code	Description
4:00	9:59	010101	Sleeping
10:00	20:59	120101	Socializing and communicating with others
21:00	**21:59**	**120303**	**Television and movies (not religious)**
22:00	3:59	010101	Sleeping

- ***Occupant Cluster 5***: This group takes 15% of all occupants. This schedule is similarly from weekday diaries (44%) and weekend diaries (56%). This group consists of more female occupants (70%) than male occupants (30%), and the ratio of occupants who have a job or not is similar (45%:55%). They spend time for interior cleaning, eating, napping, and food preparation during the day time, then watch TV in the evening. In this cluster, *Washing, dressing, and grooming, Interior cleaning, Food and drink preparation* and *Television and movies* are the energy usage-related activities, which infers that residential energy consumption might be higher during the time of these activities. *Washing, dressing,*

and grooming, Food and drink preparation and *Television and movies* are the habitual energy usage-related activity at home (Table 3-12).

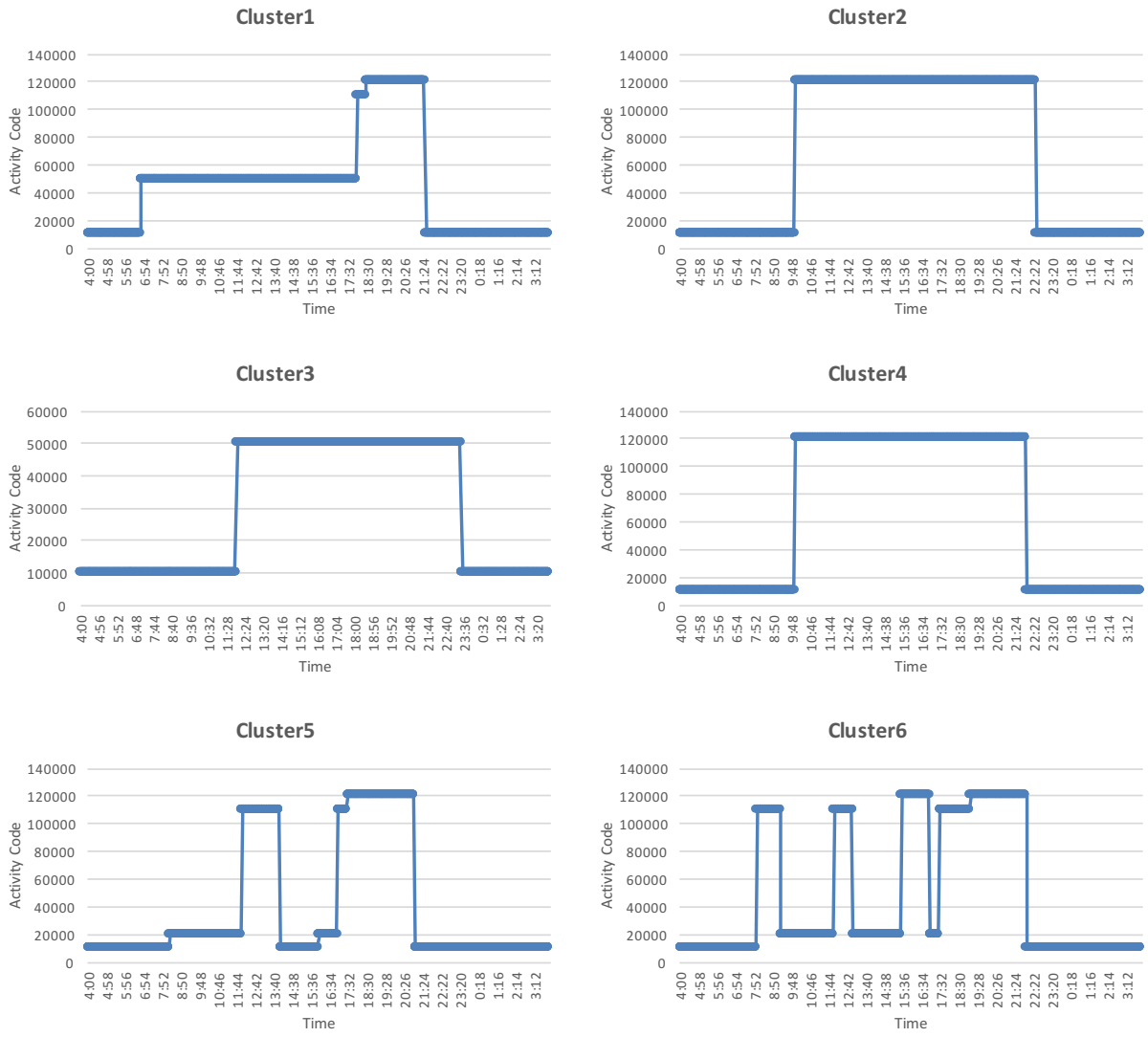
Table 3-12. Centroid of Occupant Cluster 5

Start	End	Code	Description
4:00	7:59	010101	Sleeping
8:00	**8:14**	**010201**	**Washing, dressing, and grooming oneself**
8:15	**11:59**	**020101**	**Interior cleaning**
12:00	13:59	110101	Eating and drinking
14:00	15:59	010101	Sleeping
16:00	**16:59**	**020201**	**Food and drink preparation**
17:00	17:29	110101	Eating and drinking
17:30	**20:59**	**120303**	**Television and movies (not religious)**
21:00	3:59	010101	Sleeping

- ***Occupant Cluster 6***: This group takes 10% of all occupants. This schedule is more from weekend diaries (61%) than weekday diaries (39%). This group consists of similar ratio of male and female occupants (44%:56%), and the ratio of occupants who have a job or not is same (50%:50%). In this cluster, *Lawn, garden, and houseplant care, Food and drink preparation* and *Television and movies* are the energy usage-related activities, which infers that residential energy consumption might be higher during the time of these activities. *Food and drink preparation and Television and movies* are the habitual energy usage-related activities at home, and it means that the occupants in this group tend to cook around from 16:36 to 17:30, and watch TV around from 18:46 to 22:30 (Table 3-13).

Table 3-13. Centroid of Occupant Cluster 6

Start	End	Code	Description
4:00	7:59	010101	Sleeping
8:00	9:14	110101	Eating and drinking
9:15	**11:59**	**020501**	**Lawn, garden, and houseplant care**
12:00	12:59	110101	Eating and drinking
13:00	**15:29**	**020501**	**Lawn, garden, and houseplant care**
15:30	**16:59**	**120303**	**Television and movies (not religious)**
17:00	**17:29**	**020201**	**Food and drink preparation**
17:30	19:04	110101	Eating and drinking
19:05	**21:59**	**120303**	**Television and movies (not religious)**
22:00	3:59	010101	Sleeping



** *Y-axis values are the 3rd tier ATUS activity codes which are nominal values in the form of numbers. 10000 is not "higher" than 8000, and same Activity Code values indicate same activities and different values indicate different activities.*

Figure 3-2. Daily Activity Routines of Occupant Clusters

Television and movies (120303) is one of the most habitual energy usage-related activities, and it is included in 5 clusters: Cluster 1 (from 18:27 to 21:32), Cluster 2 (from 10:00 to 22:29), Cluster 4 (from 21:00 to 21:59), Cluster 5 (from 17:30 to 20:59), and Cluster 6 (from 15:30 to 16:59, from 19:05 to 21:59). Based on the overlapping time from these 5 clusters, the occupants watch TV around from 18:30 to 21:30, and it means that one of the most habitual energy usage-related

activities strongly tend to happen during this time. It infers that most of the occupants in the U.S. watch TV or movie in the late evening regardless of occupant groups (day of the week, gender, job status, etc.). This activity is strongly habitual in their ordinary daily schedule, and the occupants keep this activity except for the cases when they have other special occasions. It also infers that this strong habit is difficult to be changed with other occupant interventions or education. In addition, the energy consumption for watching TV or movie during this time can be stably estimated.

3.4.2. Comparative Analysis for Energy Usage-Related Activities

Energy usage-related activities at home are compared by several factors to identify which factors influence on the similarity and difference of the activities. The result can guide how to set energy control or intervention/education strategies by different sub-groups of the factors examined in this section. The mean values (*Frequency, Duration per Act, Duration per Day, Start Time, End Time*) and mode values (*Partner*) of the energy usage-related activities at home (Table 3-3) are compared in this section as follows: (1) activities by the regions, (2) activities by the day of the week (weekdays vs. weekends), (3) activities by gender (male vs. female), and (4) activities by job status (yes vs. no). The mean or mode values calculated from all data are used as the baseline to compare other subcategory values. The line of the mean/mode values from all data are included in every figures under this 3.4.2. section. The detailed results of the comparative analysis are listed in the tables in Appendix.

3.4.2.1. Activities by Region

Frequency, Durations, Start Time, End Time, and *Partner* of energy usage-related activities less differ by regions compared to clusters (the previous subsection). It infers that relatively similar energy strategies can be used in the different regions for those selected activities. In this section, which components more affect on differences of the energy usage-related activities, and which activities are more different by the region.

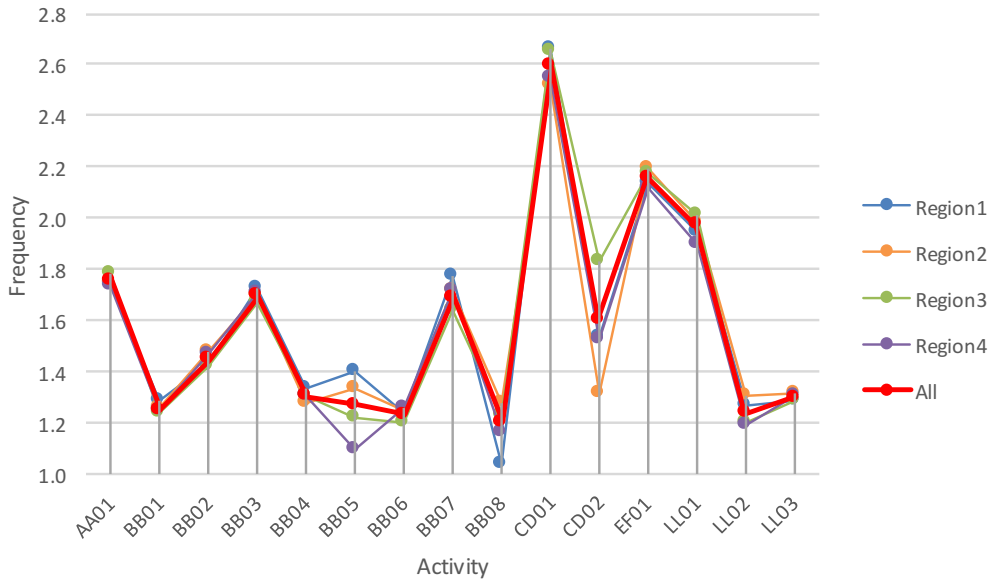


Figure 3-3. Comparison of Frequency by Region

Frequency (Figure 3-3) of *Physical care for children* (CD01) in Region 3 (South) is the highest (2.65 times) among all regions and activities. *Frequency* of *Vehicle repair and maintenance* (BB08) in Region 1 (Northeast) is the lowest (1.04 times) among all regions and activities. *Frequency* of *Heating and cooling* (BB05), *Physical care for/helping adults* (CD02) and *Vehicle repair and maintenance* (BB08) show greater differences among regions, which means that these activities need different energy strategies regarding *Frequency* in different regions. The rest of the activities

are similar among regions, which means that energy strategies for these activities can be similar regarding *Frequency* for different regions.

Duration per act (Figure 3-4) of *Watching TV* (LL01) is the longest (123.36 minutes) in Region 3 (South) among all regions and activities. *Duration per act* of *Care for animals and pets* (BB07) in Region 2 (Midwest) is the shortest (15.55 minutes). *Duration per act* of *Heating and cooling* (BB05) and *Vehicle repair and maintenance* (BB08) show greater differences among regions, which means that these activities need different energy strategies regarding *Duration per act* in different regions. The rest of the activities are similar among regions, which means that energy strategies for these activities can be similar regarding *Duration per act* for different regions.

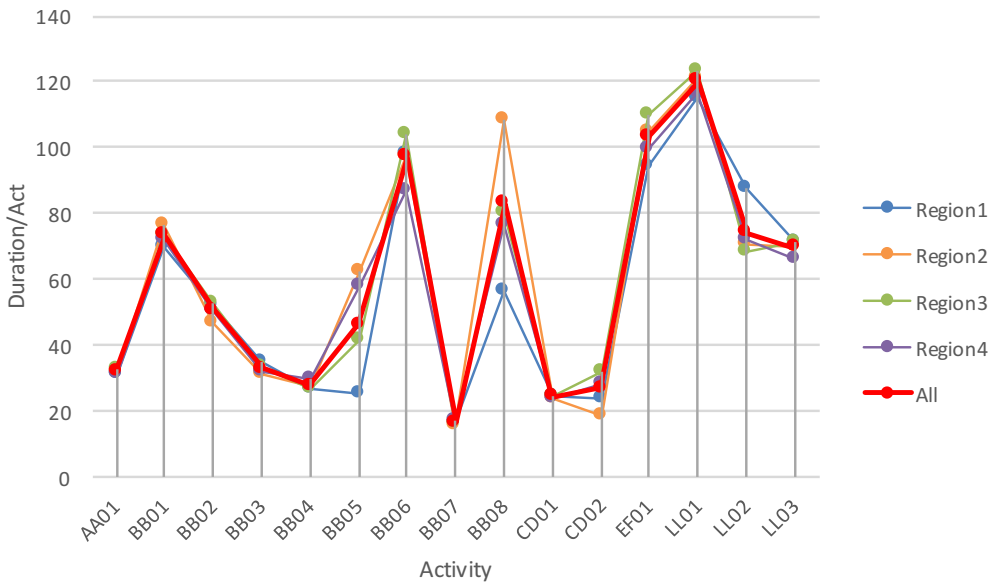


Figure 3-4. Comparison of Duration per Act by Region

Duration per day (Figure 3-5) of *Watching TV* (LL01) in Region 3 (South) is the longest (207.95 minutes) among all regions and activities. *Duration per day* of *Physical care for/helping adults* (CD02) in Region 2 (Midwest) is the shortest (24.25 minutes). *Vehicle repair and maintenance*

(BB08) show a greater difference among regions, which means that this activity needs different energy strategies regarding *Duration per day* in different regions. The rest of the activities are similar among regions, which means that energy strategies for these activities can be similar regarding *Duration per day* for different regions.

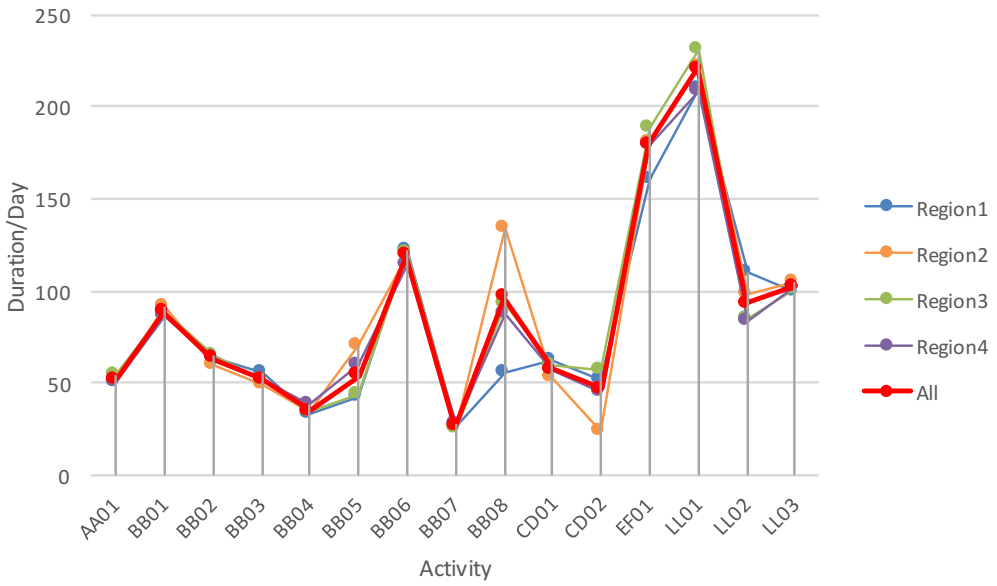


Figure 3-5. Comparison of Duration per Day by Region

Start Time (Figure 3-6) and *End Time* (Figure 3-7) of Physical care for/helping adults (CD02) in Region 3 (South) are the earliest (*Start Time* 11:41 – 701.49 in minutes, *End Time* 12:13 – 733.35 in minutes) among all regions and activities. *Start Time* of *Watching TV* (LL01) in Region 1 (Northeast) is the latest (17:06 – 1026.86 in minutes), and *End Time* of Watching TV (LL01) in Region 4 (West) is the latest (17:51 – 1071.83 minutes). *Start Time* and *End Time* of *Heating and cooling* (BB05), *Vehicle repair and maintenance* (BB08), *Physical care for/helping adults* (CD02), and *Listening to/playing radio or music* (LL02) show greater differences among regions, which means that these activities need different energy strategies regarding *Start Time* and *End Time* in different regions. *Interior cleaning* (BB01), *Laundry* (BB02), and *Watching TV* (LL01) are similar

among regions, which means that energy strategies for these activities can be similar regarding *Start Time* and *End Time* for different regions.

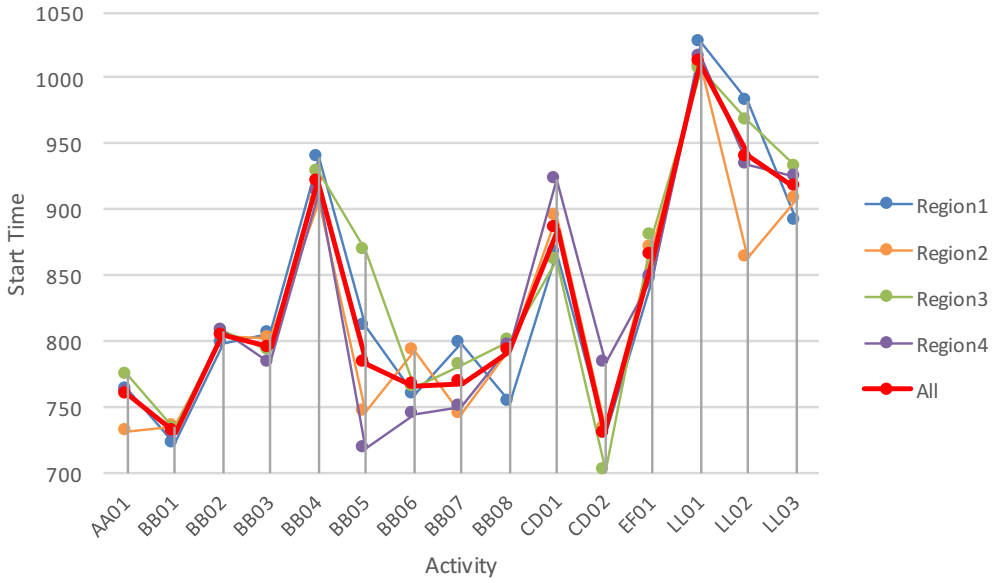


Figure 3-6. Comparison of Start Time by Region

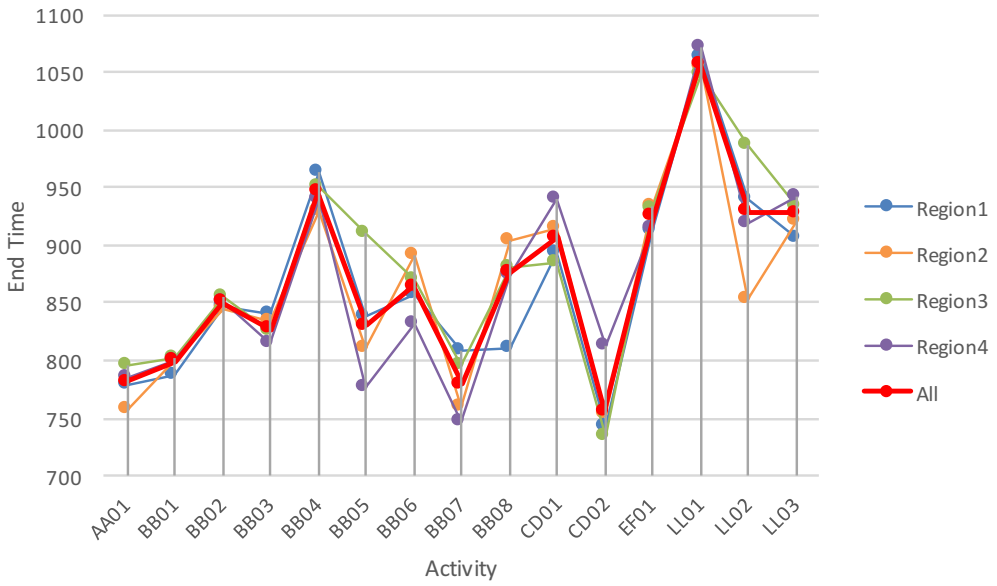
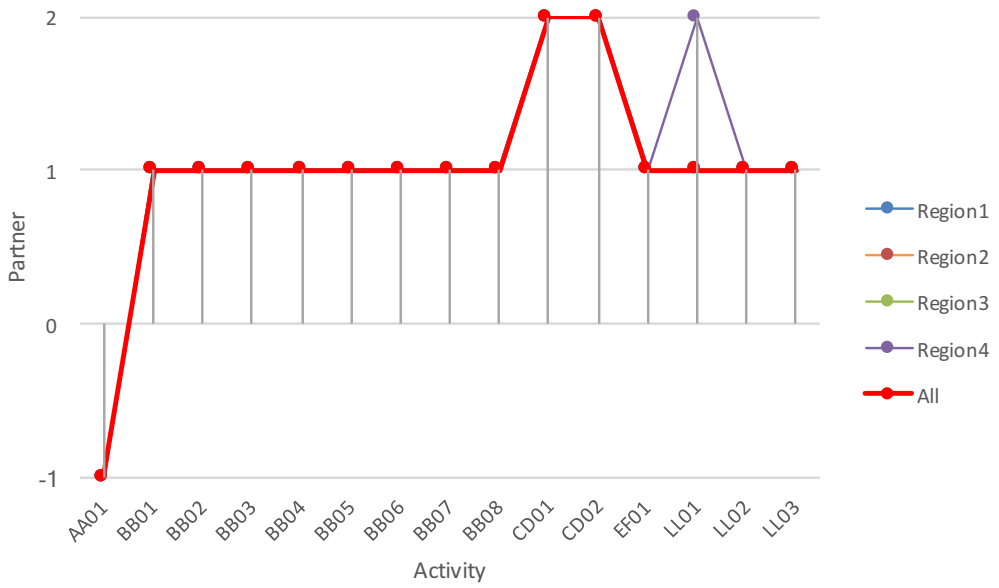


Figure 3-7. Comparison of End Time by Region

Generally, most of the activities in all regions have same mode values (Figure 3-8). However, *Partner* of *Watching TV* (LL01) shows differences among regions, which means that this activity

needs different energy strategies regarding *Partner* in different regions. *Partner* of the rest of the activities are same among clusters, which means that energy strategies for these activities can be similar regarding *Partner* for different regions.



*** Partner Code: Not Recorded (-1), Alone (1), with Family (2)*

Figure 3-8. Comparison of Partner by Region

Table 3-14 summarizes the difference of activities by region which is derived from the *ANOVA*. In general, *Duration per act, Duration per day, Start Time*, and *End Time* more affect on the differences of the energy usage-related activities in different regions, and *Frequency* and *Partner* less affect on them. Considering all the given six components, *Washing, dressing, and grooming* (AA01), *Food and drink preparation* (BB03), and *Watching TV* (LL01) are generally different by regions. The rest of the activities are not significantly different by regions, which means that these activities are more similar in all regions. In sum, the four census regions have less influences on the difference of activities.

Table 3-14. Difference in Activities by Region

Act.	Freq	Dur/a	Dur/d	Start	End	Partner
AA01	1.21 (0.30)	4.73 (0.00)*	8.10 (0.00)*	8.85 (0.00)*	7.72 (0.00)*	n/a (n/a)
BB01	0.83 (0.48)	1.01 (0.39)	0.53 (0.66)	0.31 (0.81)	0.43 (0.73)	5.21 (0.00)*
BB02	0.61 (0.61)	1.64 (0.18)	1.05 (0.37)	0.11 (0.95)	0.23 (0.88)	4.30 (0.00)*
BB03	0.94 (0.42)	3.29 (0.02)*	3.30 (0.02)*	2.16 (0.09)	2.38 (0.07)	5.47 (0.00)*
BB04	1.05 (0.37)	2.24 (0.08)	1.95 (0.12)	1.78 (0.15)	1.70 (0.16)	1.30 (0.27)
BB05	1.19 (0.32)	1.06 (0.37)	0.57 (0.64)	0.79 (0.50)	0.53 (0.66)	1.24 (0.30)
BB06	0.66 (0.58)	1.95 (0.12)	0.26 (0.85)	2.43 (0.06)	3.83 (0.01)*	0.97 (0.41)
BB07	0.82 (0.48)	0.31 (0.82)	0.30 (0.83)	1.87 (0.13)	2.36 (0.07)	1.47 (0.22)
BB08	1.90 (0.13)	2.81 (0.04)*	4.10 (0.01)*	0.28 (0.84)	0.97 (0.41)	1.31 (0.27)
CD01	0.53 (0.66)	0.11 (0.95)	1.86 (0.13)	4.47 (0.00)*	3.78 (0.01)*	0.63 (0.59)
CD02	0.94 (0.42)	1.10 (0.35)	1.49 (0.22)	0.46 (0.71)	0.42 (0.74)	0.13 (0.94)
EF01	0.22 (0.88)	1.97 (0.12)	1.31 (0.27)	1.40 (0.24)	0.44 (0.73)	0.36 (0.78)
LL01	4.25 (0.01)*	3.32 (0.02)*	8.54 (0.00)*	3.05 (0.03)*	2.68 (0.05)*	5.28 (0.00)*
LL02	0.58 (0.63)	1.31 (0.27)	1.34 (0.26)	1.58 (0.19)	1.44 (0.23)	1.01 (0.39)
LL03	0.19 (0.90)	0.92 (0.43)	0.17 (0.92)	1.97 (0.12)	1.10 (0.35)	1.10 (0.35)

*** ANOVA: F-value (p-value), * denotes significant p-value*

3.4.2.2. Activities by Day of the Week

Durations, Start Time and End Time of energy usage-related activities more differ by day of the week, and *Frequency* and *Partner* less differ by day of the week. It infers that the energy strategies need to consider the energy usage-related activities on weekdays vs. weekends more focusing on when and how long those activities happen.

Frequency (Figure 3-9) of *Physical care for children* (CD01) on weekdays is the highest (2.61 times) among all activities, and *Frequency* of *Work for job(s)/research/homework* (EF01) on weekdays is the 2nd highest (2.52 times). *Frequency* of *Vehicle repair and maintenance* (BB08) on weekdays is the lowest (1.20 times). *Frequency* of *Work for job(s)/research/homework* (EF01) shows a greater difference between weekdays and weekends, which means that this activity needs different energy strategies regarding *Frequency* for weekdays and weekends. The rest of the activities are similar on weekdays and weekends, which means that energy strategies for these activities can be similar regarding *Frequency* for weekdays and weekends.

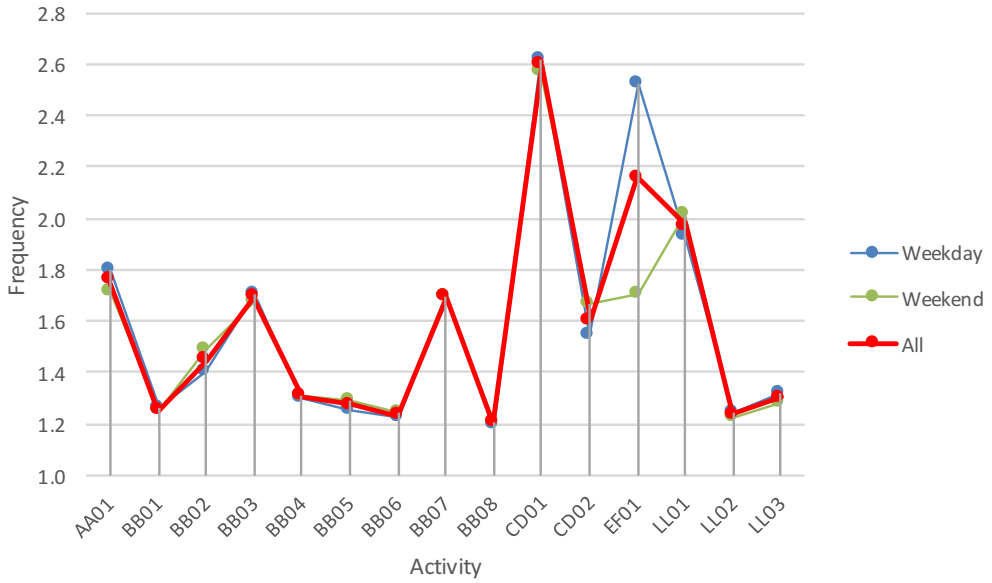


Figure 3-9. Comparison of Frequency by Day of the Week

Duration per act (Figure 3-10) of *Watching TV* (LL01) on weekends is the longest (129.81 minutes) among all activities, and *Work for job(s)/research/homework* (EF01) on weekdays (107.13) and *Gardening, ponds, pools, and hot tubs* (BB06) on weekends (105.76 minutes) are next longest ones other than *Watching TV*. *Care for animals and pets* (BB07) weekdays (15.99 minutes) is the shortest. *Duration per act* of *Gardening, ponds, pools, and hot tubs* (BB06) and *Watching TV* (LL01) show greater differences between weekdays and weekends, which means that these activities need different energy strategies regarding *Duration per act* for weekdays and weekends. The rest of the activities are similar on weekdays and weekends, which means that energy strategies for these activities can be similar regarding *Duration per act* for weekdays and weekends.

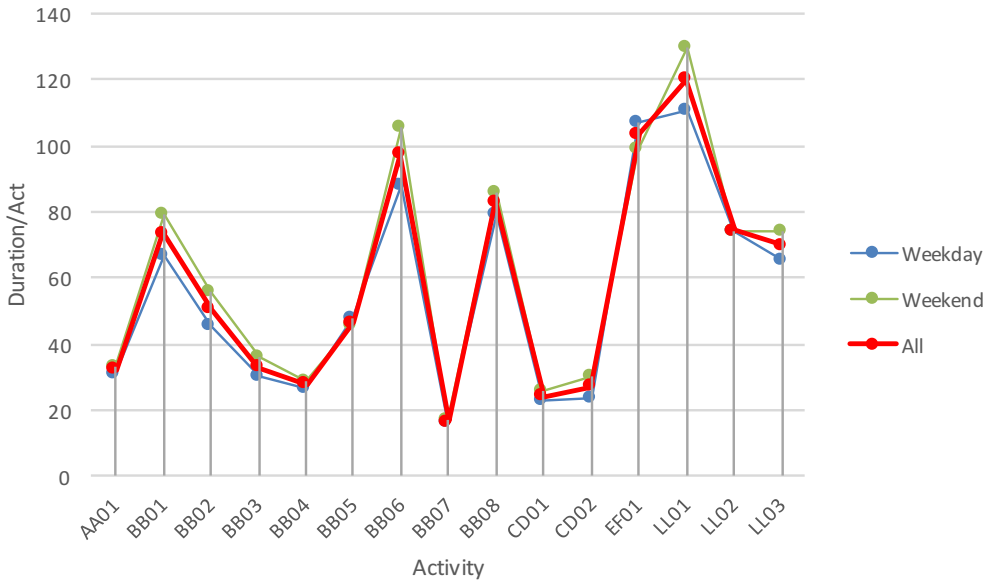


Figure 3-10. Comparison of Duration per act by Day of the Week

Duration per day (Figure 3-11) of *Watching TV* (LL01) on weekends is the longest (239.92 minutes) among all activities, and *Care for animals and pets* (BB07) weekdays (25.61 minutes) is the shortest. *Duration per day* of *Gardening, ponds, pools, and hot tubs* (BB06), *Work for job(s)/research/homework* (EF01) and *Watching TV* (LL01) show greater differences between weekdays and weekends, which means that these activities need different energy strategies regarding *Duration per day* for weekdays and weekends. The rest of the activities are similar on weekdays and weekends, which means that energy strategies for these activities can be similar regarding *Duration per day* for weekdays and weekends.

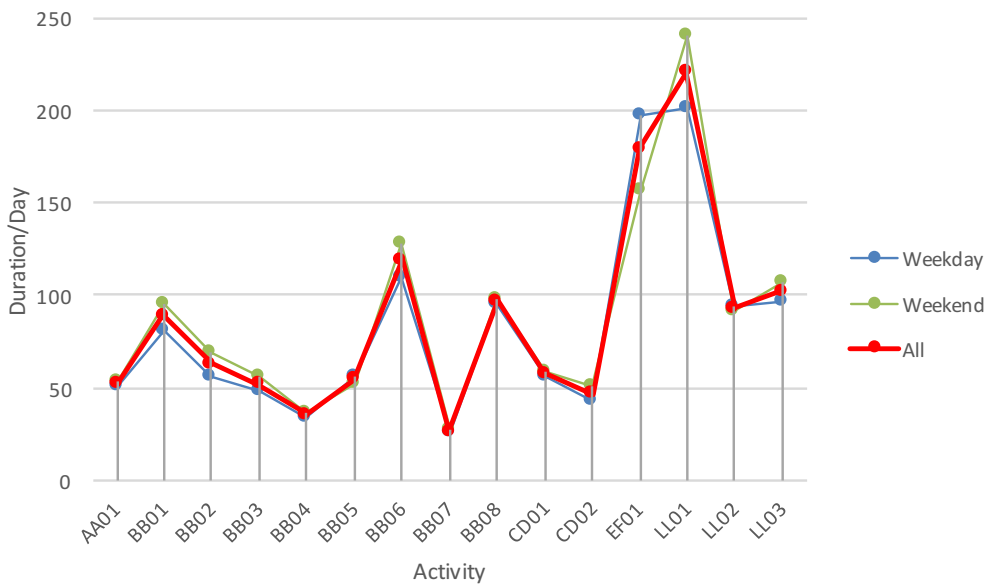


Figure 3-11. Comparison of Duration per Day by Day of the Week

Start Time (Figure 3-12) and *End Time* (Figure 3-13) of Physical care for/helping adults (CD02) on weekdays are the earliest (*Start Time* 11:06 – 666.13 in minutes, *End Time* 11:36 – 696.18 in minutes) among all activities. *Start Time* and *End Time* of Watching TV (LL01) on weekdays are the latest (*Start Time* 17:13 – 1033.02 minutes, *End Time* 17:53 – 1073.30 minutes). *Start Time* and *End Time* of *Physical care for children* (CD01) and *Physical care for/helping adults* (CD02) show greater differences between weekdays and weekends compared to other activities, which means that these activities need different energy strategies regarding *Start Time* and *End Time* for weekdays and weekends. The rest of the activities are similar on weekdays and weekends, which means that energy strategies for these activities can be similar regarding *Start Time* and *End Time* for weekdays and weekends.

Figure 3-12. Comparison of Start Time by Day of the Week

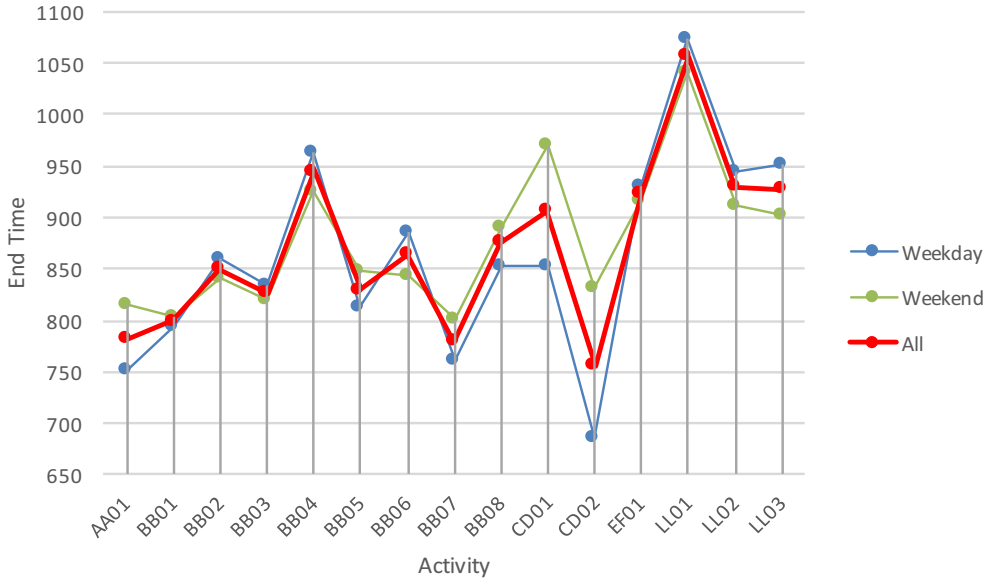
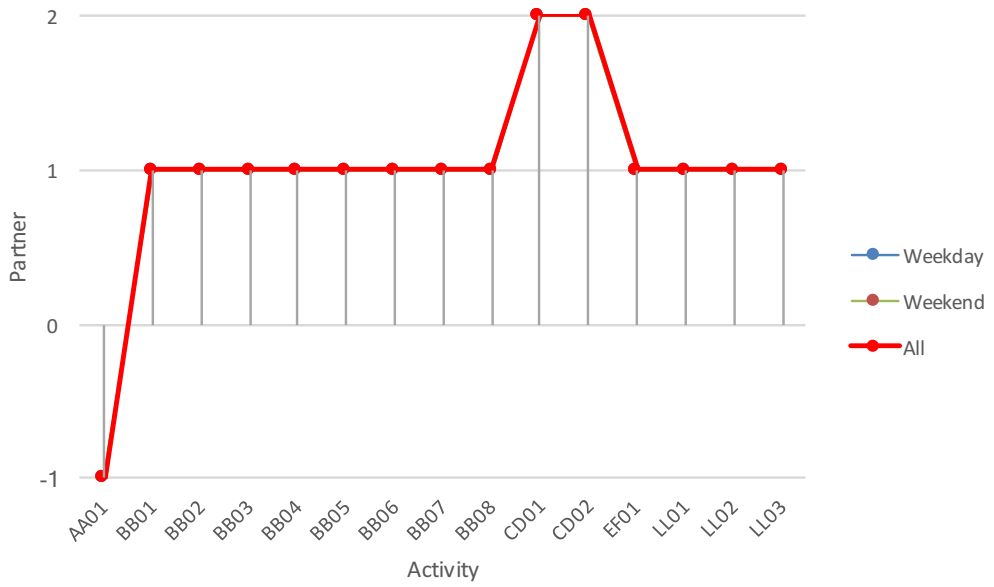


Figure 3-13. Comparison of End Time by Day of the Week

Partner of all activities are same on weekdays and weekends, which means that energy strategies for these activities can be similar regarding *Partner* for weekdays and weekends (Figure 3-14).



*** Partner Code: Not Recorded (-1), Alone (1), with Family (2)*

Figure 3-14. Comparison of Partner by Day of the Week

Table 3-15 summarizes the difference of activities by day of the week which is derived from the *t-test*. In general, *Duration per act, Duration per day, Start Time*, and *End Time* more affect on the differences of the energy usage-related activities in different day of the week (weekday or weekend), and *Frequency* and *Partner* less affect on them. Considering all the given six components, *Washing, dressing, and grooming* (AA01), *Interior cleaning* (BB01), *Laundry* (BB02), *Food and drink preparation* (BB03), *Kitchen and food clean-up* (BB04), *Gardening, ponds, pools, and hot tubs* (BB06), *Physical care for children* (CD01), *Watching TV* (LL01), and *General computer use* (LL03) are generally different by day of the week. The rest of the activities are not significantly different by day of the week, which means that these activities are more similar in weekdays and weekends. In sum, day of the week has strong influences on the difference of activities.

Table 3-15. Differences in Activities by Day of the Week

Act.	Freq	Dur/a	Dur/d	Start	End	Partner
AA01	4.59 (0.00)*	-5.24 (0.00)*	-2.44 (0.01)*	-10.49 (0.00)*	-10.43 (0.00)*	n/a (n/a)
BB01	0.45 (0.65)	-4.98 (0.00)*	-4.35 (0.00)*	-0.27 (0.78)	-0.94 (0.35)	-5.74 (0.00)*
BB02	-1.99 (0.05)*	-4.53 (0.00)*	-5.19 (0.00)*	3.25 (0.00)*	1.80 (0.07)*	-3.63 (0.00)*
BB03	1.08 (0.28)	-7.63 (0.00)*	-6.62 (0.00)*	3.06 (0.00)*	2.08 (0.04)*	-5.77 (0.00)*
BB04	-0.51 (0.61)	-2.09 (0.04)*	-2.40 (0.02)*	3.26 (0.00)*	3.50 (0.00)*	-1.89 (0.06)
BB05	-0.27 (0.79)	0.12 (0.91)	0.23 (0.82)	-0.52 (0.61)	-0.48 (0.63)	-1.09 (0.28)
BB06	-0.72 (0.47)	-3.44 (0.00)*	-2.72 (0.01)*	4.52 (0.00)*	3.29 (0.00)*	-4.65 (0.00)*
BB07	0.03 (0.98)	-1.01 (0.31)	-1.18 (0.24)	-2.07 (0.04)*	-2.23 (0.03)*	-2.44 (0.01)*
BB08	-0.22 (0.83)	-0.55 (0.58)	-0.15 (0.88)	-0.97 (0.33)	-1.17 (0.25)	-2.76 (0.01)*
CD01	0.46 (0.65)	-3.06 (0.00)*	-0.84 (0.40)	-9.23 (0.00)*	-8.98 (0.00)*	-0.70 (0.48)
CD02	-0.53 (0.60)	-1.24 (0.22)	-0.63 (0.53)	-2.64 (0.01)*	-2.79 (0.01)*	-0.29 (0.77)
EF01	12.57 (0.00)*	1.70 (0.09)	4.02 (0.00)*	-1.32 (0.19)	0.83 (0.41)	0.35 (0.72)
LL01	-3.60 (0.00)*	-9.51 (0.00)*	-10.17 (0.00)*	8.63 (0.00)*	5.08 (0.00)*	-6.00 (0.00)*
LL02	0.28 (0.78)	0.02 (0.98)	0.25 (0.80)	-0.04 (0.97)	0.72 (0.47)	-1.44 (0.15)
LL03	1.63 (0.10)	-3.58 (0.00)*	-2.48 (0.01)*	3.29 (0.00)*	3.56 (0.00)*	-2.10 (0.04)*

*** t-test: t-value (p-value), * denotes significant p-value*

3.4.2.3. Activities by Gender

Frequency and *Durations* (especially, *Durations per day*) of energy usage-related activities more differ by gender, *Start Time, End Time* and *Partner* less differ by gender. It infers that the energy strategies need to consider the energy usage-related activities on male vs. female more focusing on how often those activities happen.

Frequency (Figure 3-15) of *Physical care for children* (CD01) by female is the highest (2.88 times) among all activities, and *Frequency* of *Vehicle repair and maintenance (BB08)* by female is the lowest (1.09 times). *Frequency* of *Physical care for children* (CD01), *Food and drink preparation* (BB03), *Laundry* (BB02) show greater differences between male and female, which means that these activities need different energy strategies regarding *Frequency* for male and female. The rest of the activities are similar, which means that energy strategies for these activities can be similar regarding *Frequency* for male and female.

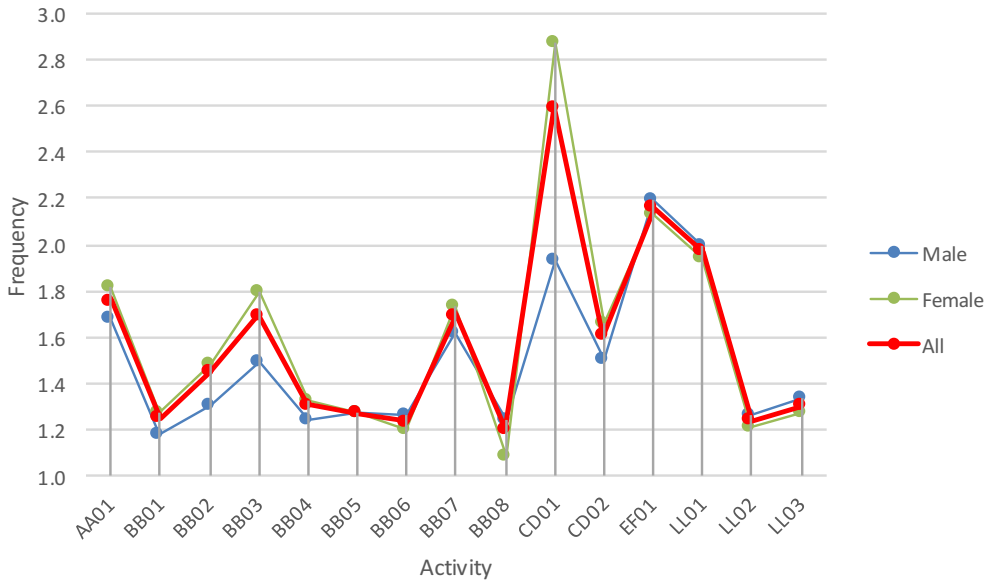


Figure 3-15. Comparison of Frequency by Gender

Duration per act (Figure 3-16) of *Watching TV* (LL01) by male is the longest (128.62 minutes) among all activities, and *Care for animals and pets* (BB07) by female is the shortest (15.88 minutes). *Duration per act* of *Heating and cooling* (BB05), *Gardening, ponds, pools, and hot tubs* (BB06), *Vehicle repair and maintenance* (BB08), *Watching TV* (LL01), *Listening to/playing radio or music* (LL02), *General computer use* (LL03) show greater differences between male and female compared to other activities, which means that these activities need different energy strategies regarding *Duration per act* for male and female. Generally, male occupants spend longer time (per activity) on these activities than female occupants. The rest of the activities are similar, which means that energy strategies for these activities can be similar regarding *Duration per act* for male and female.

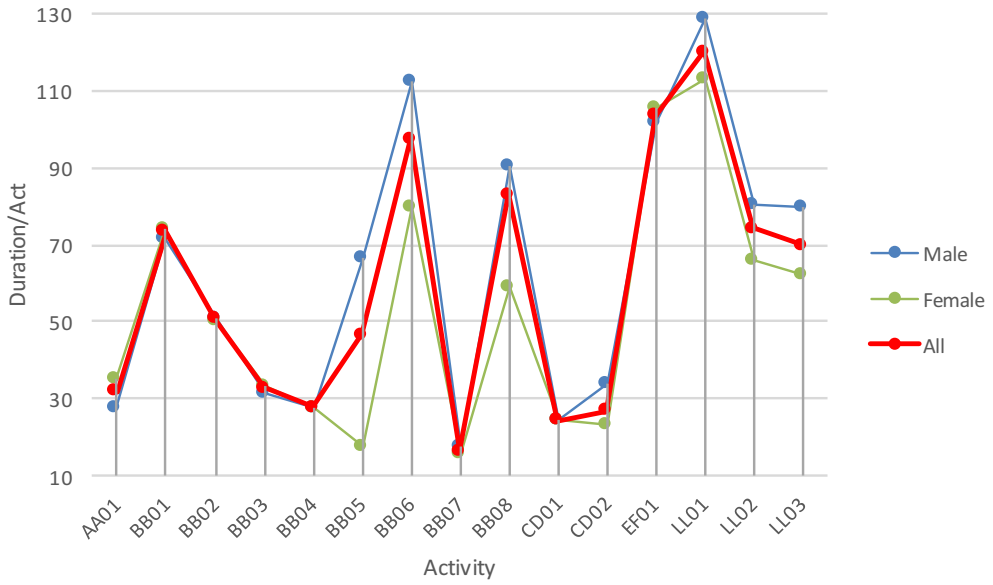


Figure 3-16. Comparison of Duration per Act by Gender

Duration per day (Figure 3-17) of Watching TV (LL01) by male is the longest (238.84 minutes) among all activities, and Care for animals and pets (BB07) by female is the shortest (26.57 minutes). *Duration per day* of *Heating and cooling* (BB05), *Gardening, ponds, pools, and hot tubs* (BB06), *Vehicle repair and maintenance* (BB08), *Physical care for children* (CD01), *Physical care for/helping adults* (CD02), *Watching TV* (LL01), *Listening to/playing radio or music* (LL02), *General computer use* (LL03) show greater differences between male and female compared to other activities, which means that these activities need different energy strategies regarding *Duration per day* for male and female. Male occupants spend longer time on these activities during a day than female occupants except for *Physical care for children* (CD01). Female occupants spend longer time for *Physical care for children* (CD01) than male occupants. The rest of the activities are similar, which means that energy strategies for these activities can be similar regarding *Duration per day* for male and female.

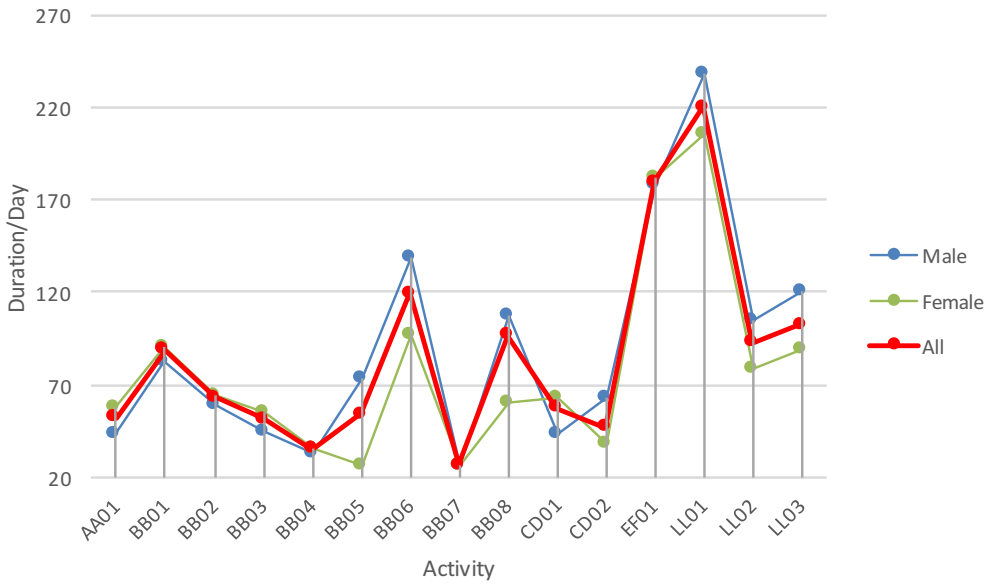


Figure 3-17. Comparison of Duration per Day by Gender

Start Time (Figure 3-18) and *End Time* (Figure 3-19) of *Heating and cooling* (BB05) by female are the earliest (*Start Time* 11:55 – 715.40 in minutes, *End Time* 12:13 – 733.29 in minutes) among all activities. *Start Time* and *End Time* of *Watching TV* (LL01) by female are the latest (*Start Time* 17:02 – 1022.01 minutes, *End Time* 17:42 – 1062.16 minutes). *Start Time* and *End Time* of *Heating and cooling* (BB05) and *Physical care for children* (CD01) show greater differences between male and female compared to other activities, which means that these activities need different energy strategies regarding *Start Time* and *End Time* for male and female. Male occupants start and end these activities later than female occupants. The rest of the activities are similar, which means that energy strategies for these activities can be similar regarding *Start Time* and *End Time* for male and female.

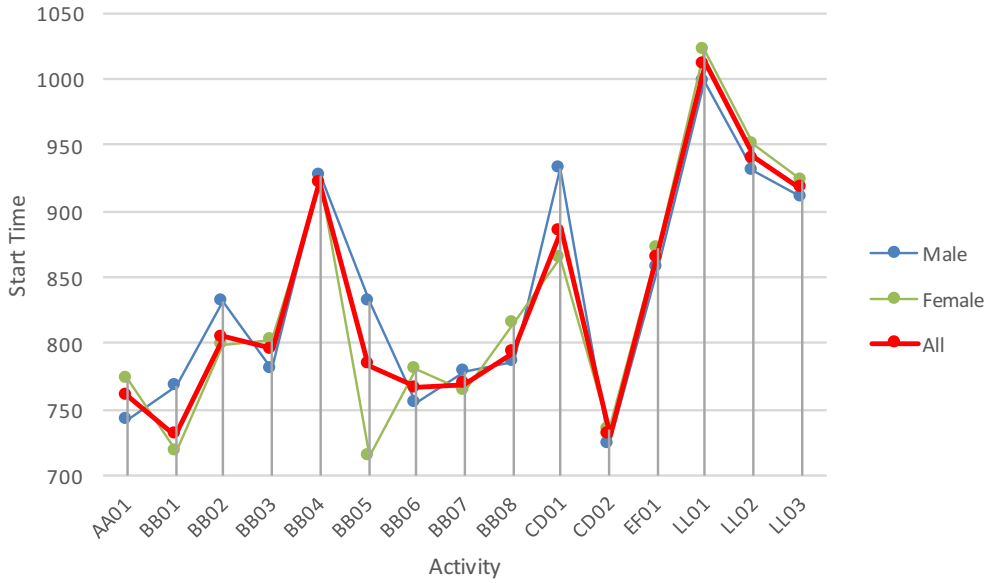
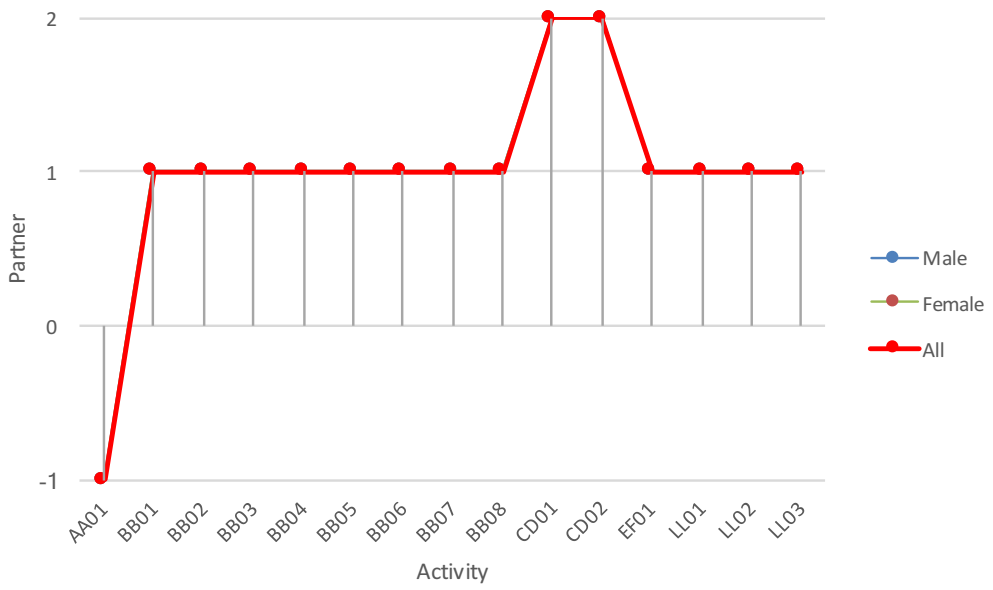


Figure 3-18. Comparison of Start Time by Gender

Figure 3-19. Comparison of End Time by Gender

Partner of all activities are same between male and female, which means that energy strategies for these activities can be similar regarding *Partner* for male and female. (Figure 3-20).



*** Partner Code: Not Recorded (-1), Alone (1), with Family (2)*

Figure 3-20. Comparison of Partner by Gender

Table 3-16 summarizes the difference of activities by gender which is derived from the *t-test*. In general, *Frequency, Duration per act,* and *Duration per day* more affect on the differences of the energy usage-related activities in different gender (male or female), and *Start Time*, *End Time* and *Partner* less affect on them. Considering all the given six factors, *Washing, dressing, and grooming* (AA01), *Interior cleaning* (BB01), *Laundry* (BB02), *Food and drink preparation* (BB03), *Kitchen and food clean-up* (BB04), *Heating and cooling* (BB05), *Vehicle repair and maintenance* (BB08), *Physical care for children* (CD01), *Watching TV* (LL01), and *General computer use* (LL03) are generally different by gender. *Gardening, ponds, pools, and hot tubs* (BB06), *Care for animals and pets* (BB07), *Physical care for/helping adults* (CD02), *Work for job(s)/research/homework* (EF01), and *General computer use* (LL03) are not significantly different by gender, which means that these activities are more similar between male and female. In sum, gender has significant influences on the difference of activities.

Table 3-16. Differences in Activities by Gender

Act.	Freq	Dur/a	Dur/d	Start	End	Partner
AA01	-6.67 (0.00)*	-15.08 (0.00)*	-18.96 (0.00)*	-4.79 (0.00)*	-5.34 (0.00)*	n/a (n/a)
BB01	-4.26 (0.00)*	-0.69 (0.49)	-2.08 (0.04)*	4.56 (0.00)*	4.25 (0.00)*	2.25 (0.02)*
BB02	-3.80 (0.00)*	0.17 (0.87)	-1.78 (0.07)	2.43 (0.02)*	2.68 (0.01)*	-2.19 (0.03)*
BB03	-11.74 (0.00)*	-2.05 (0.04)*	-8.18 (0.00)*	-3.41 (0.00)*	-3.81 (0.00)*	-3.09 (0.00)*
BB04	-3.16 (0.00)*	-0.30 (0.76)	-2.01 (0.04)*	0.44 (0.66)	0.83 (0.41)	3.31 (0.00)*
BB05	-0.07 (0.95)	2.87 (0.01)*	2.59 (0.01)*	1.61 (0.11)*	2.29 (0.02)*	1.93 (0.06)
BB06	1.84 (0.07)	6.35 (0.00)*	6.25 (0.00)*	-1.82 (0.07)	0.67 (0.50)	-18 (0.08)
BB07	-1.74 (0.08)	1.55 (0.12)	0.15 (0.88)	0.75 (0.45)	1.09 (0.28)	0.06 (0.95)
BB08	2.19 (0.03)*	2.46 (0.01)*	2.95 (0.00)*	-0.78 (0.44)	0.08 (0.93)	-1.83 (0.07)
CD01	-8.94 (0.00)*	0.29 (0.77)	-6.88 (0.00)*	4.59 (0.00)*	4.32 (0.00)*	-1.81 (0.07)
CD02	-0.61 (0.54)	1.78 (0.08)	1.91 (0.06)	-0.21 (0.83)	-0.12 (0.90)	-0.33 (0.74)
EF01	0.96 (0.34)	-0.72 (0.47)	-0.38 (0.70)	-1.02 (0.31)	-1.25 (0.21)	-2.38 (0.02)*
LL01	2.07 (0.04)*	7.51 (0.00)*	8.64 (0.00)*	-4.61 (0.00)*	-1.88 (0.06)	-0.60 (0.55)
LL02	0.72 (0.47)	1.98 (0.05)*	2.53 (0.01)*	-0.48 (0.63)	1.27 (0.21)	0.03 (0.97)
LL03	2.52 (0.01)*	6.71 (0.00)*	7.31 (0.00)*	-1.02 (0.31)	-0.18 (0.86)	-0.83 (0.40)

*** t-test: t-value (p-value), * denotes significant p-value*

3.4.2.4. Activities by Job Status

Start Time and *End Time* of energy usage-related activities more differ by job status, *Duration, Frequency* and *Partner* less differ by job status. It infers that the energy strategies need to consider the energy usage-related activities on job status more focusing on when those activities start and end. In Figure 3-21, 3-22, 3-23, 3-24, 3-25 and 3-26, *Yes* indicates the occupants who have a job, and *No* indicates the occupants who have no job.

Frequency (Figure 3-21) of *Physical care for children* (CD01) by the occupants who have no job is the highest (2.99 times) among all activities, and *Frequency* of *Vehicle repair and maintenance (BB08)* by the occupants who have no job is the lowest (1.19 times). *Frequency* of *Physical care for children* (CD01), *Physical care for/helping adults* (CD02), *Work for job(s)/research/homework* (EF01) and *Watching TV* (LL01) show greater differences between the occupants who have a job and who have no job, which means that these activities need different energy strategies regarding *Frequency* depending on the occupants' job status. The rest of the activities are similar, which

means that energy strategies for these activities can be similar regarding *Frequency* for the occupants who have a job and who have no job.

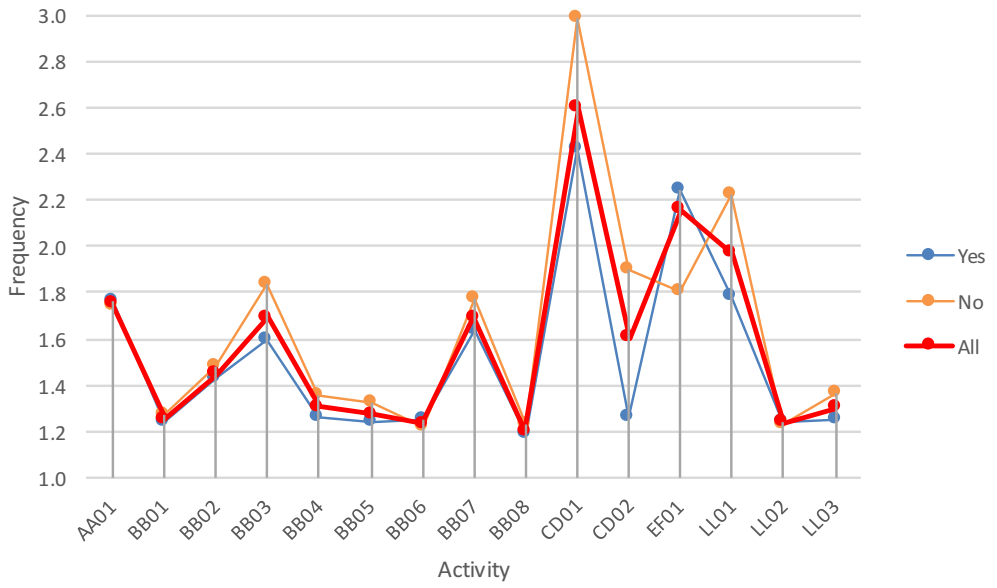


Figure 3-21. Comparison of Frequency by Job Status

Duration per act (Figure 3-22) of *Watching TV* (LL01) by the occupants who have no job is the longest (132.77 minutes) among all activities, and *Care for animals and pets* (BB07) by the occupants who have a job is the shortest (15.79 minutes). *Duration per act* of *Heating and cooling* (BB05), *Watching TV* (LL01), *Listening to/playing radio or music* (LL02), *General computer use* (LL03) show greater differences between the occupants who have a job and who have no job compared to other activities, which means that these activities need different energy strategies regarding *Duration per act* depending on the occupants' job status. The rest of the activities are similar, which means that energy strategies for these activities can be similar regarding *Duration per act* for the occupants who have a job and who have no job.

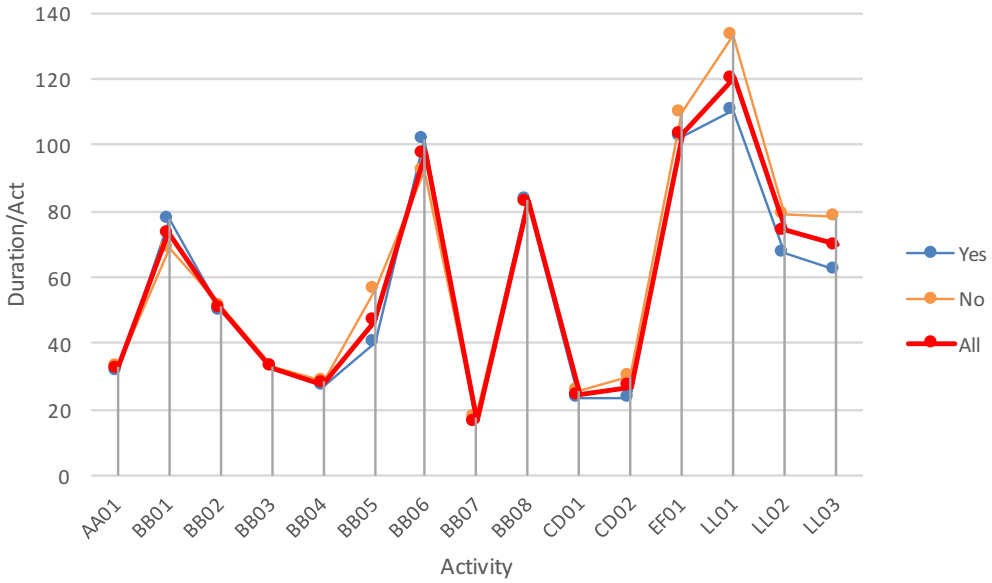


Figure 3-22. Comparison of Duration per Act by Job Status

Duration per day (Figure 3-23) of Watching TV (LL01) by the occupants who have no job is the longest (271.22 minutes) among all activities, and Care for animals and pets (BB07) by the occupants who have a job is the shortest (24.08 minutes). *Duration per day* of *Physical care for/helping adults* (CD02), *Watching TV* (LL01), and *General computer use* (LL03) show greater differences between the occupants who have a job and who have no job compared to other activities, which means that these activities need different energy strategies regarding *Duration per day* depending on the occupants' job status. The rest of the activities are similar, which means that energy strategies for these activities can be similar regarding *Duration per day* for the occupants who have a job and who have no job.

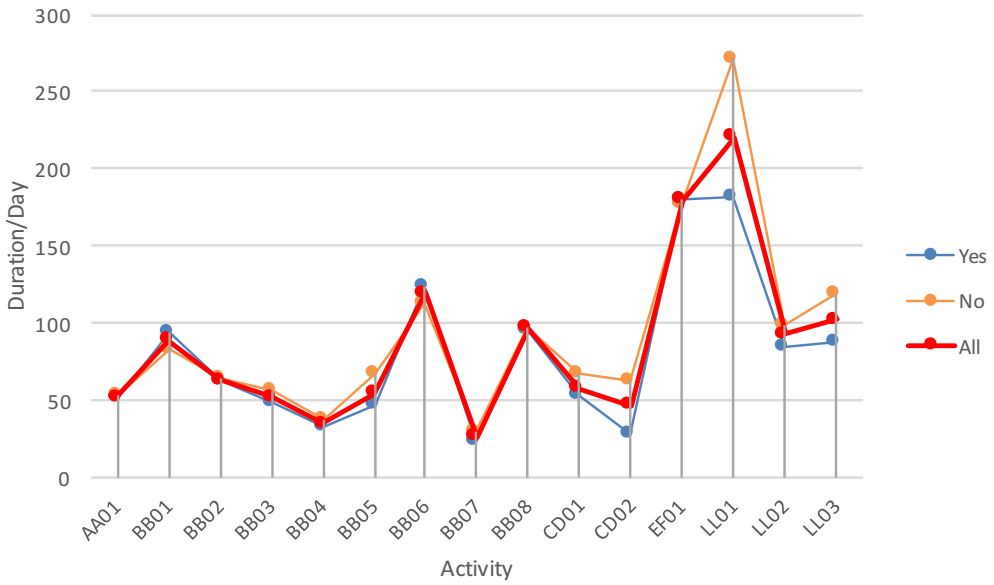


Figure 3-23. Comparison of Duration per Day by Job Status

Start Time (Figure 3-24) and *End Time* (Figure 3-25) of *Heating and cooling* (BB05) by the occupants who have no job are the earliest (*Start Time* 11:10 – 670.64 in minutes, *End Time* 12:06 – 726.77 in minutes) among all activities. *Start Time* and *End Time* of *Watching TV* (LL01) by the occupants who have a job are the latest (*Start Time* 17:18 – 1038.95 minutes, *End Time* 17:57 – 1077.63 minutes). *Start Time* and *End Time* of *Interior cleaning* (BB01), *Laundry* (BB02), *Heating and cooling* (BB05), *Physical care for children* (CD01), *Watching TV* (LL01), and *General computer use* (LL03) show greater differences between the occupants who have a job and who have no job compared to other activities, which means that these activities need different energy strategies regarding *Start Time* and *End Time* the occupants who have a job and who have no job. The occupants who have a job start and end these activities later than the occupants who have no job. The rest of the activities are similar, which means that energy strategies for these activities can be similar regarding *Start Time* and *End Time* for the occupants who have a job and who have no job.

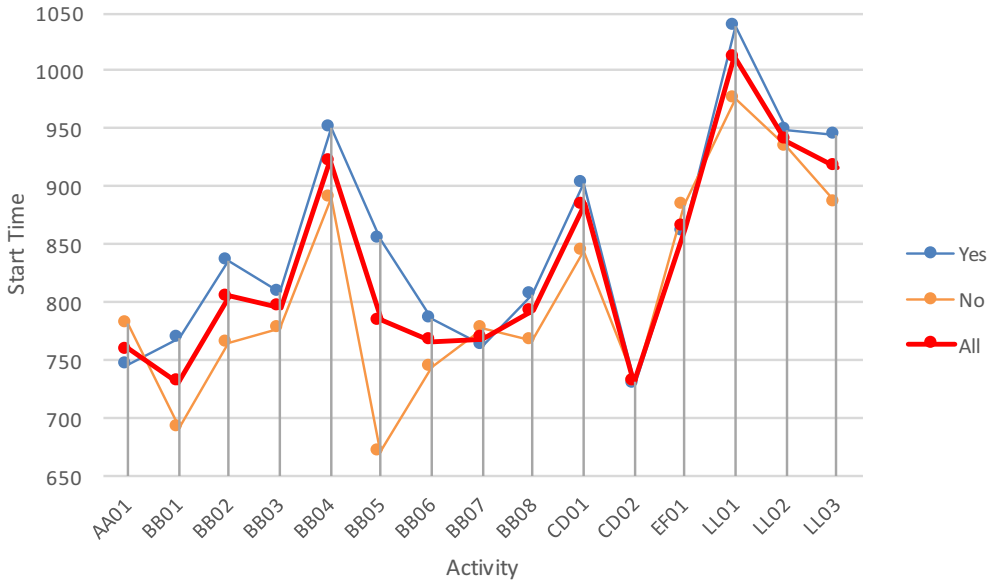


Figure 3-24. Comparison of Start Time by Job Status

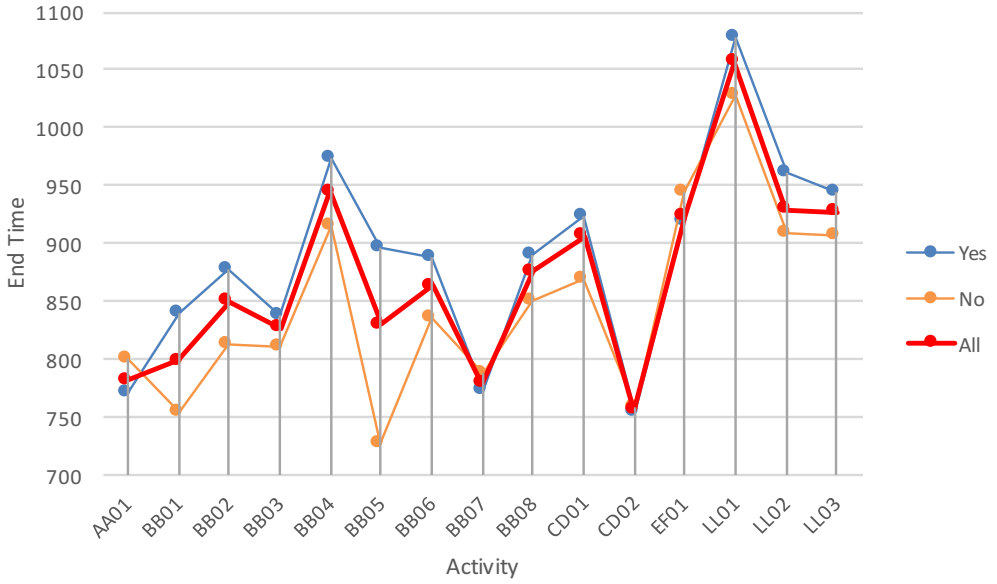
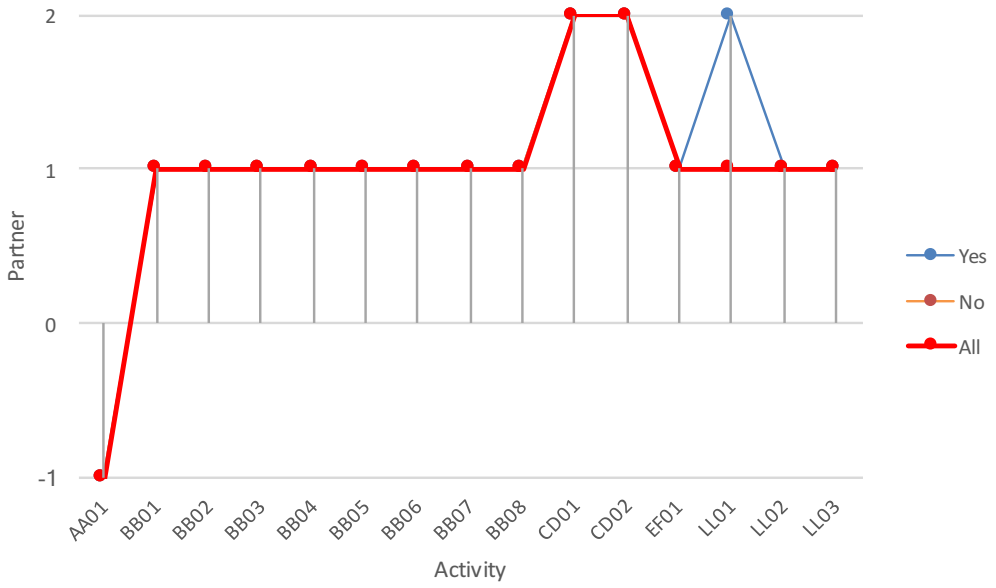


Figure 3-25. Comparison of End Time by Job Status

Generally, most of the activities by the occupants who have a job and who have no job have same mode values (Figure 3-26). However, *Partner* of *Watching TV* (LL01) shows differences depending on the occupants' job status, which means that this activity needs different energy strategies regarding *Partner* for the occupants who have a job and who have no job. *Partner* of the

rest of the activities are same, which means that energy strategies for these activities can be similar regarding *Partner* for the occupants who have a job and who have no job.



** *Partner Code: Not Recorded (-1), Alone (1), with Family (2)*

Figure 3-26. Comparison of Partner by Job Status

Table 3-17 summarizes the difference of activities by job status which is derived from the *t-test*. In general, *Start Time* and *End Time* more affect on the differences of the energy usage-related activities by the occupants who have a job and who have no job, and the rest of the activities less affect on them. Considering all the given six factors, *Washing, dressing, and grooming* (AA01), *Interior cleaning* (BB01), *Food and drink preparation* (BB03), *Kitchen and food clean-up* (BB04), *Gardening, ponds, pools, and hot tubs* (BB06), *Physical care for children* (CD01), *Watching TV* (LL01), and *General computer use* (LL03) are generally different by occupants' job status. *Laundry* (BB02), *Heating and cooling* (BB05), *Care for animals and pets* (BB07), *Vehicle repair and maintenance* (BB08), *Physical care for/helping adults* (CD02), *Work for job(s), research/ homework* (EF01), and *Listening to/playing radio or music* (LL02) are not significantly different by occupants' job status, which means that these activities are more similar between the occupants

who have a job and who have no job. In sum, job status has significant influences on the difference of activities.

Table 3-17. Differences in Activities by Job Status

Act.	Freq	Dur/a	Dur/d	Start	End	Partner
AA01	-1.42 (0.16)	3.43 (0.00)*	2.86 (0.00)*	5.26 (0.00)*	4.69 (0.00)*	n/a (n/a)
BB01	1.37 (0.17)	-3.29 (0.00)*	-3.04 (0.00)*	-8.21 (0.00)*	-9.21 (0.00)*	-4.78 (0.00)*
BB02	1.42 (0.16)	0.73 (0.47)	0.91 (0.36)	-6.22 (0.00)*	-5.84 (0.00)*	0.01 (0.99)
BB03	9.90 (0.00)*	0.20 (0.84)	6.69 (0.00)*	-4.75 (0.00)*	-4.35 (0.00)*	-6.54 (0.00)*
BB04	4.23 (0.00)*	1.87 (0.06)	3.57 (0.00)*	-5.82 (0.00)*	-5.58 (0.00)*	-5.49 (0.00)*
BB05	0.65 (0.52)	0.88 (0.38)	1.09 (0.28)	-2.60 (0.01)*	-2.33 (0.02)*	-0.85 (0.40)
BB06	-0.88 (0.38)	-1.85 (0.06)	-1.64 (0.10)	-3.25 (0.00)*	-4.18 (0.00)*	-2.10 (0.04)*
BB07	2.05 (0.04)*	1.50 (0.13)	2.99 (0.00)*	0.77 (0.44)	0.91 (0.37)	-1.20 (0.23)
BB08	0.69 (0.49)	-0.07 (0.94)	0.12 (0.90)	-1.17 (0.24)	-1.19 (0.23)	0.35 (0.73)
CD01	5.29 (0.00)*	2.19 (0.03)*	5.04 (0.00)*	-4.08 (0.00)*	-3.71 (0.00)*	2.38 (0.02)*
CD02	2.74 (0.01)*	1.12 (0.26)	3.01 (0.00)*	0.02 (0.98)	0.07 (0.94)	0.63 (0.53)
EF01	-5.00 (0.00)*	1.22 (0.22)	-0.18 (0.85)	1.21 (0.22)	1.30 (0.19)	-0.27 (0.79)
LL01	18.22 (0.00)*	10.76 (0.00)*	23.50 (0.00)*	-12.80 (0.00)*	-7.39 (0.00)*	-10.88 (0.00)*
LL02	-0.13 (0.90)	1.51 (0.13)	1.38 (0.17)	-0.34 (0.73)	-1.14 (0.26)	-2.72 (0.01)*
LL03	4.50 (0.00)*	6.13 (0.00)*	7.34 (0.00)*	-4.86 (0.00)*	-2.86 (0.00)*	-2.06 (0.04)*

*** t-test: t-value (p-value), * denotes significant p-value*

In sum, day of the week, gender, and job status have strong influences on the difference of energy usage-related activities regarding *Frequency, Duration, Start Time, End Time,* and *Partner*. Region has a less influence on the activities compared to other factors. However, four census regions are used in this section, since it might be too complicated to use more detailed geographical levels of locations (such as state or county level) for comparative analysis (50 or more geographical locations for each of 15 energy usage-related activities). Therefore, most habitual activities are selected and each of them is further examined by States in the next section.

3.4.3. Spatial Analysis for Habitual Energy Usage-Related Activities

Habitual energy usage-related activities are compared in this section in order to identify similarities and differences of the activities by States. The GIS Grouping Analysis and map visualization are used to find how geographical location affects the characteristic (*Frequency, Duration, Start Time, End Time, Partner*) of the activity.

The predictability of activities was examined using the Occupant Behavior Prediction Model in the previous study (Mo, 2018)(Chapter 2). Table 3-5 summarizes the top five most predictable and habitual energy usage-related activities (AA01: *Washing, dressing, and grooming*, LL01: *Watching television*, CD01: *Physical care for children*, BB03: *Food and drink preparation*, BB04: *Kitchen and food clean-up*). For these activities, the Grouping Analysis using K-means clustering is performed, and mean values of *Frequency, Duration, Start Time, End Time*, and *Partner* for each activity were compare by States. Among the activities, *Watching TV* (LL01) will be explained in detail in this section, since it shows more distinctive differences by States. Also, occupants generally spend longer time for *Watching TV* (LL01) than other four activities (Section 3.4.1 and 3.4.2), which means that *Watching TV* (LL01) more influences on occupant behavior and energy consumption. All of the GIS maps for the top five most predictable and habitual energy usage-related activities are listed in Appendix.

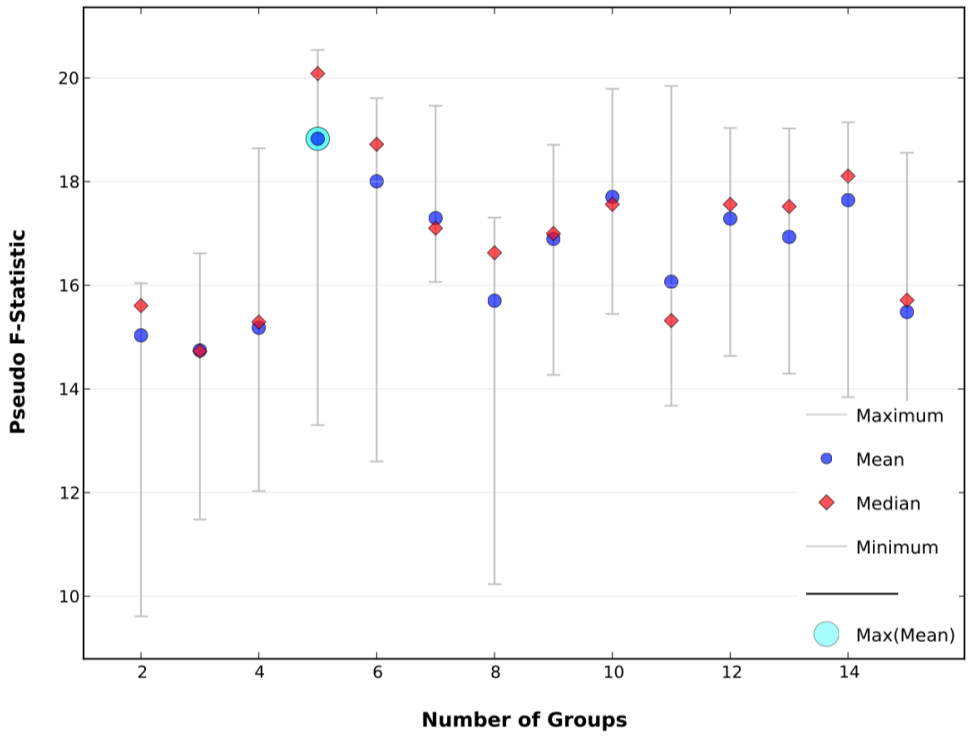


Figure 3-27. LL01 Number of K

Before running K-means clustering, pseudo F-statistic is calculated and 5 is the most suitable number of cluster for Watching TV activity as shown in Figure 3-27.

Figure 3-28 explains the characteristics of the clusters. The boxplots represent the standardized values of the features, and dots stand for the mean values of the clusters. Figure 3-29 displays the clusters and the States on the map.

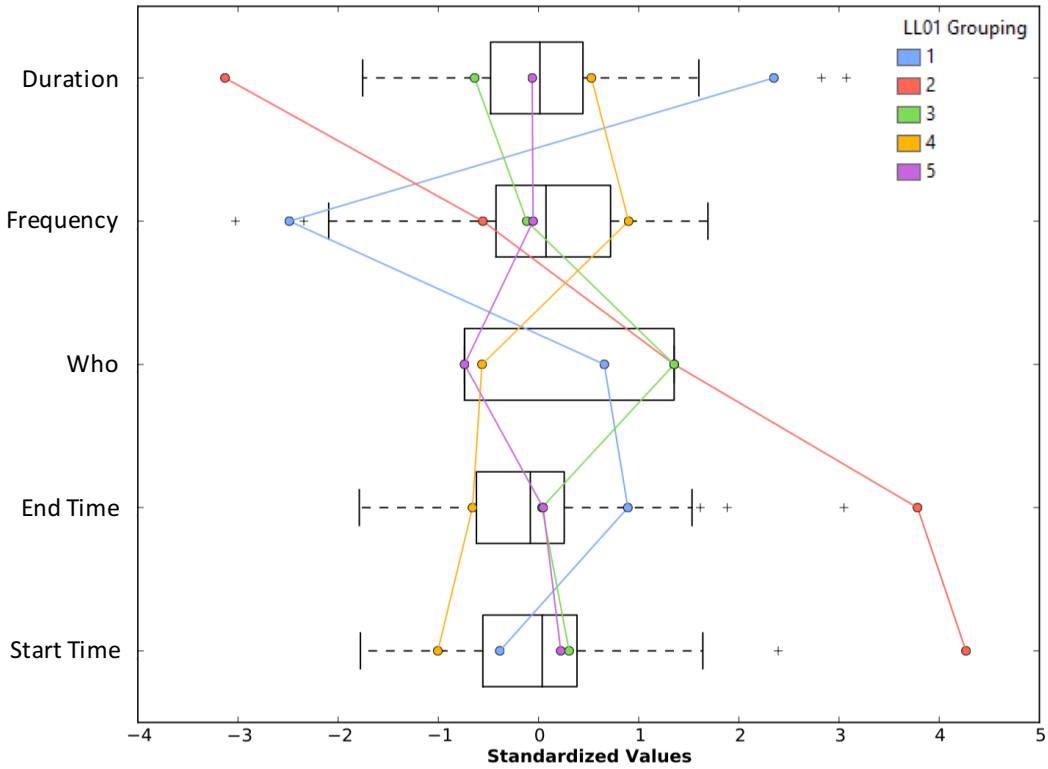


Figure 3-28. LL01 Group Analysis

- ***State Cluster 1 for LL01:*** The occupants in the States in Cluster 1 watch TV least frequently, but they spend the longest time a day compared to other clusters. They tend to watch TV more with family, and start to watch TV relatively early and end relatively late. Cluster 1 has Montana, Wyoming and Alaska which are located in colder area.

- ***State Cluster 2 for LL01:*** The occupants in the States in Cluster 2 watch TV relatively less frequently for the shortest time. They watch TV with family starting and ending latest. Cluster 2 has Hawaii.
- ***State Cluster 3 for LL01:*** The occupants in the States in Cluster 3 are in the middle for *Frequency* and *Duration* among the clusters. They watch TV with family and start relatively late and end relatively early. Cluster 3 has Oregon, Idaho, Utah, Arizona, Colorado, New Mexico, Kansas, Oklahoma, Missouri, Tennessee, North Dakota, South Dakota, Vermont, and Delaware.
- ***State Cluster 4 for LL01:*** The occupants in the States in Cluster 4 watch TV most frequently during relatively long time. They tend to watch TV alone starting and ending earliest among the clusters. Cluster 4 has Nevada, Wisconsin, Michigan, Ohio, Kentucky, West Virginia, Arkansas, Louisiana, Mississippi, Alabama, and Rhode Island.
- ***State Cluster 5 for LL01:*** The occupants in the States in Cluster 5 are in the middle for *Frequency, Duration, Start Time,* and *End Time*. They watch TV alone. Cluster 5 has Washington, California, Minnesota, Nebraska, Iowa, Illinois, Indiana, Texas, New Hampshire, Massachusetts, Maine, Connecticut, New York, New Jersey, Pennsylvania, Maryland, Virginia, North Carolina, South Carolina, Georgia, and Florida.

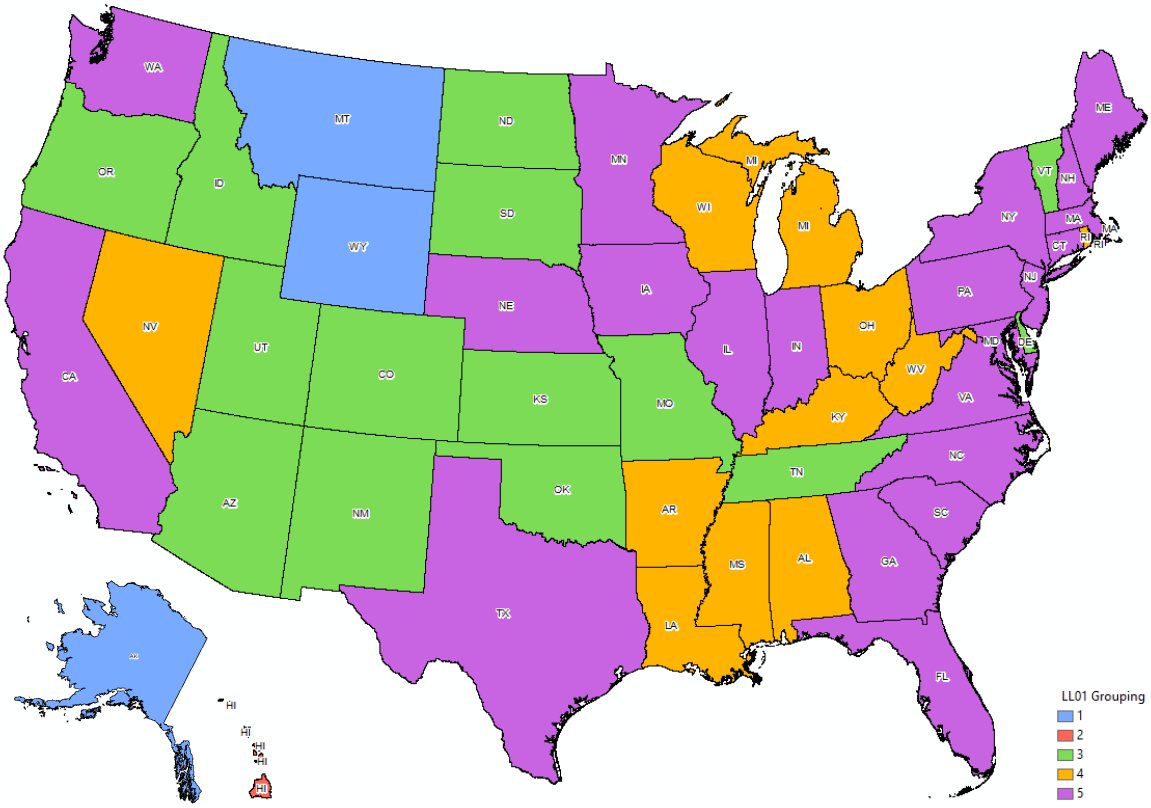


Figure 3-29. LL01 State Clusters by Grouping Analysis

In Figure 3-29, geographical distribution of the clusters shows some patterns with cold area (Cluster 1), Hawaii (Cluster 2), central area (Cluster 3), central east area (Cluster 4), and coastal, central north and central south areas (Cluster 5). Cluster 2 has only Hawaii, and it explains that the geographical location of Hawaii strongly influences the different pattern of watching TV in this location compared to other States. Hawaii is remote from the other continental States, and has unique climate, culture, economy, industry, life style and so on. These differences might affect the difference in watching TV. Each state can consider the characteristic of the cluster where the state belongs to for its energy control strategies and policy development.

While Figure 3-29 shows the result of the clustering analysis which integrated the effects of *Frequency, Duration, Start Time, End Time,* and *Partner* together, Figure 3-30, 3-31, 3-32, 3-33

and 3-34 compare the mean values of individual components of watching TV by States with 5 quantiles. These maps more simply and directly explain the difference of each component by States. Darker colors indicate more *Frequency*, longer *Duration*, later *Start Time and End Time*., and 20% of the States are indicated with the same color. Figure 3-34 compares the mode value of *Partner*, and light color indicates watching TV more alone, and dark color indicates more with family. Since Partner has two values (1: Alone, 2: with Family), the map has only 2 color levels. One of the possible explanation is that occupants in the agricultural states tend to watch TV with family and they have more flexible time to watch TV compared to other states.

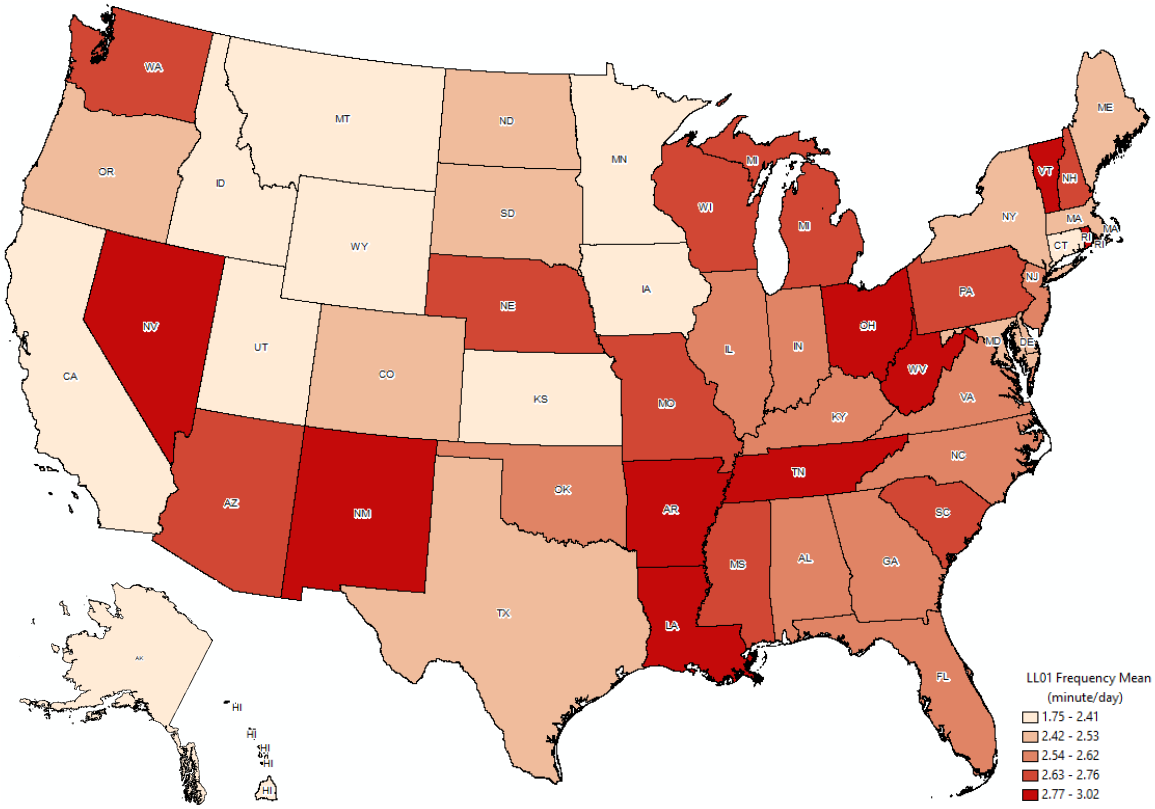


Figure 3-30. LL01 Frequency by Quantiles

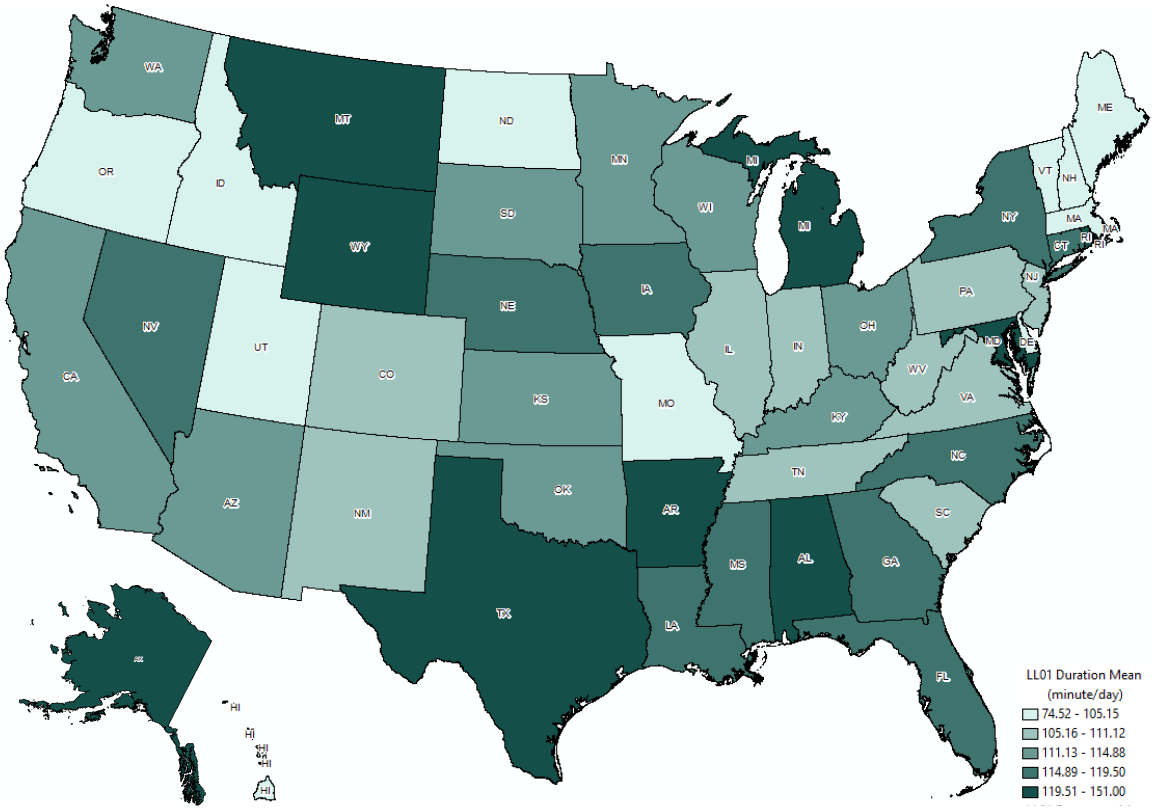


Figure 3-31. LL01 Duration by Quantiles

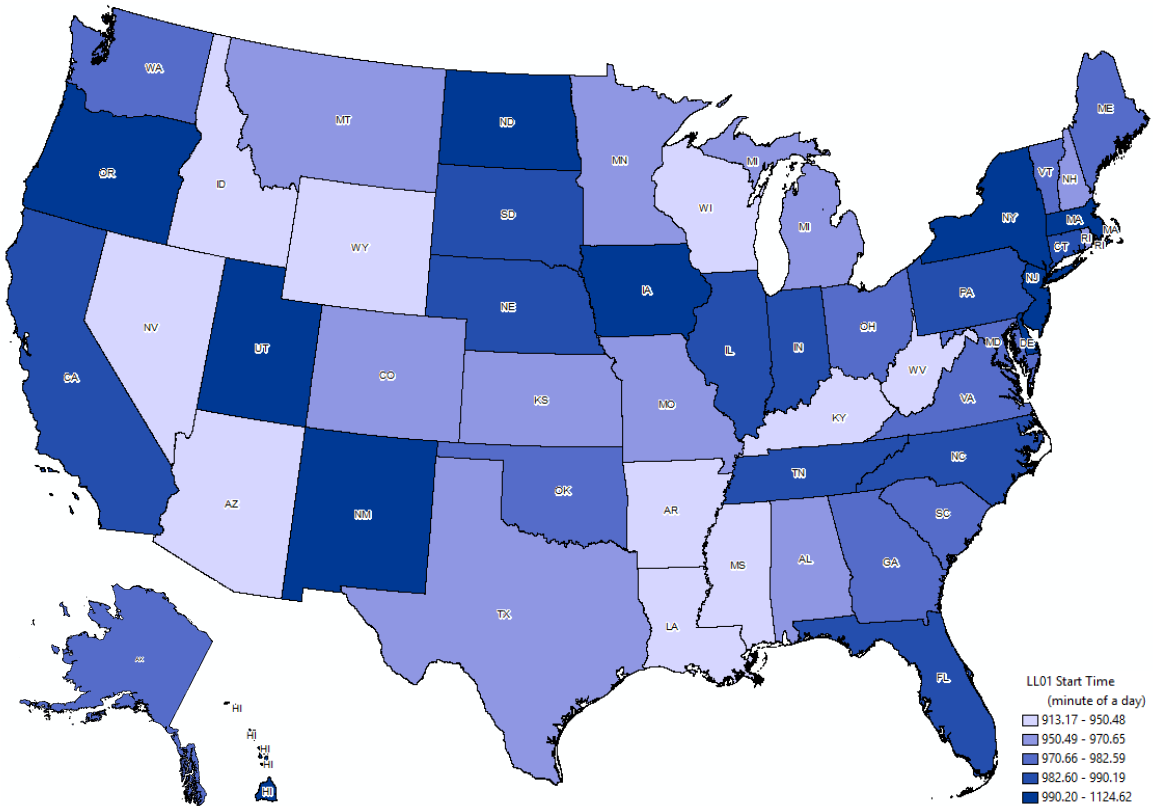


Figure 3-32. LL01 Start Time by Quantiles

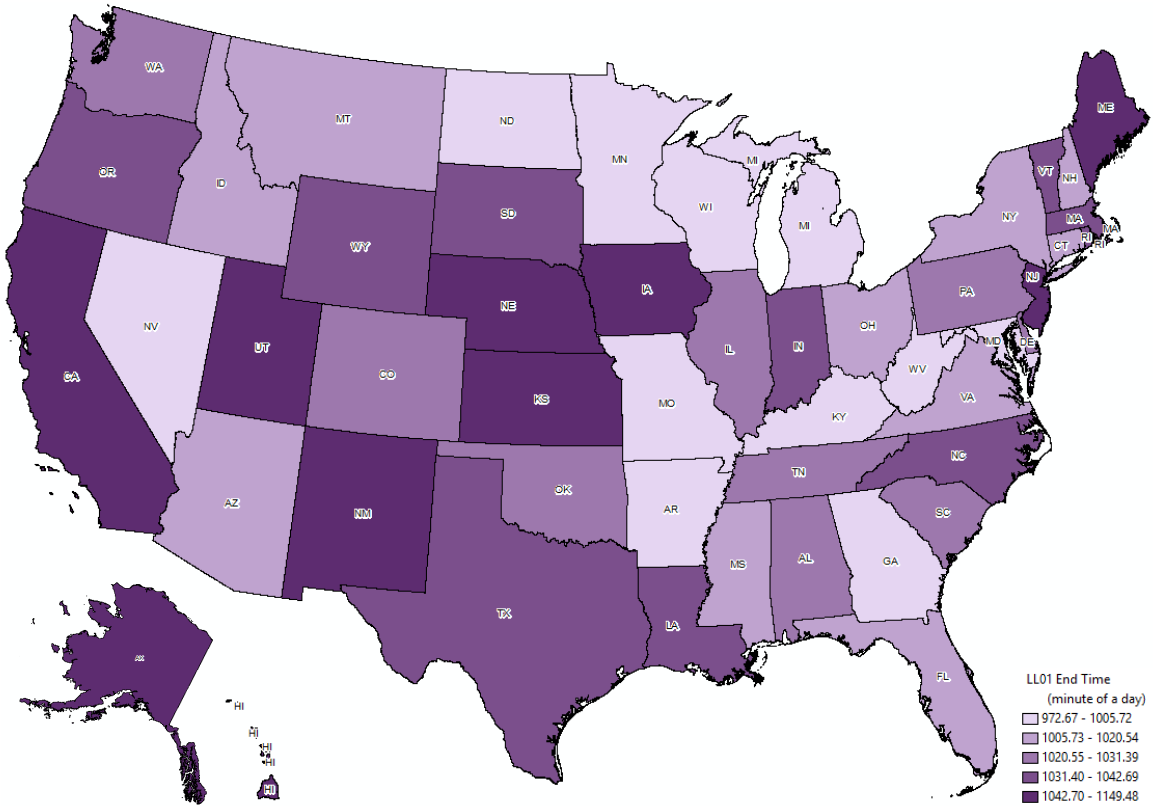


Figure 3-33. LL03 End Time by Quantiles

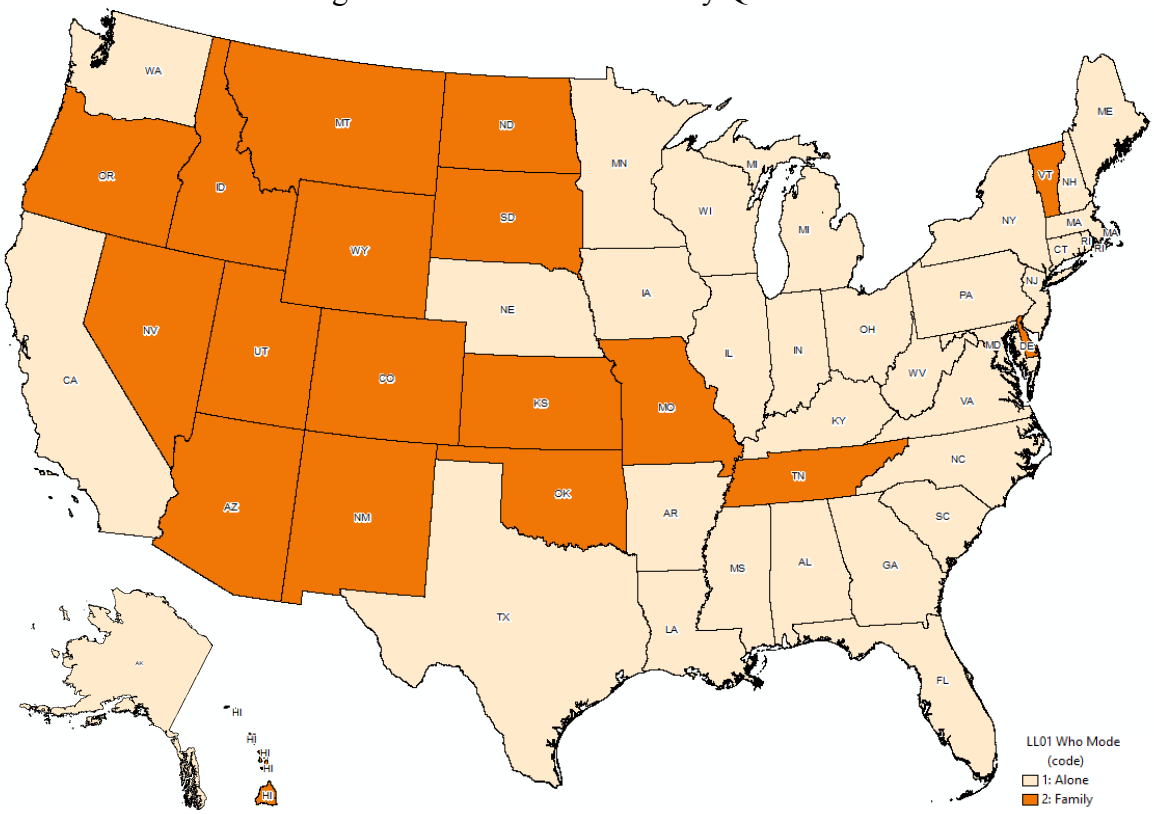


Figure 3-34. LL01 Partner

Frequency and *Duration* of *Watching TV* (LL01) is more directly related to energy consumption, and *Start Time, End Time* and *Partner* explain the life style of occupants in the State. In Figure 3-31, occupants in Alaska, Montana, Wyoming, Texas, Arkansan, Michigan, Alabama, and Maryland spend the longest time watching TV (119.51-151.00 minutes per day).

Other four habitual energy usage-related activities are analyzed in a same way to analyze *Watching TV* (LL01), using the ArcGIS Grouping Analysis with K-means clustering and mean/mode comparison, and the results are displayed from Figure 3-35 to Figure 3-38.

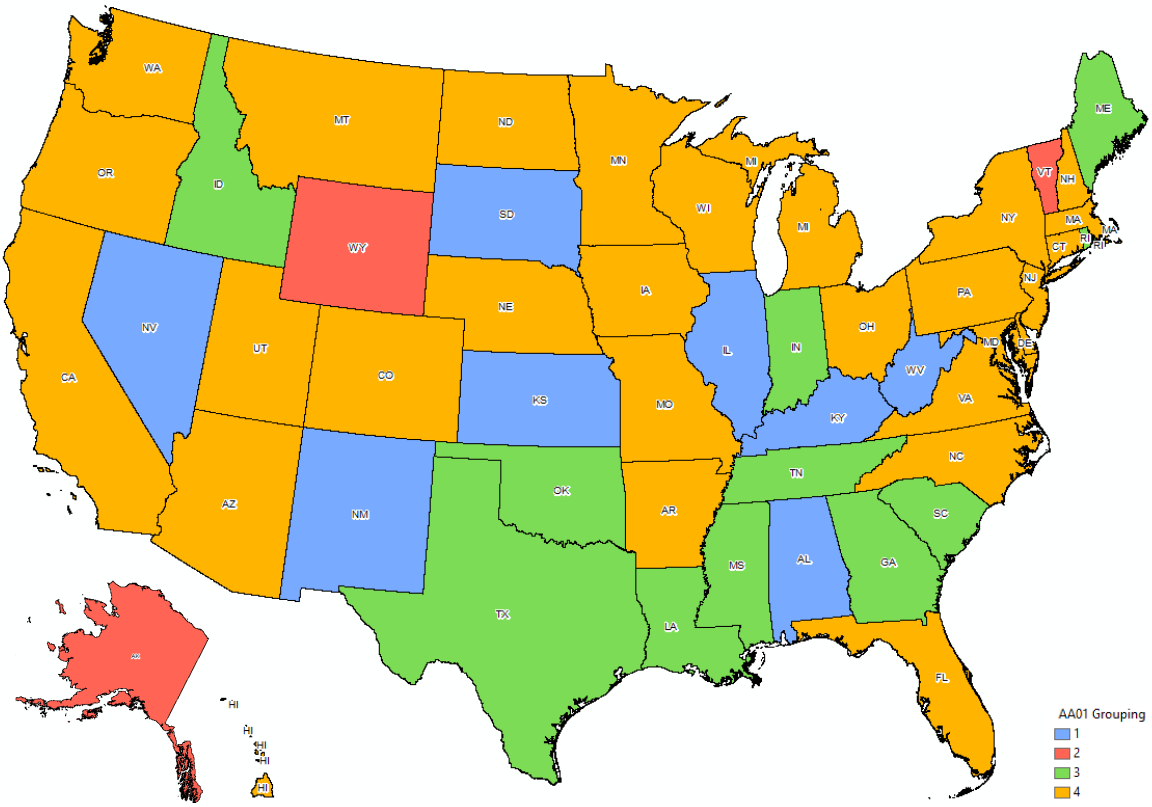


Figure 3-35. AA01 State Clusters by Grouping Analysis

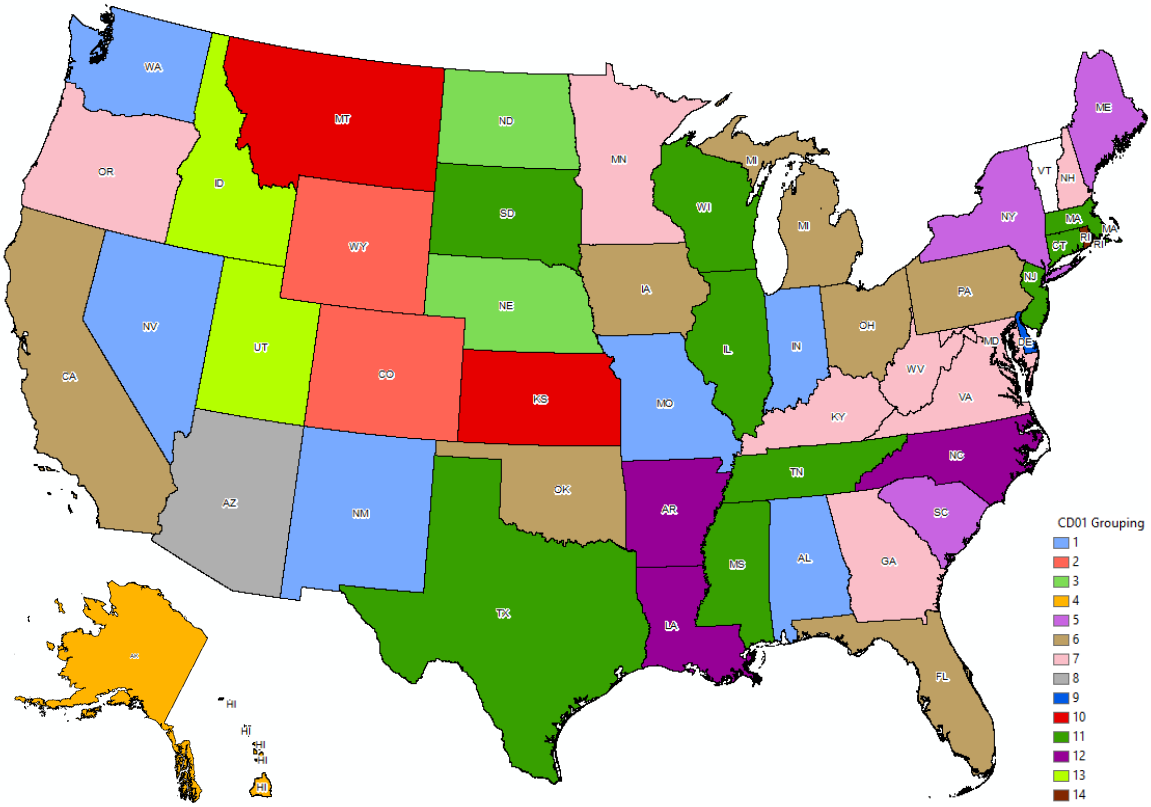


Figure 3-36. CD01 State Clusters by Grouping Analysis

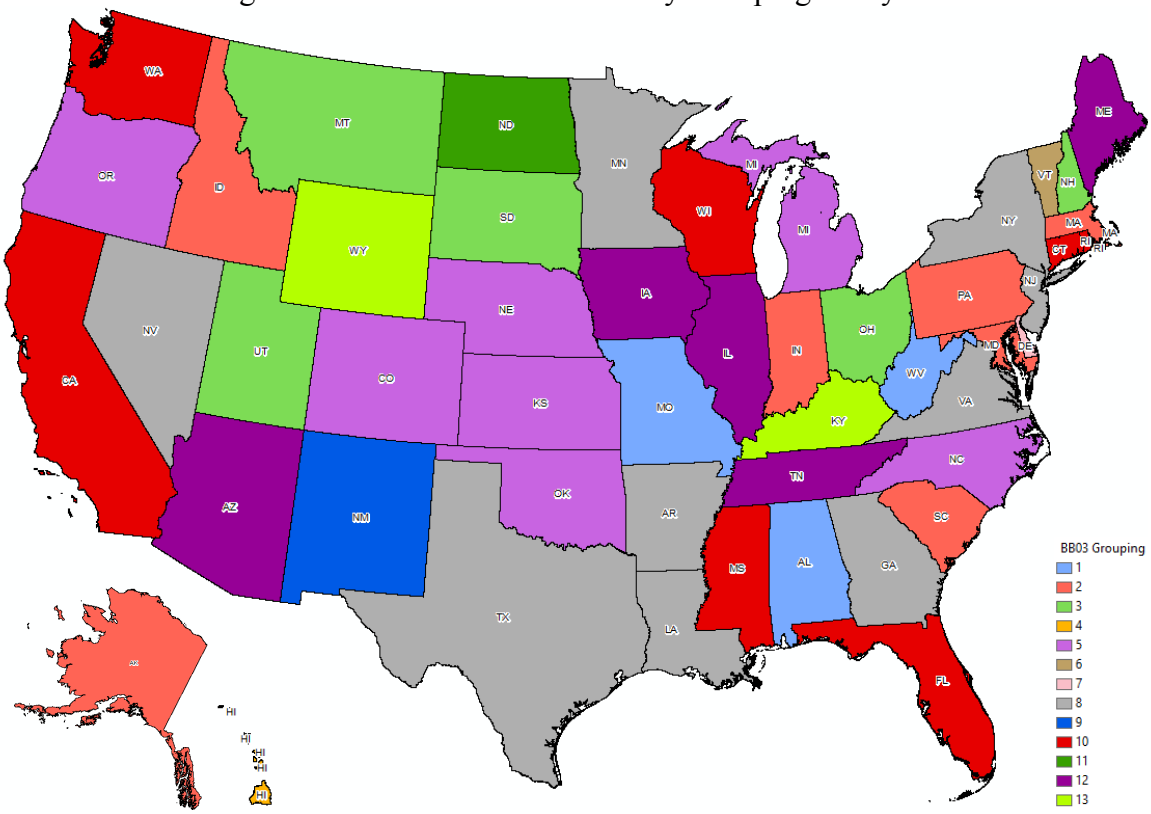


Figure 3-37. BB03 State Clusters by Grouping Analysis

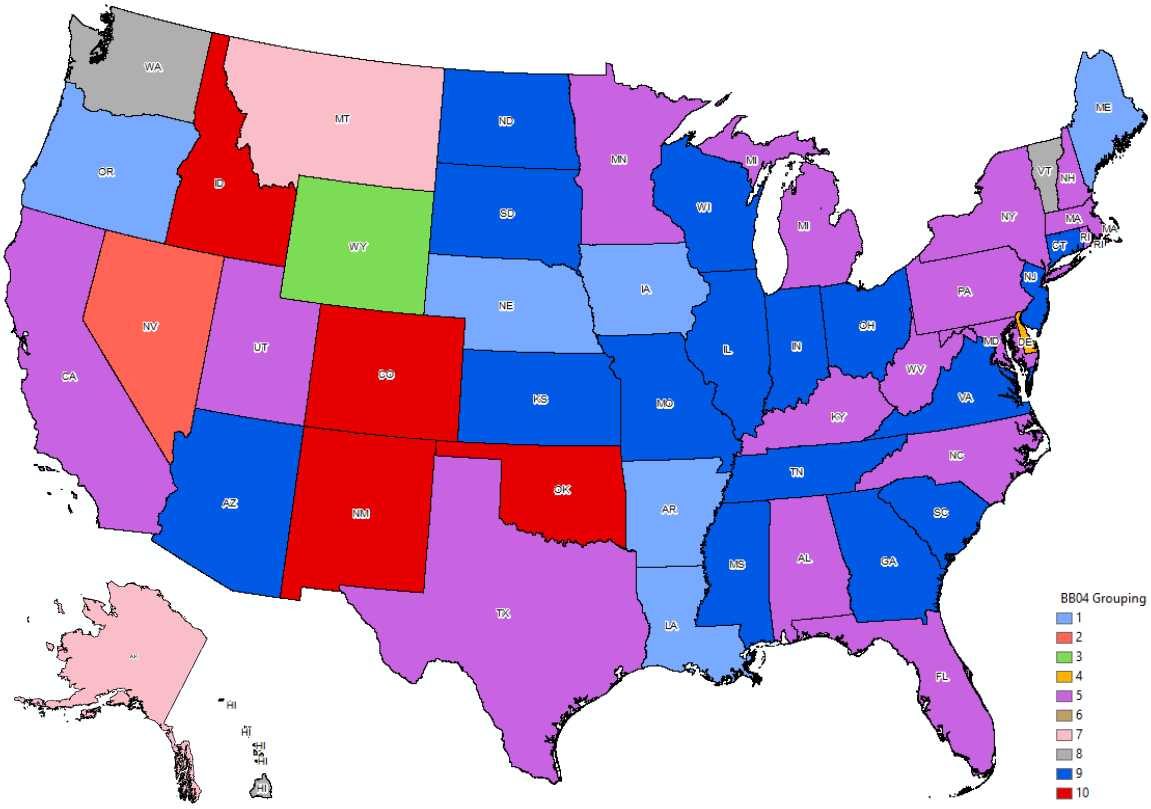


Figure 3-38. BB04 State Clusters by Grouping Analysis

Spatial analysis explains the characteristic of energy usage-related activities more efficiently, and helps to compare the activities considering the geographical locations of the occupants. It also enables to consider other environmental, social, technological factors more effectively based on the locations.

3.5. Discussion and Conclusion

In this chapter, the daily routine of occupant activities is identified by different occupant groups, and the activities are analyzed by different perspectives, including comparative analysis with the regions, day of the week, gender, job status, and geographical analysis using GIS.

The findings include (1) which factors (four census regions, day of the week, gender, job status) more affect the similarities and differences of the energy usage-related activities, (2) which habitual components are more different by the factors, and (3) which energy usage-related activities are more different by the factors.

Television and movies (120303) is one of the most habitual energy usage-related activities, and it is included in 5 clusters: Based on the overlapping time from these 5 clusters, the occupants watch TV around from 18:30 to 21:30, and it means that one of the most habitual energy usage-related activities strongly tend to happen during this time. Based on the results summarized in Table 3-14, Table 3-15, Table 3-16, and Table 3-17, day of the week, gender, and job status strongly affect on the difference of the activities. Regions less affect on them compared to other factors, but it needs further examination with more specific area division. *Frequency, Duration, Start Time* and *End Time* values generally vary by the factors, but *Partner* values vary less by the factors. *Washing, dressing, and grooming* (AA01)*, Interior cleaning* (BB01)*, Laundry* (BB02)*, Food and drink preparation* (BB03)*, Kitchen and food clean-up* (BB04)*, Gardening, ponds, pools, and hot tubs* (BB06)*, Physical care for children* (CD01)*, Watching TV* (LL01), and *General computer use* (LL03) more differ by the factors, and *Heating and cooling* (BB05)*, Care for animals and pets* (BB07)*, Vehicle repair and maintenance* (BB08)*, Physical care for/helping adults* (CD02)*, Work for job(s)/research/homework* (EF01), and *Listening to/playing radio or music* (LL02) less differ by the factors.

The findings show that when we analyze occupant behavior, proper occupant grouping or segmentation is important to understand their behavior more effectively. Since the occupant

activities and behaviors are different by day of the week, gender, and job status, these factors should be considered for daily behavior analysis.

The differences of activities can be more efficiently explained using GIS analysis. The geographical locations enable to associated other diverse factors more effectively. The habitual activities of occupants can be influenced not only by internal factors of the occupants, such as age, gender, job, income, education, number of family, etc., but also by external factors including climate, economy, industry, policies, building technology and more of the location. The finding also shows the habitual activities vary by different geographic location, although they are persistent over time in a same location. GIS analysis effectively connects these factors by providing the geographical context of the information.

The ATUS data are mainly used in this chapter. However, there exist some limitations. The ATUS data records one activity at a time by one representative of the household, while multiple activities are occurred simultaneously by multiple household members in real life. In the future, more accurate and realistic dataset can realize more refined analysis about the residential energy and behaviors.

The result can be used to provide more reliable information regarding energy and behavior to the occupants in residential buildings. Also, the result can be applied to the new energy and behavior strategies and policies about residential building energy plans. In addition, the geographical comparison using GIS and the grouping analysis can be used to develop more efficient strategies by different location of the residential buildings.

CHAPTER 4

EFFECTIVE FACTORS TO PREDICT RESIDENTIAL ENERGY CONSUMPTION USING MACHINE LEARNING

Abstract

Individual humans have a greater influence on energy consumption in residential buildings than other types of buildings. Although existing studies focus on how energy consumption is affected by building technologies and occupant demographic information, few studies have incorporated the impact of occupant energy use patterns. The goal of this study is to identify the factors that affect energy consumption in residential buildings and to measure their predictive performance. The researchers examined the impact of occupant energy use behaviors and the energy use patterns of home appliances on home energy consumption. The patterns include the combination of appliances, their use times and frequencies, and the configurations set by users. Data from the Residential Energy Consumption Survey (RECS) are analyzed to select features for prediction, using multiple machine learning algorithms including Support Vector Machine (SVM) and Random Forest. The results provide a list of factors that efficiently predict energy consumption in residential buildings. The selected 32 features achieve 98% of the prediction performance compared to the performance with all of 271 features. This list can be used to improve the effectiveness of energy saving programs and to educate occupants about their energy use patterns. The relationship between occupants' behavior patterns and energy use patterns revealed from this study provides the groundwork for researchers to further explore the prediction of occupant behavior from energy consumption.

4.1. Introduction

The residential sector accounts for 39% of the total electricity consumption in the U.S., according to the U.S. Department of Energy (DOE) (2017). Energy consumption in individual households depends on various factors, including environmental conditions, building technology, resident demographic information, Heating, Ventilation and Air Conditioning (HVAC) systems, appliances in the home, and the individual power rating of each appliance. Among these factors, the usage patterns of HVAC systems and appliances are more dependent upon occupant behavior, such as temperature settings, frequency and duration of uses, time of day they are used, etc.

Human factors have a greater impact on energy consumption in residential buildings than in other types of buildings. While the existing studies have focused more on the impact of appliances' technological characteristics on energy consumption, fewer studies have incorporated the impact of the appliances' usage patterns by occupants.

A comprehensive understanding of the main factors affecting household electricity consumption, including building technology, occupant demographic information, appliances, and occupant behavioral patterns, is needed to develop effective energy efficiency programs and provide relevant educational information to occupants. These factors will explain energy consumption more effectively, and their relationships can be used to uncover the energy consumption factors and behavioral patterns behind energy consumption data.

The goal of this chapter is to identify the main factors affecting electricity consumption in residential buildings and measure the predictive performance of these factors. In particular, behavior-related factors from appliances and their usage patterns are separately examined to see the effects of occupant behavior on energy consumption.

4.2. Background

National survey datasets are a good source of occupant behavior and building characteristics. One example is a study by Santin et al. (2009), which used national survey data from the Netherlands for 15,000 households, with questions about household characteristics and building attributes. In the United States, the Residential Energy Consumption Survey (RECS) and the American Time

Use Survey (ATUS) data are frequently used in research regarding occupant behavior and energy usage.

Sanquist et al. (2012) performed a lifestyle analysis of electricity consumption in residential buildings with a multivariate statistical approach using the 2005 RECS data. They identified five lifestyle factors associated with specific behavioral patterns: air conditioning, use of laundry machines, use of personal computers, climate zone, and use of TVs. These factors explain about 40% of variance in electricity consumption, and the explained variance is increased to about 54% by adding household and market characteristics, such as residents' income, access to natural gas, and local electricity prices.

Diao et al. (2017) identified and classified occupant behaviors with energy consumption outcomes. They extracted occupant features of five typical house types in New York state from the 2009 RECS data. The features included number of occupants, number of rooms, floor area, heated area, and number of windows in heated areas for the house types of single family (detached), single family (attached), apartment (2-4 units), apartment (5+ units), and mobile home. The features were applied to the behavior clusters from the ATUS by mapping the demographic information of the ATUS and the RECS.

Aksanli et al. (2016) developed a residential energy modeling framework based on human activities to estimate the energy consumption in residential buildings. They extracted appliance-related parameters from the RECS, including the types, numbers, and frequencies of usage, and associated them with specific actions and activities. They grouped the activities based on

occupants' demographic information, such as age, gender, employment status, and number of household members. They aimed to capture the energy use activities based on the probabilistic time-series nature which is dependent on demographic variables and time variables (time of a day, day of the week etc.).

Existing studies focused more on only parts of the RECS data, but this study will approach the variable analysis with a holistic view, incorporating the categories of electrical appliances at home, building technologies, occupant behaviors, and occupant demographic information. First, the critical factors for electricity consumption will be examined from all of the variables in the RECS data. Then, the factors will be analyzed within the individual categories of the RECS data, and the factors from behavior-related categories will be assessed.

4.3. Data

4.3.1. Overview of RECS Data

The Residential Energy Consumption Survey (RECS) is a national energy survey for residential buildings conducted by the Energy Information Administration (EIA) under the U.S. Department of Energy (US DOE). The survey has been conducted every three years since 1978 (Sanquist et al., 2012). The RECS mainly collects households' total energy consumption data over one year with energy fuel types, building geometry information, household demographics, and appliance information (Diao et al., 2017). Lifestyle patterns can be derived from a subset of the RECS variables. These variables include geographic location, household equipment and appliances, family structure, income, and local electricity price. (Sanquist et al., 2012).

The survey data were collected in 2015, which was the 14th iteration of the RECS program. Traditionally, the EIA used in-person interviews to collect the data, but they started to use online and mailed forms in addition to the in-person interviews in 2015. They combined these responses with data from the energy suppliers to these residential units to estimate the energy consumption and costs of appliances, heating, cooling, and other end uses (EIA, 2018).

4.3.2. Data Pre-Process

In this study, the latest 2015 RECS microdata file is used. It contains household characteristics, household energy insecurity data, and energy consumption and expenditures data from 5686 households. The original 2015 RECS microdata file has 736 features. After removing 309 features (imputation flags and replicate weights), the remaining 427 features are grouped by their characteristics as described in Table 4-1. As expected from the number of features, the RECS has detailed data about appliances, building technology, occupant demographics, and energy consumption information (kWh, Btu, Cost), but has less detailed data about occupant behavior.

Table 4-1. Categories of RECS Data

Category	Feature Examples	Count
ID	Unique identifier for each respondent	1
Appliance	Appliances, Lighting, Internet, Number, Size, Type, Age, Fuel type for appliances, Energy star appliances	81
Behavior	Frequency, Duration, Number of days/months used, Heating/cooling temperature set-point, Dishwasher, washer, dryer temperature and cycle setting, Smart meter data check	32
Technology	Building envelope, HVAC, Water heater, Fuel type for Tech, Thermostat, Light controller, Sensor, Smart meter install, Building audit, Pool, Hot tub	117
Demographic	Occupant/family characteristics, Who pays bill, Receive/participate in home energy assistance program	41
kWh	Electricity usage in kWh	27
Btu	Energy consumption in Btu, Conversion factor	57
Cost	Usage cost for electricity, propane, oil/kerosene	53
Other	Natural gas, Propane, Oil/kerosene information	18

Among the categories, 271 features from the Appliance, Behavior, Technology, and Demographic categories are used as input features in feature selection and machine learning algorithms to predict total electricity consumption in kWh, which is a numeric variable.

- ***Independent Variables (Xs):*** Features from Appliance, Behavior, Technology, and Demographic categories
- ***Dependent Variable (Y):*** Total electricity consumption in kWh

4.4. Methodology

The methodology follows the machine learning features selection and algorithm selection process. Features are selected from different categories, and the selected features are used to predict total energy consumption using various ML algorithms. The efficiency of the selected features is evaluated by comparing the prediction performance of the selected features and the prediction performance of all features together. The categories for feature selection and energy consumption prediction are as follows: (1) All, (2) Appliance, (3) Behavior, (4) Technology, (5) Demographic, and (6) Appliance + Behavior.

4.4.1. Feature Selection

Feature selection is the process of selecting a subset of features to be used for model construction in machine learning and statistics. It is also called attribute selection, variable selection, or variable subset selection (James, Witten, Hastie, & Tibshirani, 2013). It aims to find faster and more cost-effective predictors, improve the prediction performance of the predictors, and help researchers understand the underlying process better (Guyon & Elisseeff, 2003).

In this study, Correlation-based Feature Selection (CFS) with Greedy Stepwise method is used for the feature selection. It evaluates the worth of a subset of features by considering the single predictive ability of each feature and the degree of redundancy between them. Greedy Stepwise performs a greedy search forward or backward through the features subset. It stops when the addition or deletion of any remaining features results in a decreased performance evaluation (Hall, 1998).

First, feature selection is performed for all features from the Appliance, Behavior, Technology, and Demographic categories to identify the most critical and efficient features for predicting energy consumption among all possible features in the RECS data. Then, feature selection is performed for each category to find the most efficient features in each category. In case only limited features of data are available from the real-world datasets, this feature selection will be useful to find the most efficient factors for predicting energy consumption from a limited dataset. Finally, feature selection is performed for the combination of the Appliance and Behavior categories. Appliance and Behavior are determined by the occupants more than Technology or Demographic, and they reflect the behavioral patterns of the occupants. Thus, the effect of this combination of features is examined separately.

4.4.2. Algorithm Selection

To predict electricity energy consumption using the RECS features, Linear Regression, Support Vector Machine, Random Forest, M5P Trees, and M5 Rules are tested.

Linear Regression is simple and one of the most common algorithms for numeric prediction, so it is used as the baseline. However, if the data show a nonlinear dependency, the predictive line will not fit well. The predictive line is expressed as the linear combination of the features with pre-determined weights as follows (Witten et al., 2016).

$$x = w_0 + w_1 a_1 + w_2 a_2 + \ldots + w_k a_k$$

where x is the class, a_1, a_2, ..., a_k are the feature values, and w_0, w_1, ..., w_k are weights.

Sequential minimal optimization (SMO) regression implements the Support Vector Machine (SVM) for regression. It produces a model that can be expressed with support vectors and can be applied to nonlinear datasets using kernel functions (Witten et al., 2016). Its predictive performance is influenced by the kernels and parameter settings, and Radial Basis Function (RBK) kernel with C value 1 and gamma value 0.01 is used in this study.

Decision trees and rules work more naturally with nominal features, but they can be extended to numeric features by combining with numeric-value tests into the decision tree or rule-induction scheme, with pre-discretization of numeric features into nominal ones (Witten et al., 2016).

Random Forest is an ensemble learning method that builds a randomized decision tree in each iteration of the training, and outputs the mean prediction of the individual trees (Breiman, 2001; Witten et al., 2016).

M5P Trees combines a conventional decision tree with the possibility of linear regression at the nodes (Quinlan, 1992; Wang & Witten, 1996). M5 Rules generate a decision list for regression

problems with a divide-and-conquer approach by constructing a model tree using M5 and developing the best leaf into a rule in each iteration (Holmes, Hall, & Prank, 1999).

Correlation Coefficient and Root Mean Squared Error (RMSE) are used to evaluate performance. Correlation Coefficient measures the statistical correlation between the actual values and the predicted values. It ranges from -1 to 1, where 1 indicates perfect positive correlation and -1 indicates perfect negative correlation. 0 means that there is no correlation. It can be calculated as follows (Witten et al., 2016):

$$Correlation\ Coefficient = \frac{S_{pa}}{\sqrt{S_p S_a}}$$

where, $S_{pa} = \frac{\sum_{i=1}^{n}(p_i-\overline{p})(a_i-\overline{a})}{n-1}$, $S_p = \frac{\sum_{i=1}^{n}(p_i-\overline{p})^2}{n-1}$, $S_a = \frac{\sum_{i=1}^{n}(a_i-\overline{a})^2}{n-1}$, p_1, p_2, ..., p_n are the predicted values, a_1, a_2, ..., a_n are the actual values, $\bar{p}$ is the mean value over the predicted data, $\bar{a}$ is the mean value over the test data, and n is the number of data.

Root Mean Squared Error (RMSE) is the square root of Mean Squared Error (MSE), which is the principal and one of the most commonly used measures of performance. RMSE is always non-negative, and an RMSE of 0 would indicate a perfect fit. Lower values are better than higher values in general, but it is not a valid way to compare different types of data because RMSE is dependent on the scale of the numbers in a given dataset. It can be calculated as follows (Witten et al., 2016):

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(p_i - a_i)^2}{n}}$$

where, $p_1, p_2, ..., p_n$ are the predicted values, $a_1, a_2, ..., a_n$ are the actual values, and n is the number of data.

4.5. Result

4.5.1. Main Factors of Energy Consumption

The feature selection results suggest key features for predicting energy consumption. Among all 271 features, 32 features are selected using CFS with Greedy Stepwise method as summarized in Table 4-2. The features are selected from the Appliance, Behavior, Technology, and Demographic categories. From the Appliance category, the number and size of refrigerators and freezers, and the number of ovens, televisions, and ceiling fans are selected. This means that the usage patterns of these appliances are good predictors of the total electricity consumption in a residential building. From the Behavior category, duration of swimming pool usage, frequency of clothes dryer usage, and duration of TV usage on weekends are selected. From the Technology category, climate, the location and type of the house, and Heating, Ventilation and Air Conditioning (HVAC)-related features including fuel type are selected. From the Demographic category, the number of total household members and number of adults are selected. Two economic features were also selected from this category: if a household could not afford repair or replacement of broken cooling equipment, and the number of days covered by Energy Supplier Survey natural gas billing data. While these features indicate the economic status of the household, they indirectly provide cooling and heating information about the household as well. In summary, HVAC, refrigerator/freezer and TV, climate, location, house type, and number of household members are the main features for electricity consumption prediction in residential buildings.

Table 4-2. Selected Features from All

Factor	Description
Appliance	
NUMFRIG	Number of refrigerators used
SIZRFRI1	Size of most-used refrigerator
ICE	Through-the-door ice on most-used refrigerator
SIZFREEZ	Size of most-used freezer
OVEN	Number of separate ovens
TVCOLOR	Number of televisions used
NUMCFAN	Number of ceiling fans used
Behavior	
MONPOOL	Months swimming pool used in the last year
DRYRUSE	Frequency clothes dryer used
TVONWE1	Most-used TV usage on weekends
Technology	
UATYP10	Census 2010 Urban Type
TYPEHUQ	Type of housing unit
NCOMBATH	Number of full bathrooms
TOTROOMS	Total number of rooms in the housing unit, excluding bathrooms
UGASHERE	Natural gas available in neighborhood
POOL	Heated swimming pool
FUELTUB	Fuel used for heating hot tub
FUELHEAT	Main space heating fuel
AIRCOND	Air conditioning equipment used
COOLTYPE	Type of air conditioning equipment used
CENACHP	Central air conditioner is a heat pump
FUELH2O	Fuel used by main water heater
FUELH2O2	Fuel used by secondary water heater
ELWARM	Electricity used for space heating
ELWATER	Electricity used for water heating
ELFOOD	Electricity used for cooking
FOWATER	Fuel oil used for water heating
CLIMATE_REGION_PUB	Building America Climate Zone
Demographic	
NHSLDMEM	Number of household members
NUMADULT	Number of household members age 18 or older
NOACBROKE	Unable to use cooling equipment in the last year because equipment was broken and could not afford repair or replacement
PERIODNG	Number of days covered by Energy Supplier Survey natural gas billing data and used to calculate annual consumption and expenditures

When selecting features only from the Appliance category (Table 4-3), 19 out of 81 features are selected. The selected features include numbers and sizes of the refrigerators and freezers, and numbers of cooktops, ovens, coffee makers, and other small appliances. Numbers of televisions, cable or satellite boxes, computers, and smartphones are also selected. Numbers of ceiling fans and light bulbs are included as well. In summary, refrigerators and freezers, cooling appliances,

TVs, computers, smartphones, and light bulbs are the main factors to predict electricity consumption from the Appliance category.

Table 4-3. Selected Features from Appliances

Factor	Description
NUMFRIG	Number of refrigerators used
SIZRFRI1	Size of most-used refrigerator
ICE	Through-the-door ice on most-used refrigerator
NUMFREEZ	Number of separate freezers used
SIZFREEZ	Size of most-used freezer
STOVE	Number of separate cooktops
OVEN	Number of separate ovens
MICRO	Microwave oven used
COFFEE	Coffee maker used
APPOTHER	Other small appliance used
DRYRFUEL	Fuel used by clothes dryer
TVCOLOR	Number of televisions used
CABLESAT	Number of cable or satellite boxes without DVR
COMBODVR	Number of cable or satellite boxes with DVR
DESKTOP	Number of desktop computers
NUMSMPHONE	Number of smart phones
NUMCFAN	Number of ceiling fans used
LGTINNUM	Number of light bulbs installed inside the home
LGTOUTNUM	Number of light bulbs installed outside the home

9 out of 32 factors are selected from the Behavior category (Table 4-4). The months when swimming pools and hot tubs were used in the last year, oven usage, dishwasher usage, clothes dryer usage, TV usage on weekends or weekdays, water temperature for dishwasher rinse cycles, and the set-point temperature on summer nights are selected as important factors for predicting electricity consumption.

Table 4-4. Selected Features from Behavior

Factor	Description
MONPOOL	Months swimming pool used in the last year
MONTUB	Months hot tub used in the last year
SEPOVENUSE	Frequency of separate oven use
DWASHUSE	Frequency of dishwasher used
DRYRUSE	Frequency clothes dryer used
TVONWE1	Most-used TV usage on weekends
TVONWD2	Second most-used TV usage on weekdays
RNSETEMP	Water temperature used for rinse cycle
TEMPNITEAC	Summer temperature at night

The RECS has most detailed information about the Technology category, and 17 out of 117 features are selected (Table 4-5). They include climate zone, census urban type, housing type, numbers of rooms and bathrooms, presence of a heated or unheated swimming pool, types of air conditioning equipment, and fuel types for space heating, cooling, water heating, and cooking. It is noticeable that the features selected focus mainly on HVAC and fuel types for heating, cooling, and cooking compared to factors about building envelopes, which implies that HVAC-related factors predict electricity consumption more efficiently.

Table 4-5. Selected Features from Technology

Factor	Description
UATYP10	Census 2010 Urban Type
TYPEHUQ	Type of housing unit
NCOMBATH	Number of full bathrooms
TOTROOMS	Total number of rooms in the housing unit, excluding bathrooms
UGASHERE	Natural gas available in neighborhood
SWIMPOOL	Swimming pool
POOL	Heated swimming pool
FUELTUB	Fuel used for heating hot tub
COOLTYPE	Type of air conditioning equipment used
CENACHP	Central air conditioner is a heat pump
FUELH2O	Fuel used by main water heater
FUELH2O2	Fuel used by secondary water heater
ELWARM	Electricity used for space heating
ELFOOD	Electricity used for cooking
USENG	Natural gas used
FOWATER	Fuel oil used for water heating
CLIMATE_REGION_PUB	Building America Climate Zone

6 out of 41 factors are selected from the Demographic category (Table 4-6). They are: if the house is owned or rented, numbers of household members and household adults, and factors indicating the economic status of the household, including it they participated in home energy assistance programs, if they could afford to repair or replace broken cooling equipment, and the number of days covered by energy supplier survey billing data.

Table 4-6. Selected Features from Demographic

Factor	Description
KOWNRENT	Own or rent
NHSLDMEM	Number of household members
NUMADULT	Number of household members age 18 or older
ENERGYASST	Participated in home energy assistance program
NOACBROKE	Unable to use cooling equipment in the last year because equipment was broken and could not afford repair or replacement
PERIODNG	Number of days covered by Energy Supplier Survey natural gas billing data and used to calculate annual consumption and expenditures

Table 4-7. Selected Features from Application and Behavior

Factor	Description
Appliance	
NUMFRIG	Number of refrigerators used
SIZRFRI1	Size of most-used refrigerator
ICE	Through-the-door ice on most-used refrigerator
SIZRFRI2	Size of second most-used refrigerator
NUMFREEZ	Number of separate freezers used
SIZFREEZ	Size of most-used freezer
STOVE	Number of separate cooktops
OVEN	Number of separate ovens
MICRO	Microwave oven used
APPOTHER	Other small appliance used
DRYRFUEL	Fuel used by clothes dryer
TVCOLOR	Number of televisions used
COMBODVR	Number of cable or satellite boxes with DVR
DESKTOP	Number of desktop computers
NUMSMPHONE	Number of smart phones
NUMCFAN	Number of ceiling fans used
LGTINNUM	Number of light bulbs installed inside the home
LGTOUTNUM	Number of light bulbs installed outside the home
Behavior	
MONPOOL	Months of swimming pool used in the last year
MONTUB	Months of hot tub used in the last year
WASHLOAD	Frequency clothes washer used
TVONWE1	Most-used TV usage on weekends
TEMPNITEAC	Summer AC temperature at night

When selecting features from the Appliance and Behavior categories, 23 out of 113 features are selected (Table 4-7). Previously, 19 features had been selected from Appliance only, and 9 features had been selected from Behavior only. When combining these 2 categories, *Use of coffee maker* (COFFEE) and *Number of cable or satellite boxes without DVR* (CABLESAT) are excluded, and *Size of second most-used refrigerator* (SIZRFRI2) is added from the Appliance category. Also,

Frequency of separate oven use (SEPOVENUSE), *Frequency of dishwasher use* (DWASHUSE), *Frequency of clothes dryer use* (DRYRUSE), *Water temperature used for rinse cycle* (RNSETEMP), and *Summer temperature at night* (TEMPNITEAC) are excluded, and *Frequency of clothes washer use* (WASHLOAD) is added in Behavior category.

4.5.2. Energy Consumption Prediction

As feature selections are performed for each category, the electricity consumption predictions are also performed. The predictions' performances between all features and the selected features are compared using different algorithms.

As summarized in Table 4-8, SVM shows the best performance. The correlation coefficient is 0.8024 with all 271 features, and 0.7848 with the selected 32 features. It is notable that using the selected features, which are only 12% of the total number of features, still achieves 98% of the predictive performance reached with all of the features. This implies that the selected features are an efficient way to predict electricity consumption.

Table 4-8. Algorithm Performance with All Features

All Features	All (271)		Selected (32)	
Algorithm	Cor.Coeff.	RMSE	Cor.Coeff.	RMSE
SVM	0.8024	4222.66	0.7848	4404.92
Linear Regression	0.7853	4367.89	0.7826	4444.12
Random Forest	0.7825	4533.78	0.7809	4405.09
M5P Trees	0.7794	4418.81	0.7755	4451.09
M5 Rules	0.7794	4418.81	0.7637	4550.76

For the Appliance category, SVM shows the best performance, with a correlation coefficient of 0.6369 with all of 81 Appliance features and 0.5945 with the selected 19 features (Table 4-9). The selected features achieve 93% of baseline predictive performance with 23% of the number of Appliance category features.

Table 4-9. Algorithm Performance with Appliance Features

Application	All (81)		Selected (19)	
Algorithm	Cor.Coeff.	RMSE	Cor.Coeff.	RMSE
SVM	0.6369	5508.87	0.5945	5670.69
Linear Regression	0.6297	5477.69	0.5929	5677.64
Random Forest	0.6217	5549.93	0.5904	5692.82
M5P Trees	0.6185	5542.48	0.5889	5697.35
M5 Rules	0.6179	5545.44	0.5877	5823.52

For the Behavior category, SVM shows the best performance with a correlation coefficient of 0.5719 with all 32 Behavior features, and M5P Trees shows the best performance with a correlation coefficient of 0.5444 with the selected 7 features (Table 4-10). The selected features achieve 95% of baseline predictive performance with 28% of the number of features.

Table 4-10. Algorithm Performance with Behavior Features

Behavior	All (32)		Selected (9)	
Algorithm	Cor.Coeff.	RMSE	Cor.Coeff.	RMSE
SVM	0.5719	4173.85	0.5414	4300.46
Random Forest	0.5708	4318.83	0.4848	4648.84
M5P Tress	0.5685	4293.45	0.5444	4398.60
M5 Rules	0.5685	4293.45	0.5434	4401.48
Linear Regression	0.5685	4293.45	0.5432	4402.76

For the Technology category, SVM shows the best performance with a correlation coefficient of 0.7492 with all 117 Technology features, and 0.7240 with the selected 17 features (Table 4-11). The selected features achieve 97% of baseline predictive performance with 15% of the number of features.

Table 4-11. Algorithm Performance with Technology Features

Technology	All (117)		Selected (17)	
Algorithm	Cor.Coeff.	RMSE	Cor.Coeff.	RMSE
SVM	0.7492	4704.85	0.7240	4863.29
Random Forest	0.7399	4789.11	0.7170	4980.07
M5 Rules	0.7333	4793.41	0.7124	4954.03
Linear Regression	0.7332	4794.91	0.7045	5002.87
M5P Trees	0.7273	4840.76	0.6746	5233.95

For the Demographic category, M5 Rules shows the best performance with a correlation coefficient of 0.5815 with all 41 Demographic features, and 0.5316 with the selected 6 features (Table 4-12). Unlike the other categories, SVM is not the best algorithm for this category. The performance difference is not big when using all features (M5 Rules is 0.5815 and SVM is 0.5734, which is 98.6% of M5 Rules), but it shows greater differences with the selected features (M5 Rules is 0.5316 and SVM is 0.5085, which is 95.7% of M5 Rules). Using M5 Rules, the selected features achieve 92% of baseline predictive performance with 15% of the number of features.

Table 4-12. Algorithm Performance with Demographic Features

Demographic	All (41)		Selected (6)	
Algorithm	Cor.Coeff.	RMSE	Cor.Coeff.	RMSE
M5 Rules	0.5815	5735.67	0.5316	5971.21
Random Forest	0.5806	5745.83	0.5311	5973.50
M5P Trees	0.5787	5750.49	0.5288	5997.25
SVM	0.5734	5876.67	0.5085	6183.11
Linear Regression	0.5637	5823.14	0.4982	6112.20

For the Appliance and Behavior categories together, SVM shows the best performance with a correlation coefficient of 0.6831 with all 113 features, and 0.6429 with the selected 23 features (Table 4-13). The selected features achieve 94% of baseline predictive performance with 20% of the number of features.

Table 4-13. Algorithm Performance with Application and Behavior Features

Application+Behavior	All (113)		Selected (23)	
Algorithm	Cor.Coeff.	RMSE	Cor.Coeff.	RMSE
SVM	0.6831	5208.85	0.6429	5522.75
Linear Regression	0.6801	5169.57	0.6245	5505.81
M5P Trees	0.6759	5197.26	0.6373	5433.78
M5 Rules	0.6759	5197.26	0.6293	5481.00
Random Forest	0.6585	5368.32	0.6410	5413.05

SVM shows the best performance for most of the categories except for the Demographic category (all features and selected features) and the Behavior category (selected features). In the Demographic category, the correlated coefficient values of M5 Rules are the best with all features

(0.5815) and with the selected features (0.5316). Compared to M5 Rules, SVM reaches 99% of its performance with all features and 96% of its performance with the selected features in the Demographic category. In the Behavior category, the correlated coefficient value of M5P Trees is the best with the selected features (0.5444). Compared to M5P Trees, SVM reaches 99% of its performance with the selected features in Behavior category. Thus, the performance of the SVM algorithm in each category is compared in Table 4-14. The best performance is 0.8024 with all 271 features. However, it is meaningful that the selected 32 features from all categories still achieve 98% of the all-feature baseline performance. This performance is even better with 117 Technology features (correlation coefficient 0.7492) or 113 Appliance and Behavior features (correlation coefficient 0.6831). In addition, it is noticeable that the Demographic category shows weaker performance to predict energy consumption compared to other categories. The smaller number of the Demographic features can be one reason of the lower performance, but it raises new questions about using demographic characteristics for energy strategies in many existing studies and policies. The selected features generally achieve more than 90% of the performance achieved with all features (Table 4-14 and Figure 4-1). Given the number of features and the performance achieved, this demonstrates that the selected features are an efficient way to predict electricity consumption.

Table 4-14. Performance Comparison by Different Features

Features (SVM)	Cor.Coeff.		# Features		Ratio (Selected/All)	
	All	Selected	All	Selected	#Features	Cor.Coeff.
All Features	0.8024	0.7848	271	32	12%	98%
Technology	0.7492	0.7240	117	17	15%	97%
Appliance + Behavior	0.6831	0.6429	113	23	20%	94%
Appliance	0.6369	0.5945	81	19	23%	93%
Demographic	0.5734	0.5085	41	6	15%	92%
Behavior	0.5719	0.5414	32	9	28%	95%

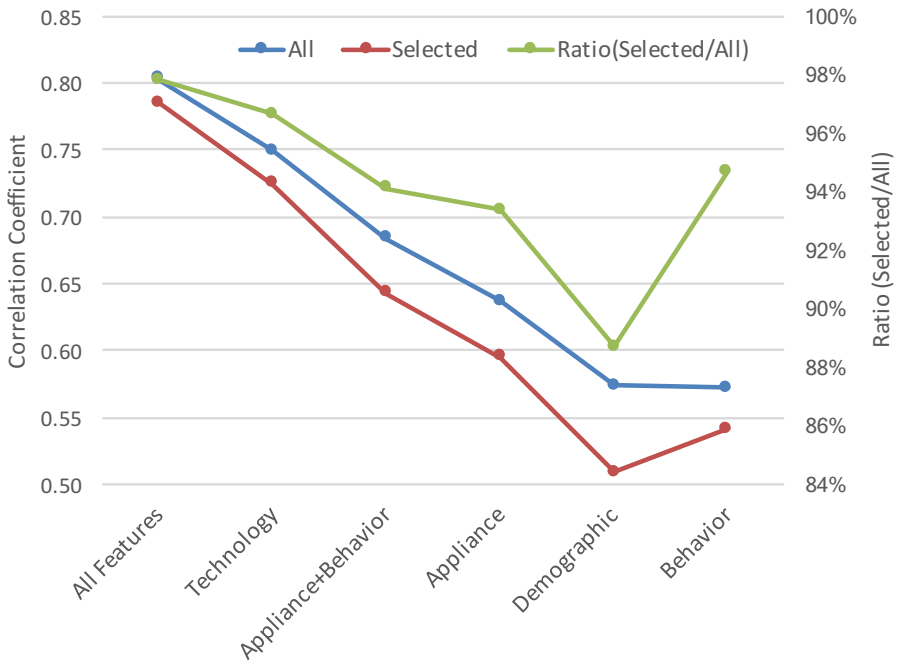


Figure 4-1. Performance Comparison by Different Features

4.6. Conclusion

The findings include the efficient features to predict energy consumption, and the prediction of total energy consumption in residential buildings. The main selected features are refrigerator, freezer, oven and television from the Appliance category, TV, cloth dryer, and swimming pool usage from the Behavior category, housing type, number of rooms from Technology category, and number of household members and number of young (under 18 years old) household members from the Demographic category. The selected 32 features predict the total electricity consumption with 78% accuracy, which almost reaches 80% accuracy with all 271 features. It shows that the selected features keep 98% of the prediction power compared to all of the 271 features.

This study provides lists of the most efficient factors for predicting electricity consumption in residential buildings. The lists can be used to predict energy consumption for more effective energy saving programs and to help residential occupants understand important factors regarding their

energy consumption patterns. Furthermore, the relationships between behavior-related factors (a set of appliances, the frequencies and times they are used, and the determining the usage by the occupants) and electricity usage provides the groundwork to predict occupant behavior from energy consumption data.

The limitation of the RECS dataset is that its data about behavior-related activities are less detailed compared to its more detailed data about appliances, building technology, demographic information, and energy usage. In the future, behavior-related factors can be further examined in detail by using datasets with more detailed behavior information.

CHAPTER 5

VALIDATION OF THE OCCUPANT BEHAVIOR PREDICTION MODEL USING REAL-WORLD HOME ENERGY SENSORS

Abstract

There have been a number of studies about occupant behavior and energy consumption in residential buildings. However, most of the studies tried to predict energy consumption from occupant behavior or building technology, and rarely predicted occupant behavior from energy consumption or technology. This study aims to identify the relationship between energy consumption, occupant behavior, and building technologies, and to predict occupant behavior with energy consumption data by applying the Occupant Behavior Prediction Model. The actual household's energy consumption data is used to predict appliances and associated occupant activities using machine learning (ML) numeric prediction algorithms. In addition, the American Time Use Survey (ATUS) and the Residential Energy Consumption Survey (RECS) national survey data are further examined using clustering and descriptive analysis to support the sensor data analysis. The results show that the Occupant Behavior Prediction Model with the ML Decision Tree algorithm can achieve 96 percent accuracy when predicting appliances and associated activities. These findings can be further used to set efficient energy saving strategies in residential buildings.

5.1. Introduction

A number of studies have been conducted regarding occupant behavior and energy consumption, but each study approaches this topic with different perspectives. Some studies focus more on finding the structural relationship between occupant behavior and energy usage, some studies are interested in building a model to predict energy usage, and some try to save energy by promoting certain occupant behavior.

The fundamental objective of energy saving through occupant behavior is to understand their relationship in a systematic and structural way. Yu et al. (2011) developed a methodology to examine the influences of occupant behavior on building energy consumption using the data mining technique called cluster analysis. Diao et al. (2017) presented an unsupervised clustering

method to identify and classify the relationship between occupant behavior and energy use over time. Santin (2011) studied the association of behavioral patterns and heating energy consumption in order to identify the building characteristics that contribute to energy use. Chen et al. (2015) addressed three levels of residential occupant behavior by assessing different complexities of the parameters in each level: the simple level for simple descriptive analysis, the intermediate level for statistical analysis, and the complex level for energy simulation.

Other studies developed frameworks or models to understand the relationship between energy and behavior in a more quantitative way. Aksanli et al. (2016) developed a residential energy modeling frameworks based on human activities, and implemented power demand profiles based on the characteristics of occupants. They developed a user-behavior model to predict the energy consumption of a residential building, based on detailed activity sequences of occupants and the relationship between the activities and appliances used. Chen et al. (2012) developed an agent-based computational model for individual occupant energy consumption behavior. They collected energy consumption data during an experiment on a residential building. They leveraged how the energy consumption behavior in the dataset can be modified with the network relations, and explored how energy consumption patterns can be related to the structural properties of peer networks. Sanquist et al. (2012) developed a model of lifestyle factors by quantitative, multivariate methods with regard to U.S. residential electricity consumption. Santin et al. (2009) developed a statistical model of residential occupant behavior to estimate occupants' energy consumption. They explored the effect of occupant behavior on space heating energy consumption while controlling for building characteristics.

Some other studies examined ways to improve occupant behavior. One example is a study conducted by Diao et al. (2017), which aimed to evaluate the energy saving potential of promoting the improvement of occupant behavior using the energy consumption patterns identified by a cluster analysis.

However, few studies established clear relationships between energy consumption and occupant behavior in residential buildings. Kavousian et al. (2013) pointed out the limitations of existing bottom-up research, including use of low-resolution energy consumption data, limited sets of explanatory variables, no clear distinctions between peak energy consumption and idle energy consumption, and use of energy intensity as the only evaluating indicator for energy usage. Thus, a more comprehensive study is still needed.

The goal of this study is to identify the relationship between energy consumption, occupant behavior, and technologies, and to predict occupant behavior with the actual household's energy consumption data by applying the Occupant Behavior Prediction Model in order to verify the model. In addition, national survey data from the American Time Use Survey (ATUS) and the Residential Energy Consumption Survey (RECS) are further analyzed to support the model and the sensor data analysis.

5.2. Background

5.2.1. Data Used for Existing Studies

Existing studies have used data from various sources to understand occupant behavior regarding energy consumption in residential buildings. Main sources include measured data, surveys, the

RECS, the ATUS, and the ASHERAE occupant schedule, and most studies used data from multiple sources. Some examples are discussed below.

5.2.1.1. Measured Data (Energy, Occupant Behavior)

Most measured data are energy consumption data. Very few studies include measured occupant behavior data, since it is difficult to observe and measure occupant behavior for multiple households for a long duration at a detailed level. Most of the behavior data were self-recorded manually, and there are privacy concerns in the case of long-term observation.

The most common source of energy consumption data is monthly utility usage and bill information from energy distributors or energy companies. Kavousian et al. (2013) used 10-minute interval electricity consumption data of 1628 households for 238 days to examine behavioral determinants of electricity consumption in residential buildings. Santin et al. (2009) used energy consumption data from around 15,000 households for three years from an energy provider in the Netherlands. Ouyang and Hokao (2009) used monthly electricity usage data from 124 households in Hangzhou, China for 17 months. Vassileva et al. (2012a) used monthly electricity usage data from 24 multi-residential households with 40 residents in Sweden.

However, monthly energy consumption data tend to only have aggregated energy consumption, and do not include appliance-level usage or energy consumption in more granular time intervals. Thus, some studies use more detailed energy consumption data measured by various sensors or experiments. Chen et al. (2012) used electricity consumption data collected from an experiment with 45 occupants for 46 days to develop an agent-based model to assess individual energy

consumption behavior. Yu et al. (2011) used measured energy data for 80 houses to identify the effects of occupant behavior on energy consumption of residential buildings. Higashino et al. (2014) used sensor-measured electricity consumption data in each circuit level for 586 apartments in Osaka, Japan.

Most existing studies used measured energy data, and they difficultly used measured occupant behavior data. One example of a study that used measured occupant behavior data is a study conducted by Chen et al. (2015), which used one year of real-time monitoring data of occupant behavior in a family to verify their list structure of occupant behavior at an intermediate level.

5.2.1.2. Survey Data

Surveys are a common way to collect occupant behavior data and their socioeconomic information. A number of studies create their own questionnaires about occupant behavior. Yu et al. (2011) used survey data about occupants' lifestyles, annual incomes, utilization of appliances, and basic building information for 80 households. Kavousian et al. (2013) used a survey of household data with 114 questions, including the location, climate, building attributes, appliances, and occupants.

Bartusch et al. (2012) used survey data from 595 households about household-specific features, building properties, main heating systems, supplementary heating devices, energy saving installations, and energy-consuming installations. Chen et al. (Chen et al., 2015) used survey data from 73 families in a city to verify their list structure of occupant behavior at a simple level. Ouyang and Hokao (2009) used survey results from 124 households in three typical residential buildings in a Chinese city. The questions included occupant energy usage-related behaviors and

building characteristics. Vassileva et al. (2012b) used behavioral survey data from 24 households and matching monthly energy data in Swedish multi-residential buildings.

National survey data are also a good source of occupant behaviors and building characteristics. One example is a study by Santin et al. (2009), which used national survey data from 15,000 households in the Netherlands, with questions about household characteristics and building attributes. In the United States, the RECS and the ATUS data are frequently used in research about occupant behavior and energy consumption. These will be further examined in the following subsections.

5.2.1.3. RECS

The Residential Energy Consumption Survey (RECS) is a national energy survey for residential buildings conducted by the Energy Information Administration (EIA) under the U.S. Department of Energy (US DOE). They have been conducted every three years since 1978 (Sanquist et al., 2012). The RECS mainly collects the total energy consumption data of a household over one year with energy fuel types, building geometry information, household demographics, and appliance information (Diao et al., 2017). Lifestyle patterns can be derived from a subset of the RECS variables. These variables include geographic location, household equipment and appliances, family structure, income, and local electricity price (Sanquist et al., 2012).

Sanquist et al. (2012) used the 2005 RECS data, collected from 4382 housing units representing 111.1 million housing units in the U.S. of that year. They focused on 2165 single houses with annual electricity bill data collected from utility companies. Diao et al. (2017) used the 2009 RECS

data in New York State, which included 938 households sub-grouped by five typical house types: (1) single family (detached), (2) single family (attached), (3) apartment (2-4 units), (4) apartment (5+ units), and (5) mobile home. Aksanli et al. (2016) used appliance information from the RECS including the types, numbers, and use frequencies of the appliances.

5.2.1.4. ATUS

The American Time Use Survey (ATUS) is a survey conducted by the U.S. Bureau of Labor Statistics every year. The purpose of the survey is to record respondents' activities, locations, and demographic information on a regular day from 4 AM to 4 AM of the next day (Diao et al., 2017). The ATUS provides (1) population measurement and (2) participant measurement. The population measurement provides the average time of day that participants do an activity for a particular population. The participant measurement estimates the average time spent on an activity per day (Diao et al., 2017). While the time use surveys conducted in other countries, such as the United Kingdom and Sweden, require respondents to record their activity with 5- or 10-minute intervals, the ATUS asks respondents to report the start and end times of an activity.

Diao et al. (2017) used the 2009 ATUS data collected from New York State, including 738 instances of activities. They summarized the activities in the ATUS as follows: (1) personal care, (2) household activities, (3) caring for and helping household members, (4) caring for and helping non-household members, (5) work and work-related activities, (6) education, (7) consumer purchases, (8) professional and personal care services, (9) household services, (10) government services and civic obligations, (11) eating and drinking, (12) socializing, (13) sports, (14) religious and spiritual activities, (15) volunteer activities, (16) telephone calls, and (17) traveling. Then,

they used lower individual-level activity data from the ATUS and connected them with higher-level family and appliance statistics from the RECS.

Aksanli et al. (2016) calculated activity graph parameters with detailed activity information from the ATUS. They then combined it with appliance information from the RECS. Johnson et al. (2014) used the ATUS data to develop behavioral models that show the interaction between an individual occupant and the major residential energy consumption loads in a day. However, they argued that the activity categories in the ATUS are too broad and not directly associated enough with the participants' energy consumption.

5.2.2. Methods Used for Existing Studies

Diverse methods have been employed in past studies to explain occupant behavior and energy usage. The main methods used include (1) machine learning / data mining, (2) statistics, and (3) simulation.

5.2.2.1. Machine Learning / Data Mining

A number of recent studies used data mining techniques to understand the effects of user behavior. Clustering is one of the most popular pattern recognition methods, and groups data into unsupervised clusters, keeping more similar data within a single group.

Yu et al. (2011) used data mining techniques and clustering analysis to categorize the data. Min-max normalization is utilized to deal with data inconsistencies in the data pre-processing step. Grey relational grades were used as weighted coefficients of different attributes to measure the

relatedness between two factors. Diao et al. (2017) used a method that integrated K-modes clustering and probability neural networks to identify ten distinctive behavior patterns within the ATUS data demographic information. In this study, K-modes clustering with Hemming distance was selected to process categorical data of the behavior schedule. The center of a K-modes cluster is set with the most frequently appearing value for each attribute. The ideal number of k was decided with Akaike Information Criterion (AIC). Diao et al. (2017) applied a hierarchical clustering method to recognize occupancy patterns. Yu et al. (2011) used a decision tree method for modeling building energy demand and applied the model to historical data from residential buildings in Japan.

5.2.2.2. Statistics

Statistical analysis is one of the most popular methods used to analyze data about occupant behavior and energy consumption. Various statistical methods have been used in existing studies.

Sanquist et al. (2012) used a multivariate statistical method to analyze the lifestyle regarding residential electricity consumption. Factor analysis was utilized for the selected variables from the 2005 RECS data in order to identify five lifestyle factors explaining social and behavioral patterns regarding air conditioning, laundry, computer usage, TV usage, and climate zone. Kavousian et al. (2013) used factor analysis (FA) to eliminate multicollinearity of the variables, and to identify latent variables that are not revealed by direct behavioral questions. FA decreases the number of variables while keeping as much information as possible.

5.2.2.3. Simulation / Modeling

Modeling or simulation is another method frequently used to understand the relationship between occupant behavior and energy consumption in a more detailed and quantitative way.

Diao et al. (2017) used a first-order inhomogeneous Markov chain to synthesize the individual activity schedule of an RECS respondent based on the activity schedules of all occupants who performed the same behavior pattern in the ATUS data. Johnson et al. (2014) implemented Markov chain–based statistical models with time-varying minute resolution to analyze different types of occupant behaviors. They simulated individual occupants using these behavioral models to show the interactions between an occupant's behavior and the major energy-consuming loads throughout a day within the residential sector. Chen et al. (2012) adopted agent-based modeling to simulate the decision-making process of building occupants and the information transmission process. Agent-based modeling enables the analysis and manipulation of agents interacting within a given environmental condition. It can integrate situations that are not in the status of equilibrium and directly manage the results of interactions between agents. Thus, it is well suited to these studies, to emphasize the process and its consequences. Aksanli et al. (2016) developed a graph-based model to explain the chain of occupant activities. They probabilistically captured user behavior with its time-series nature. The probabilities are derived from people-related variables (e.g. number of other household members, gender, age, employment, etc.) and non–people-related variables (e.g. time of day, day of the week, etc.).

5.3. Data

5.3.1. Sensor Measured Data

The actual energy consumption data from two testbeds were measured using home energy sensors from Smappee, which is an international smart energy monitoring system. Appliance usage and electricity consumption data are collected from June 2017 till November 2018, and one-year data between 7/1/2017 and 6/30/2018 are used in this study. Figure 5-1 explains the data collection process. The sensor is installed in an electricity box to measure the electricity consumption of home appliances. It is connected to a Wi-Fi router to communicate with the sensor cloud. The cloud stores the measured data and current energy consumption can be monitored by other devices (like smartphones, tablets, and computers). The data in the cloud can be downloaded as a .csv file.

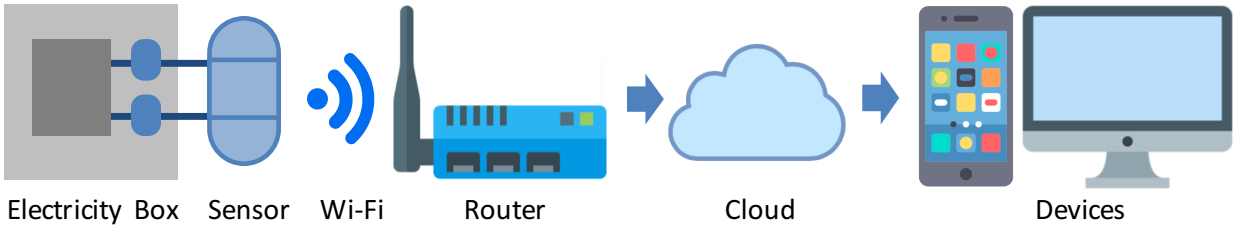


Figure 5-1. Data Collection Process

The characteristics of the testbeds are as follows:

- ***Case1:*** Main data
  - Location: Okemos, Michigan
  - House Type: Single family detached house
  - Number of family members: 4 (2 adults and 2 children)
  - Job: At least one occupant has at least one main job
  - Data collection: Started from 6/5/2017
- ***Case2:*** Supplementary data with similar conditions to Case1, collected starting 1/20/2018.

The sensors measured detailed electricity usage data including the total electricity usage and the appliance usage pattern of the household.

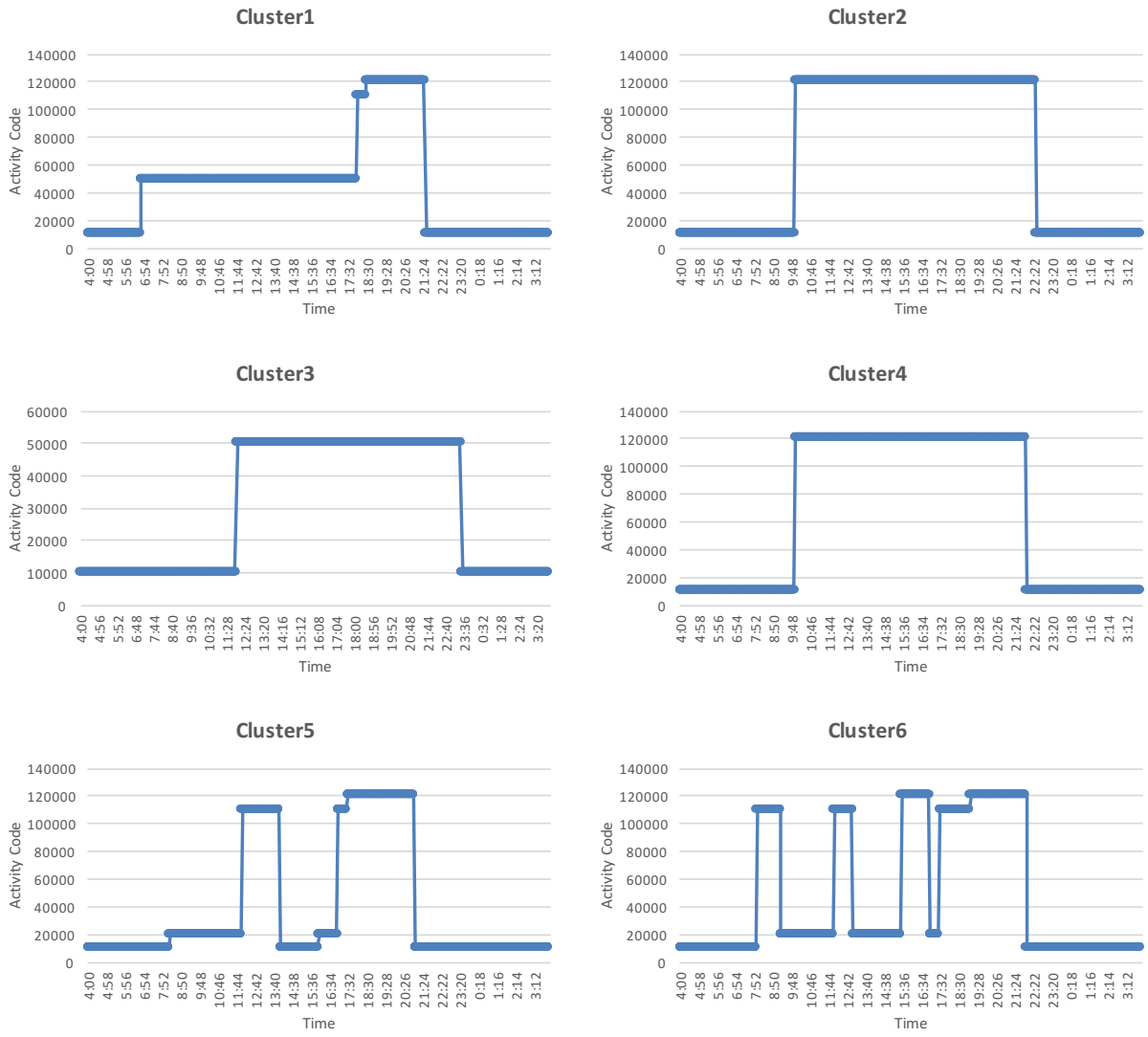
- ***Event data:*** Appliance on/off data with timestamps and wattage of the appliance
- ***Electricity data:*** 5-minute intervals of total electricity usage in kWh

The monitoring system used non-intrusive load monitoring (NILM), which is a load disaggregation technology used to identify appliances. Each appliance has a specific watt range and uses the electrical current in a unique way. The system learns the unique patterns and identifies each of the appliances (Smappee, 2018).

5.3.2. Other Data

Other datasets are used to support the analysis of the sensor-measured data.

- ***The American Time Use Survey (ATUS)***
  - Daily activities by 1-minute intervals of selected instances are extracted from the 2015 ATUS data.
  - Daily activity routines of 6 occupant clusters (Figure 5-2) are used for further activity data analysis (Mo, 2018)(Chapter 3).



*** Y-axis values are the 3rd tier ATUS activity codes which are nominal values in the form of numbers. 10000 is not "more" than 8000, and same Activity Code values indicate same activities and different values indicate different activities.*

Figure 5-2. Daily Activity Routines of Occupant Clusters (=Figure 3-2)

- ***The Residential Energy Consumption Survey (RECS)***
  - Overall energy consumption and appliance information of selected instances are extracted from the 2015 RECS data.
  - Appliance features (Table 5-1) from the feature selection process are used for further electricity and appliance data analysis (Mo, 2018)(Chapter 4).

Table 5-1. Selected Features from Appliances (=Table 4-3)

Factor	Description
NUMFRIG	Number of refrigerators used
SIZRFRI1	Size of most-used refrigerator
ICE	Through-the-door ice on most-used refrigerator
NUMFREEZ	Number of separate freezers used
SIZFREEZ	Size of most-used freezer
STOVE	Number of separate cooktops
OVEN	Number of separate ovens
MICRO	Microwave oven used
COFFEE	Coffee maker used
APPOTHER	Other small appliance used
DRYRFUEL	Fuel used by clothes dryer
TVCOLOR	Number of televisions used
CABLESAT	Number of cable or satellite boxes without DVR
COMBODVR	Number of cable or satellite boxes with DVR
DESKTOP	Number of desktop computers
NUMSMPHONE	Number of smart phones
NUMCFAN	Number of ceiling fans used
LGTINNUM	Number of light bulbs installed inside the home
LGTOUTNUM	Number of light bulbs installed outside the home

- ***Weather Data***
  - Cooling degree days (CDD) and heating degree days (HDD) data, mapped with the location of the sensor-measured data (Okemos, MI), are downloaded from degreedays.net.
  - CDD and HDD are used for appliance prediction and further electricity usage analysis.

5.3.3. Data Pre-Process

In this study, the sensor-measured data from 1 year (from 7/1/2017 to 6/30/2018) at Case1 is used for the main data analysis. It includes two different datasets: (1) event data with appliance information, and (2) electricity data with the total household electricity consumption. The raw data includes noise, such as outliers or repeated values due to the error of the sensor, and the noisy data is removed during data pre-process.

Sensor Measured Event Data

The total number of instances of event data is 599,019, and they are used to predict appliances. The raw event data have the type and name of each appliance, its power in watts, its action (on/off), and the timestamps of each action. The variables are defined based on the Occupant Behavior Prediction Model (Figure 5-3) (Mo, 2018)(Chapter 2).

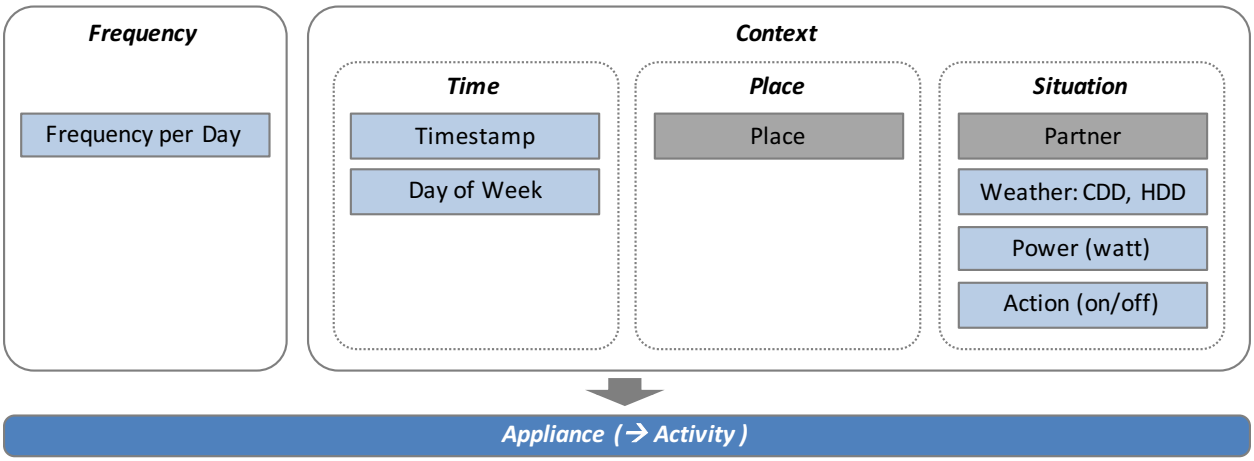


Figure 5-3. Appliance/Activity Prediction with Occupant Behavior Prediction Model

The sensor-measured event data are used to predict appliance names using the variables based on the Occupant Behavior Prediction Model. Since the data are measured from the same household, Place is excluded. Partner is also excluded due to limitations of the data collection method, which could not record or distinguish which household member(s) used the appliances.

- ***Independent variables (Xs)***
  - **Frequency:** Calculated by counting the number of "on" actions of each appliance during each day.
  - **Timestamp:** Minute-level timestamps of each appliance action in number format. For example, 4:00 AM is converted to 240 minutes.

- **Day of the Week:** Weekday or weekend. Weekday is coded as 0, and weekend as 1. Occupant activities differ significantly between weekdays and weekends (Mo, 2018)(Chapter 3), and thus this variable is added to the prediction features.
- **Cooling Degree Days (CDD) and Heating Degree Days (HDD):** Since weather affects activities, heating/cooling, and appliance usage, the CDD and HDD of each day are added to the prediction features.
- **Power:** Wattage of the appliance. Power and Action are important contexts of the appliance usage, and they are added to the prediction features.
- **Action:** The action of turning the appliance on or off. On is coded as 1, and off is coded as 0.

- ***Dependent variable (Ys)***
  - **Appliance:** The name of the appliance.
  - **Activity/Behavior:** This is derived from the table showing the association between appliances and specific activities/behaviors (Further explained in subsection 5.4.3).

Sensor-Measured Electricity Data

The raw electricity data have 5-minute interval timestamps and show the total electricity consumption of the household in kWh. The dataset is converted to a matrix format: rows are days (total 365 days) and columns are the 5-minute intervals of a day (288 interval timestamps in minutes). Missing data are assigned the average value of each column. Electricity data are used to identify energy consumption patterns through clustering.

5.4. Methodology

Earlier, the Occupant Behavior Prediction Model was defined and the model was applied to ATUS data to predict energy usage–related activities and to identify habitual activities (Mo, 2018)(Chapter 2). Occupants' habitual activities were further examined in the following study (Mo, 2018)(Chapter 3), and then efficient appliances and other factors affecting residential electricity usage were identified (Mo, 2018)(Chapter 4). As described in Figure 5-4, in this chapter, the Occupant Behavior Prediction Model is applied to sensor-measured residential energy and appliance usage data to predict occupants' activities. Previous results (Mo, 2018)(Chapter 3 and 4) are given further descriptive analysis to support the sensor-measured data analysis. Future works for refined behavior prediction are also suggested in the latter part of this chapter.

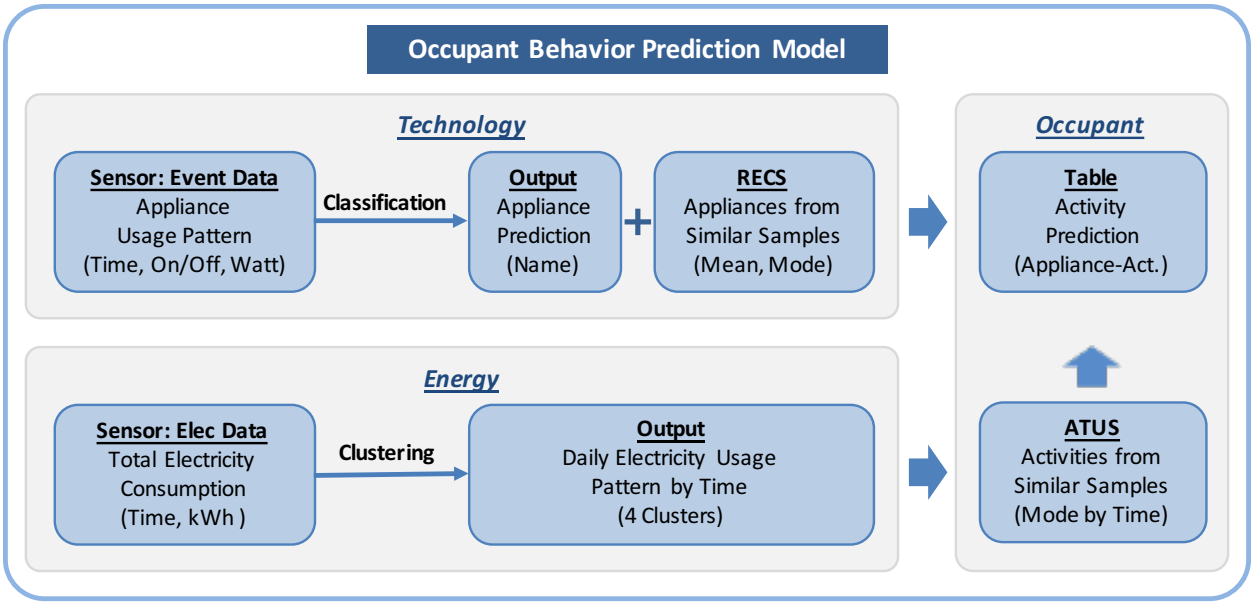


Figure 5-4. Overall Research Flow

5.4.1. Classification: Predicting Appliances

The sensor-measured event data are used to predict which appliances are in use. The features are mixed with numeric and categorical variables as follows.

- ***Numeric Variables:*** Frequency, Timestamp, CDD, HDD, Power
- ***Categorical Variables:*** Day of the week, Appliance

The numeric variables are standardized to a mean of 0 and a standard deviation of 1, and the models' performances between standardized and non-standardized features are compared. Naïve Bayes (NB), Logistic Regression (LR), K-Nearest Neighbor (KNN), Decision Tree (DT), and Support Vector Machine (SVM) are all used, and their performances are evaluated with Accuracy, Kappa, Precision, Recall, and F1-score.

5.4.2. Clustering: Grouping Electricity Usage Pattern

The sensor-measured electricity usage data are used to identify usage patterns by time and the main factors affecting a pattern. K-means clustering identifies the groups of daily electricity usage patterns by time. The number of clusters need to be decided prior to running K-means clustering, and the elbow method with distortion is used in this study.

The elbow method is a common technique to determine the appropriate number of clusters (Madhulatha, 2012). It evaluates the cost function value by increasing K by 1 in each step, starting at 2. At a certain value of K, the cost noticeably drops and the slope of the drop rate becomes smaller after that point. The K at the location of such an elbow is selected as the number of clusters for the dataset (Tibshirani, Walther, & Hastie, 2001). Distortion is determined by the within-cluster sum-of-squares, which is the sum of the squared distance between each cluster element and its cluster centroid. It can be a measure of the internal coherence of clusters, and lower values indicate that the clusters are more coherent (Kolesnikov & Trichina, 2012).

After defining the clusters, the centroid values of the clusters are plotted and analyzed, and the characteristics of the clusters (Day of the week, Month, and CDD/HDD) are examined.

5.4.3. Descriptive Analysis: Connecting Energy – Technology – Behavior

The 2015 ATUS and the 2015 RECS data are used for further descriptive analysis to support the sensor-measured data analysis by demonstrating the relationship between energy usage, technology (appliances), and occupant energy usage-related behavior (activities). In the ATUS and the RECS, responses with similar conditions to the sensor data testbed are selected and their characteristics are analyzed. The categories of the conditions in the ATUS and the RECS are not identical, so the most similar conditions are selected.

ATUS: Activity

The activity data are selected from the respondents who meet the criteria below. Out of 10,772 total data points, 63 instances are selected for this study. 26 instances are activities on weekdays, and 37 are activities on weekends.

- **Location:** Michigan state
- **Job:** Have at least one job
- **Respondent's Age:** Between 25 and 50
- **Number of household members:** 3 to 4
- **House type:** House or apartment/flat (recorded as one code)

Each respondent's 1-minute interval activities are summarized in a table (rows of respondents, and columns of minutes of a day). The mode values of activity codes are calculated for individual

minutes and the modes are summarized and plotted. Then, the results are compared with the daily activity routines of occupant clusters (Mo, 2018)(Chapter 3).

RECS: Energy and Appliance

The energy and appliance data are selected from the households that meet the criteria below. Out of 5685 total data points, 133 are selected for use.

- **Location:** East North Central division (The 2015 RECS does not have state-level location information, and census division is the most detailed level of location.)
- **Job:** Employed full-time
- **Respondent's Age:** Data not collected in the RECS
- **Number of household members:** 3-4
- **House type:** Single-family detached house

The average electricity usage is calculated from the selected data subset, and the total energy, cooling-specific energy, and heating-specific energy are compared. Then, the selected appliance features (Mo, 2018)(Chapter 4) are further examined: modes are calculated for categorical features, and averages are calculated for numeric features.

Connecting Activity and Appliance

Activities and the appliances for each activity are closely associated. Thus, when appliances are identified, associated activities can be predicted, and vice versa. Table 5-2 summarizes respondents' activities and their associated energy types and appliances (Mo, 2018)(Chapter 2).

Table 5-2. Activities and Associated Energy and Appliances (=Table 2-3)

Code	Activity	Energy	Appliances (Electricity and Gas)
AA01	Washing, dressing, and grooming	E,W,G	Lighting, Shower, Hair dryer, Shaving
BB01	Interior cleaning	E	Lighting, Vacuum
BB02	Laundry	E,W,G	Lighting, Washer, Dryer
BB03	Food and drink preparation	E,W,G	Lighting, Oven, Stove, Toaster, Blender, Coffee machine, Cooker, etc.
BB04	Kitchen and food clean-up	E,W	Lighting, Dish washer
BB05	Heating and cooling	E,G	Lighting, HVAC
BB06	Gardening, ponds, pools, and hot tubs	W,G,E	Lighting
BB07	Care for animals and pets	E,W	Lighting
BB08	Vehicle repair and maintenance	E	Lighting, Repair tools
CD01	Physical care for children	E,W	Lighting
CD02	Physical care for/helping adults	E,W	Lighting
EF01	Work for job(s)/research/homework	E	Lighting, Computer
LL01	Television	E	Lighting, TV
LL02	Listening to/playing radio or music	E	Lighting, Computer, Music player, Radio
LL03	General computer use	E	Lighting, Computer

*** E: Electricity, W: Water, G: Gas*

5.5. Result

5.5.1. Classification

The sensor detected 105 different appliances, including heating/cooling elements, lighting, other appliances for work, entertainment, cooking, cleaning, and more. Table 5-3 summarizes the list of detected appliances with the number of on/off actions ("Count"), and the mean, minimum, and maximum wattage of each appliance measured from the sensor. The appliances are sorted by their mean watt values in descending order, excluding 14 appliances with on/off counts of under 100. Heating elements generally have high watt values compared to other appliances.

Table 5-3. Appliance List from Sensor Data

Appliance	Count	Mean	Min	Max	Appliance	Count	Mean	Min	Max
Heating element 47	531	2716	2221	3339	Appliance 59	480	315	244	399
Heating element 19	4534	2665	2262	3116	Appliance 55	467	291	232	348
Heating element 17	5360	2657	2219	3147	Lights 33	2815	279	187	398
Appliance 91	3724	2653	2073	3217	Lights 99	389	269	183	352
Appliance 82	3662	2639	2049	3058	Appliance 27	2733	257	48	515
Heating element 45	747	1845	1569	2151	Appliance 97	271	242	209	285
Heating element 53	427	1829	1527	2113	Motor 23	2204	240	167	329
Heating element 16	2035	1730	1457	2088	Appliance 20	1743	206	171	245
Appliance 81	1345	1720	1503	2077	Appliance 31	8165	172	46	494
Heating element 63	356	1620	1288	1710	Appliance 57	442	170	148	184
Motor 50	1129	1491	752	2828	Appliance 30	9693	164	104	290
Heating element 22	2292	1398	1130	1624	Appliance 12	7769	162	114	229
Motor 48	381	1392	1272	1892	Appliance 88	2516	160	123	195
Heating element 15	6215	1267	1056	1566	Lights 44	2242	155	126	192
Heating element 18	5487	1261	1124	1406	Appliance 78	955	150	111	217
Appliance 86	7200	1251	1051	1490	Appliance 29	606	135	112	161
Appliance 77	5397	1245	1067	1497	Lights 95	721	129	103	156
Motor 40	510	1219	1038	1595	Motor 9	10332	125	87	273
Heating element 67	953	1195	971	1235	Motor 49	1325	124	101	166
Heating element 56	1316	964	809	1146	Lights 39	2123	122	92	146
Appliance 93	529	960	823	1214	Refrigerator	2184	116	90	163
Heating element 25	173	945	803	1132	Appliance 26	1788	115	92	139
Vacuum Cleaner	122	908	790	1113	Lights 98	663	114	96	134
Heating element 64	3548	866	763	952	Appliance 38	1070	98	86	111
Heating element 65	579	839	707	897	Appliance 52	2238	81	61	96
Appliance 94	242	816	702	894	Appliance 8	14891	80	62	97
Appliance 69	137	763	713	978	Appliance 79	6725	79	66	101
Appliance 34	703	741	618	882	Appliance 43	7034	73	61	102
Appliance 10	10617	739	535	954	Appliance 42	7428	72	51	101
Appliance 87	5242	738	535	913	Appliance 96	309	71	59	84
Appliance 46	364	710	640	849	Appliance 14	173598	65	39	103
startMicroWave	894	668	601	892	Appliance 75	78975	65	39	79
Appliance 21	7192	643	486	810	Appliance 13	4854	57	47	69
Appliance 24	4477	638	581	720	Appliance 89	1542	57	46	68
Appliance 7	3048	624	489	767	Appliance 51	2874	48	38	59
Lights 76	8923	613	453	761	Appliance 90	1301	47	37	54
Lights 7	27351	613	407	817	Appliance 68	363	46	41	53
Appliance 11	2137	603	523	660	Appliance 28	2530	43	33	58
Lights 85	17376	592	413	686	Appliance 92	789	40	30	58
Lights 11	56973	591	462	762	Appliance 71	995	38	30	44
Appliance 61	884	580	457	722	Appliance 37	405	36	25	47
Microwave	7018	478	172	804	Appliance 35	209	36	27	45
Appliance 80	1404	436	332	591	Appliance 41	1931	32	16	74
Appliance 32	1724	421	339	592	Appliance 62	2445	28	21	39
Appliance 100	626	416	343	513	Appliance 54	7342	24	19	31
Appliance 84	164	415	349	480					

*** Mean, Min (minimum), Max (maximum) in watt*

The algorithms DT, KNN, NB, SVM, and LR are used to predict the appliance names (total 105 appliances), and Table 5-4 summarizes the performance of the algorithms.

Table 5-4. Performance of Appliance Prediction

Algorithm		DT	KNN	NB	SVM	LR
Non-Scaled	Accuracy	0.96	0.90	0.68	0.63	0.52
	Kappa	0.96	0.89	0.64	0.53	0.43
	Precision	0.96	0.90	0.70	0.80	0.44
	Recall	0.96	0.90	0.68	0.63	0.52
	F1-Score	0.96	0.90	0.67	0.57	0.45
Scaled	Accuracy	0.96	0.86	0.68	0.72	0.52
	Kappa	0.96	0.84	0.63	0.68	0.43
	Precision	0.96	0.86	0.70	0.70	0.43
	Recall	0.96	0.86	0.68	0.72	0.52
	F1-Score	0.96	0.86	0.67	0.69	0.45

DT shows the highest accuracy (0.96) among the algorithms, followed by KNN (0.90) and NB. SVM and LR showed lower performance (other performance criteria in Table 5-4). There is no difference between non-scaled and scaled data for DT, NB and LR. KNN performs better with non-scaled data and SVM performs better with scaled data, but their performance scores are still lower than those of DT, KNN, and NB.

5.5.2. Clustering

Distortion values are examined by increasing the number of clusters incrementally from 2 to 10 to determine the appropriate number of clusters. Distortion drops slowly after 4 clusters (Figure 5-5), thus 4 is selected for the number of clusters (k) for K-means clustering.

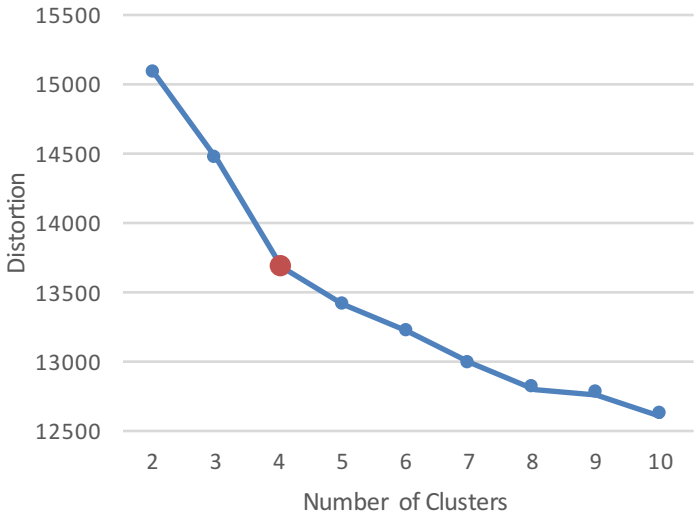


Figure 5-5. Elbow Method with Distortion

Figure 5-6 illustrates the centroid values of the clusters, and each line indicates the electricity usage of each cluster over time. Cluster 1 consumes the most electricity among all the clusters and shows big differences between daytime and nighttime. Cluster 2 also shows peak consumption in the late afternoon, but it consumes less energy than Cluster 1. Cluster 3 and Cluster 4 show relatively constant electricity consumption, and have slight peaks around 6 PM. The result is meaningful in that the energy consumption of each cluster is specified in minute-level interval. This can be used for detailed energy strategies for the households having similar conditions with this testbed.

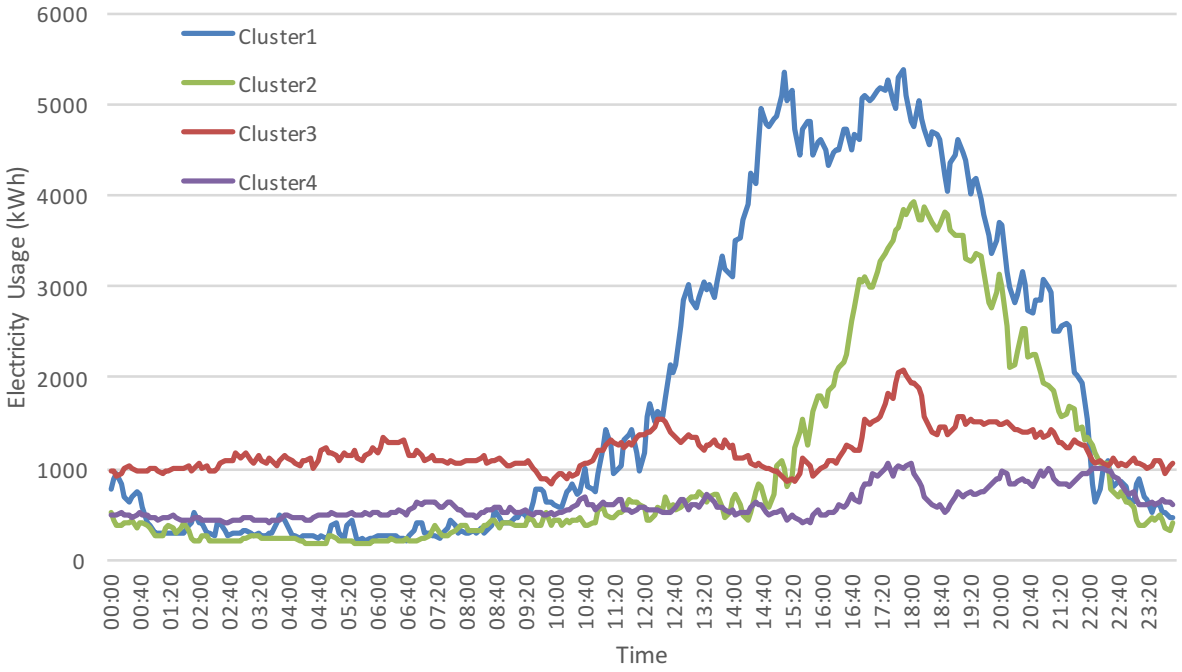


Figure 5-6. Daily Electricity Usage of Clusters

Table 5-5 summarizes the descriptive analysis of the clusters. Cluster 1 and Cluster 2 represent summer days with low HDD (1.7 and 2.2) and high electricity usage for cooling during daytime. Cluster 1 has hotter days than Cluster 2. The average CDD of Cluster 1 is 11.5, which means that Cluster 1 mainly consists of the very hot summer days from May to September. The average CDD of Cluster 2 is 6.5, which means that Cluster 2 mainly consists of the warm summer days from May to September. Cluster 3 and Cluster 4 represent non-cooling seasons with low CDD (0 and 1.9).

Table 5-5. Descriptive Analysis of Clusters

Cluster	Avg. Degree Day		Numbers by Month												Total#
	CDD	HDD	1	2	3	4	5	6	7	8	9	10	11	12	
1	**11.5**	1.7	0	0	0	0	**7**	**6**	**5**	**3**	**6**	**1**	0	0	28
2	**6.5**	2.2	0	0	0	0	**7**	**10**	**21**	**16**	**9**	**2**	0	0	65
3	0.0	**34.2**	**23**	**25**	**20**	**11**	0	0	0	0	0	**11**	**28**	**31**	149
4	1.9	**13.8**	**8**	**3**	**11**	**19**	**17**	**14**	**5**	**12**	**15**	**17**	**2**	0	123

The average HDD of Cluster 3 is 34.2, which means that Cluster 3 consists of very cold days from January to April and October to December. The average HDD of Cluster 4 is 13.8, which means that Cluster 4 consists of cool days throughout the whole year. Since the testbed uses gas for heating, heating does not heavily affect electricity usage. However, it can be inferred that the household might use extra electrical appliances for heating during very cold days, based on the differences between Cluster 3 and Cluster 4.

Over the course of the year (365 days) measured, most recorded days fall into Cluster 3 or Cluster 4 (149 for Cluster 3 and 123 for Cluster 4) since Michigan has a predominantly cool climate. Although Cluster 1 and Cluster 2 contain data from fewer unique days (28 for Cluster 1 and 65 for Cluster 2), their electricity usage for cooling is noticeable. For each cluster, the ratio of weekdays to weekend days is 5:2, which shows that there is no distinct difference between weekdays and weekends among the clusters.

5.5.3. Descriptive Analysis

The machine learning algorithms predict appliances using sensor-measured data, the activity-appliance table (Table 5-2), the ATUS activities, and the RECS appliance estimated activities. Respondent data with similar conditions to the sensor data household are selected from the ATUS and the RECS for further descriptive analysis.

5.5.3.1. ATUS: Activity

The mode values of the ATUS activity codes (in the form of the original ATUS 3rd tier activity codes) from the selected samples are calculated for each timestamp. Figure 5-7 illustrates the mode

activities by time, separated by weekdays and weekends. The weekday pattern is simpler than weekend pattern.

Figure 5-7. Mode Activities of the Selected ATUS Samples

Table 5-6 explains the weekday activities shown in Figure 5-7 and their associated appliances. Energy usage–related activity codes (Mo, 2018)(Chapter 2) are mapped to each of the ATUS 3rd tier codes, and the associated appliances are specified. In the selected samples, the respondents usually sleep until 7 AM and work until 6 PM. Then they eat dinner, take care of children, and

watch TV. They go to bed around 10 PM. The main appliances related to these activities are the computer, kitchen appliances, and TV, and their use times follow the activities.

Table 5-6. Weekday Activities and Appliances

Start Time	End Time	ATUS 3rd tier Code		Energy Usage-Related Activity		Appliance (Electricity)
		Code	Description	Code	Activity	
4:00	6:59	010101	Sleeping			
7:00	17:59	050101	Work, main job	EF01	Work for job(s) /research/homework	Computer
18:00	18:44	110101	Eating and drinking	BB03	Food and drink preparation	Oven, Stove, Toaster, Blender, Cofffee machine, Cooker, etc.
18:45	18:59	030101	Physical care for children	CD01	Physical care for children	
19:00	21:44	120303	Television and movies	LL01	Television	TV
21:45	3:59	010101	Sleeping			

Table 5-7. Weekend Activities and Appliances

Start Time	End Time	ATUS 3rd tier Code		Energy Usage-Related Activity		Appliance (Electricity)
		Code	Description	Code	Activity	
4:00	10:27	010101	Sleeping			
10:28	10:59	070101	Grocery shopping			
11:00	13:09	050101	Work, main job	EF01	Work for job(s) /research/homework	Computer
13:10	13:44	110101	Eating and drinking	BB03	Food and drink preparation	Oven, Stove, Toaster, Blender, Cofffee machine, Cooker, etc.
13:45	14:58	050101	Work, main job	EF01	Work for job(s) /research/homework	Computer
14:59	15:29	120303	Television and movies (not religious)	LL01	Television	TV
15:30	16:10	050101	Work, main job	EF01	Work for job(s) /research/homework	Computer
16:11	16:29	120101	Socializing and communicating with others			
16:30	21:49	120303	Television and movies	LL01	Television	TV
21:50	3:59	010101	Sleeping			

Table 5-7 simplified the weekend activities and appliances shown in Figure 5-7. The selected respondents sleep late until around 10 AM and then go grocery shopping. They spend the daytime working, eating, socializing, and watching TV. The main appliances used are the computer,

kitchen appliances, and TV. Their use times follow the activities, which are different by the weekdays.

Previously, clustering analysis was performed with the ATUS, and 6 occupant clusters were identified (Mo, 2018)(Chapter 3). Occupant Cluster 1 is the weekday activity pattern of the respondents who have jobs (Figure 5-2). The mode activities of the selected samples and the activities of Occupant Cluster 1 are compared in Figure 5-8. They show similar patterns with typical working hours (from approximately 7:00 AM to 5:30 PM). Activities during 1380 minutes out of 1440 minutes (96%) are identical in the selected samples and Occupant Cluster 1. This shows that the activity patterns of occupants who have jobs during weekdays are highly patterned.

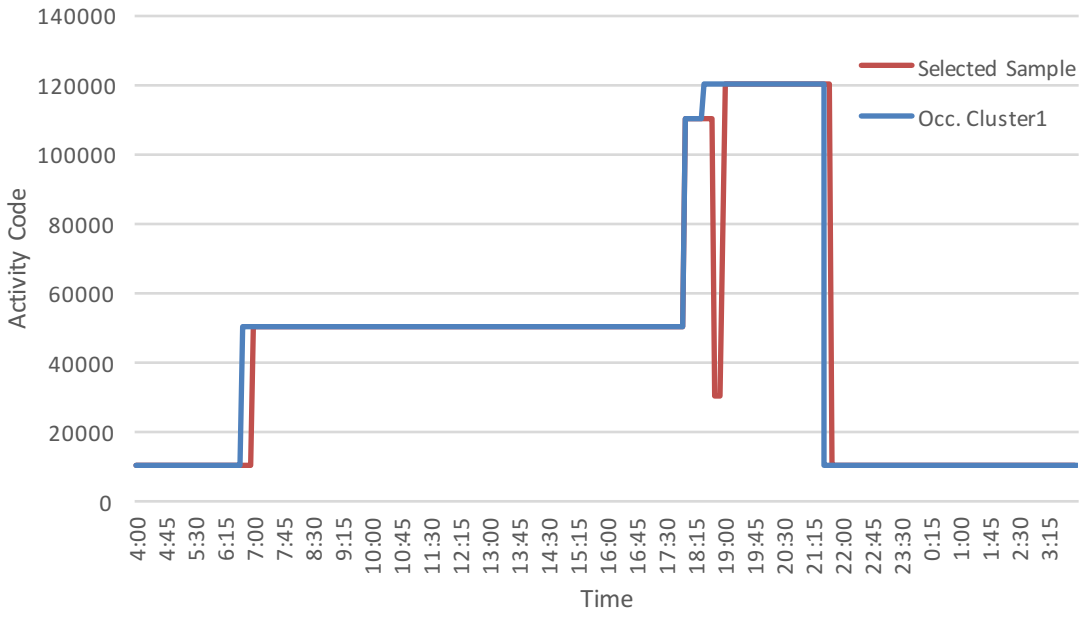


Figure 5-8. Weekday Activities

Occupant Clusters 2, 4, and 10 (Mo, 2018)(Chapter 3) are the weekend activity patterns (Figure 5-2). The mode activities of the selected samples and these Occupant Clusters are compared in Figure 5-9. The selected samples and the Occupant Clusters are different from one another, which

suggests that occupant activities during weekends are less patterned than they are during weekdays. However, while the occupants in the selected samples and Occupant Cluster 1 mostly have jobs, it is not clear if most of the occupants in Occupant Clusters 2, 4, and 10 have jobs or not.

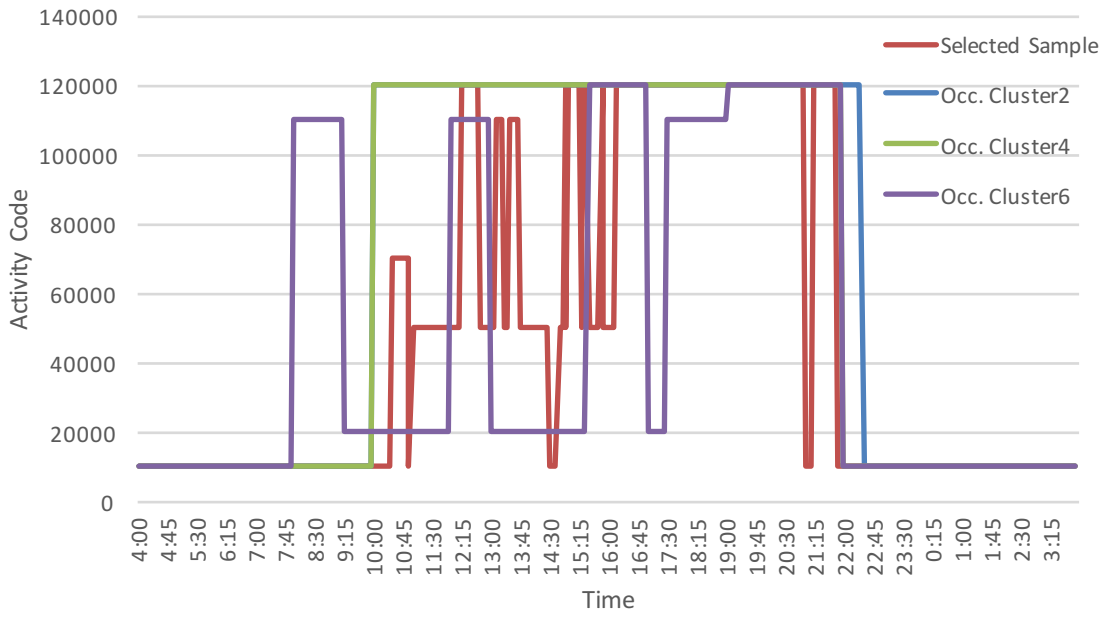


Figure 5-9. Weekend Activities

Table 5-8 summarizes the energy usage–related activities and their associated appliances with their wattage value range. The appliance wattage information is collected from several government and commercial sources (DaftLogic, 2018; Generac, 2018; HES, 2018; WholesaleSolar, 2018), and the mean, minimum, and maximum values are determined from the sources. Water heaters, air conditioners (central), clothes dryers (electricity), and stoves/ovens have high wattage values, which means that they have high impacts on the total electricity usage in the household. This list can help identify specific appliances from the sensor-measured event data.

Table 5-8. Energy Usage-Related Activities and Appliances

Code	Activity	Appliance	Mean	Min	Max
AA01	Washing, dressing, and grooming	Hair Dryer	1300	700	2500
		Shaver	15	15	20
BB01	Interior cleaning	Vacuum	847	200	2000
BB02	Laundry	Washing Machine	1005	300	3400
		Clothes Dryer: Elec	**4793**	1000	4000
		Clothes Dryer: Gas	914	300	2500
		Iron	1125	1000	1500
BB03	Food and drink preparation	Refrigerator/Freezer	775	150	2900
		Refrigerator (Room)	70	70	70
		Freezer	550	500	600
		Microwave	1103	600	1700
		Stove	2100	2100	2100
		Stove/Oven	**3000**	3000	3000
		Oven	1675	1200	2150
		Toaster	1111	800	1800
		Toaster Oven	1300	1200	1500
		Coffee Maker: Brew	1013	600	1500
		Coffee Maker: Warm	75	70	80
		Espresso Machine	580	360	800
		Electric Kettle	1660	1200	3000
		Blender	443	300	1000
		Rice Cooker	313	200	450
BB04	Kitchen and food clean-up	Dishwasher	1530	1200	3000
BB05	Heating and cooling	**Air Conditioner: Central**	**5592**	1500	15200
		Air Conditioner: Window	2833	750	13000
		Furnace Fan Blower	1510	700	3150
		Space Heater	1788	150	4000
		Water Heater	**11250**	4500	18000
		Ceiling Fan	67	25	120
		Table Fan	25	10	45
		Box Fan	200	200	200
BB06	Gardening, ponds, pools, and hot tubs	Mower	1400	1000	1500
		Pool Heater	275	275	275
		Spa (on-demand elec)	5500	5500	5500
LL01	Television	TV	223	150	500

** *Mean, Min (minimum), Max (maximum) in watt*

5.5.3.2. RECS: Energy and Appliance

Electricity usage and appliance information are derived from the selected RECS households that are similar to the household used for the sensor-measured data. Table 5-9 summarizes the yearly electricity usage of the selected households. Overall, cooling takes 12% of the total electricity

usage, and heating takes 10%. Households with gas heating use 13% of their total electricity consumption on cooling, and household with electrical heating use 8% of their total electricity consumption on cooling. It is notable that households with electrical heating use 30% of their total electricity consumption on heating. Most households use gas for heating (100 out of 133 households).

Table 5-9. Yearly Electricity Usage of the Selected RECS Samples

Selected Households	kWh			Percentage		Number
	Total	Cooling	Heating	Cooling	Heating	
All	11249.6	1303.2	1073.7	12%	10%	133
Gas Heating	10218.1	1377.5	0.0	13%	0%	100
Electricity Heating	14375.3	1078.3	4327.5	8%	30%	33

Table 5-10. Appliances of the Selected RECS Samples

Code	Description	Avg,Mode	Number or Category
***** (1) Selected Appliance Features***			
NUMFRIG	Number of refrigerators used	1.7	Number of refrigerators used
SIZRFRI1	Size of most-used refrigerator	4	1: Compact, 2: Small, 3: Medium, 4: Large, 5: Very large, -2: N/A
ICE	Through-the-door ice on most-used refrigerator	1	1: Yes, 0: No
NUMFREEZ	Number of separate freezers used	0.5	Number of separate freezers used
SIZFREEZ	Size of most-used freezer	-2	1: Compact, 2: Small, 3: Medium, 4: Large, 5: Very large, -2: N/A
STOVE	Number of separate cooktops	0.1	Number of separate cooktops
OVEN	Number of separate ovens	0.1	Number of separate ovens
MICRO	Microwave oven used	1.1	Number of microwave ovens
COFFEE	Coffee maker used	1	1: Yes, 0: No
APPOTHER	Other small appliance used	0	1: Yes, 0: No
DRYRFUEL	Fuel used by clothes dryer	5	1: Natural gas, 2: Propane, 5: Electricity
TVCOLOR	Number of televisions used	3.0	Number of televisions used
CABLESAT	Number of cable or satellite boxes without DVR	1	Number of cable or satellite boxes without DVR
COMBODVR	Number of cable or satellite boxes with DVR	1.1	Number of cable or satellite boxes with DVR
DESKTOP	Number of desktop computers	0.7	Number of desktop computers
NUMSMPHONE	Number of smart phones	2.5	Number of smart phones
NUMCFAN	Number of ceiling fans used	2.9	Number of ceiling fans used
LGTINNUM	Number of light bulbs installed inside the home	2	1: <20, 2: 20-39, 3: 40-59, 4: 60-79, 5: >=80
LGTOUTNUM	Number of light bulbs installed outside the home	1	0: None, 1: 1-4, 2: 5-9, 3: >=10

Table 5-10 (cont'd)

***** (2) Additional General Appliances***			
STOVEN	Number of stoves	1.0	Number of stoves
TOAST	Toaster used	1	1: Yes, 0: No
TOASTOVN	Toaster oven used	0	1: Yes, 0: No
CROCKPOT	Crockpot or slow cooker used	0	1: Yes, 0: No
FOODPROC	Food processor used	0	1: Yes, 0: No
RICECOOK	Rice cooker used	0	1: Yes, 0: No
BLENDER	Blender or juicer used	0	1: Yes, 0: No
DISHWASH	Have dishwasher	1	1: Yes, 0: No
CWASHER	Have clothes washer in home	1	1: Yes, 0: No
DRYER	Have clothes dryer in home	1	1: Yes, 0: No
PLAYSTA	Number of video game consoles	1.1	Number of video game consoles
DVD	Number of DVD players	1.1	Number of DVD players
VCR	Number of VCRs	0.3	Number of VCRs
NUMLAPTOP	Number of laptop computers	1.5	Number of laptop computers
NUMTABLET	Number of tablet computers	1.6	Number of tablet computers
ELPERIPH	Number of printers, scanners, fax machines, or copiers	0.9	Number of printers, scanners, fax machines, or copiers
CELLPHONE	Number of other cell phones	0.4	Number of other cell phones
MOISTURE	Humidifier used	0	1: Yes, 0: No
NUMWHOLEFAN	Number of whole house fans used	0.1	Number of whole house fans used
LGTINCAN	Portion of inside light bulbs that are incandescent	4	1: All, 2: Most, 3: About half, 4: Some, 0: None
LGTINCFL	Portion of inside light bulbs that are CFL	4	1: All, 2: Most, 3: About half, 4: Some, 0: None
LGTINLED	Portion of inside light bulbs that are LED	0	1: All, 2: Most, 3: About half, 4: Some, 0: None
ESCWASH	Energy Star qualified clothes washer	1	1: Yes, 0: No
ESDISHW	Energy Star qualified dishwasher	1	1: Yes, 0: No
ESDRYER	Energy Star qualified clothes dryer	1	1: Yes, 0: No
ESFREEZE	Energy Star qualified freezer	-2	1: Yes, 0: No, -2: N/A
ESFRIG	Energy Star qualified refrigerator	1	1: Yes, 0: No
ESLIGHT	Energy Star qualified lightbulbs	1	1: Yes, 0: No
ESWATER	Energy Star qualified water heating	0	1: Yes, 0: No
ESWIN	Energy Star qualified windows	0	1: Yes, 0: No

Table 5-10 specifies the average numbers or mode values of appliances that the selected households own. It provides a list of common appliances and their numbers in the households that are similar to the sensor data sample. The table has 2 sections: (1) appliances that are selected

features for the effective prediction of electricity usage (Mo, 2018)(Chapter 4), and (2) additional common appliances in general households.

5.6. Discussion and Conclusion

The features derived from the Occupant Behavior Prediction Model predicted residential appliances with 96% accuracy using a Decision Tree algorithm. It implies that the daily appliance usage and associated activities of a household are quite well patterned, and can be precisely predicted. This information can be further used to set efficient energy saving strategies for a household by analyzing the impact and usage time of the appliances and associated activities. Clustering analysis provided further energy consumption characteristics of the households by identifying and analyzing the days and times when energy was used. Daily energy consumption by minute-level interval is clustered with 4 groups, which is mainly influence by CDD and HDD (hot, warm, cool, and cold days). It shows that energy consumption for cooling and heating have strong influences on total energy consumption in residential buildings. Also, the minute-level daily energy consumption can be used for detailed energy strategies for the households having similar conditions with this testbed. Additional descriptive analysis of the ATUS and the RECS supplemented the sensor-measured data by providing detailed activity schedules and appliance lists from households that were similar to the household in the sensor data sample.

There are also some limitations to the datasets. First, the sensor does not identify appliances exactly. It identifies the differences between appliances, but it does not precisely recognize if, for example, an appliance is a coffee maker or a toaster. It still requires appliance names to be manually identified. As a result, the appliances from the sensor data have arbitrary names such as Appliance

12, Heating Element 7, etc. Since the names of the appliances are important to estimate the activities associated with them, the ambiguous appliance names are a barrier to precisely predicting activities by time. Once this limitation is resolved, activities can be predicted with other methodologies such as machine learning.

The individual datasets, the ATUS, the RECS, and the sensor-measured data have different data formats. The ATUS collects activities from several respondents, and records a daily diary from a single person doing one activity per timestamp during one specific day. It lacks a record of other household members' activities and simultaneous activities. While the ATUS data reflect individual activities with state-level demographic information, the RECS data are household-level with census division–level demographic information. In this study, the 2015 RECS and the 2015 ATUS are used, as they were the most recent matching years when this study started. Although the ATUS collects individual time-series activity data and the RECS collects household yearly survey data, many studies extract the structures and important concepts from both datasets and use them for further analysis.

Future research will conduct more refined behavior prediction. Once the appliance names are clearly identified with more enhanced NILM and manual detection in the sensor-measured event data, the time-series appliance data can be mapped with the time-series electricity usage. More advanced methodologies can be applied to these minute-level datasets, which can predict more precise time-series activities.

CHAPTER 6

SUMMARY AND CONCLUSION OF THE RESEARCH

6.1. Summary of Research

The purpose of this research is to identify a relationship between energy consumption and occupant behavior in detail and with consideration of building technology, and to build a model to predict behavior based on energy data using machine learning approaches, which can be potentially used to create efficient building operation and control strategies.

At the beginning of the study, the Occupant Behavior Prediction Model was developed, and this model was applied to the national survey data, the ATUS data, and the sensor-measured data (Figure 6-1). In order to define the structural relationship between occupant behavior, building technology, and energy usage, the ATUS, RECS, and sensor data are analyzed with several methods including machine learning classification, numeric prediction, clustering (K-modes clustering and K-means clustering). GIS was also used for spatial analysis and geographical representation.

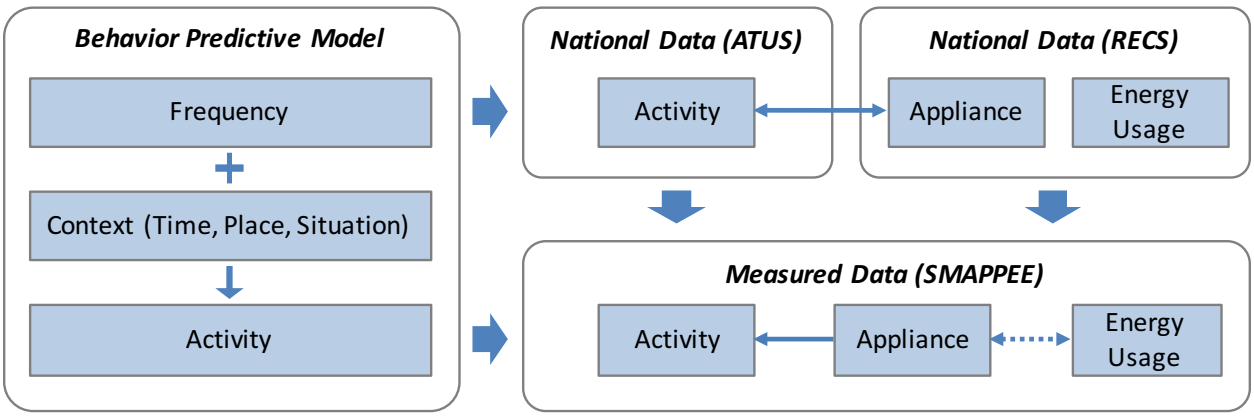


Figure 6-1. Summary of the Research

6.2. Summary of Findings

In Part I (Chapter 2), the Occupant Behavior Prediction Model was developed based on habitual behavior studies. The model was applied to the ATUS data, and findings include the prediction of occupant energy usage–related activities and behaviors with the components of the Occupant Behavior Prediction Model, and the identification of habitual energy usage–related activities. The Occupant Behavior Prediction Model can predict occupant behavior with overall 64% accuracy for the ATUS dataset, and its accuracy can reach up to 83% for a subgroup of habitual activities. Notably, the model shows 99% accuracy for predicting washing, dressing, and grooming activity and 82% accuracy for predicting watching television activity.

In Part II (Chapter 3), occupant clusters' daily routines of activities by time are derived from the ATUS data, and the time ranges of major habitual energy usage–related activities are identified. In the latter sections of this chapter, the influences of major factors (occupant clusters, day of the week, gender, region) on habitual energy usage–related activities are identified. GIS analysis identified the geographical pattern of the selected energy usage–related activities. Watching TV is one of the most habitual energy usage-related activities, and it is included in 5 clusters. Based on the overlapping time from these 5 clusters, the occupants watch TV around from 18:30 to 21:30, and it means that one of the most habitual energy usage-related activities strongly tend to happen during this time. Day of the week, gender, and job status have strong influences on the difference of energy usage-related activities.

In Part III (Chapter 4), features with significant impacts on energy consumption are selected from the RECS data, and energy consumption is predicted with the selected features. The model's

prediction performances with all features vs. with selected features are compared, and the effectiveness of the selected features are measured. The findings include the efficient features to predict energy consumption, and the prediction of total energy consumption in residential buildings. The main selected features are refrigerator, freezer, oven and television from the Appliance category, TV, cloth dryer, and swimming pool usage from the Behavior category, housing type, number of rooms from Technology category, and number of household members and number of young (under 18 years old) household members from the Demographic category. The selected 32 features predict the total electricity consumption with 78% accuracy, which almost reaches 80% accuracy with all 271 features. It shows that the selected features keep 98% of the prediction power compared to all of the 271 features.

In Part IV (Chapter 5), the Occupant Behavior Prediction Model is applied to the household sensor-measured dataset. This chapter synthesized the findings from Parts I through III. Using machine learning approaches, appliances and associated activities are predicted with electricity consumption data. The findings are as follows. The appliance names are predicted with 96 percent accuracy based on the Occupant Behavior Prediction Model using DT algorithm. Daily energy consumption by minute-level interval is clustered with 4 groups, which is mainly influence by CDD and HDD (hot, warm, cool, and cold days). It shows that energy consumption for cooling and heating have strong influences on total energy consumption in residential buildings. Also, the minute-level daily energy consumption can be used for detailed energy strategies for the households having similar conditions with this testbed.

6.3. Contributions

Unlike existing studies, which focused on predicting energy consumption based on occupant behavior, this study innovatively developed the reverse prediction model: predicting occupant behavior based on energy consumption. This model is valuable in that it provides more detailed and precise occupant behavior patterns including the daily schedule and the habitual characteristics of the occupant activities. The contribution of this research is described in Figure 6-2. The structured list and model from each research step will reveal detailed and dynamic interactions between occupant behavior, energy usage, and building technology. These findings can contribute to three different groups: residential occupants, industry companies, and researchers.

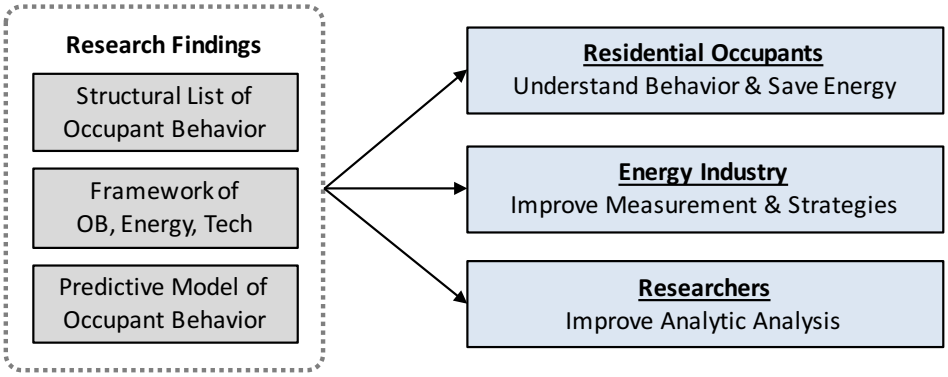


Figure 6-2. Research Contributions

First, this research will have an impact on residential occupant behavior by helping occupants better understand their own behaviors' effects on energy usage, and detect what changes would improve energy efficiency in their homes since the Occupant Behavior Prediction Model can explain the unique behavior pattern of each household based on their energy consumption data. The detailed breakdown of energy consumption will explain occupants' behavior patterns of heating, cooling, and appliance usage. Then, it will identify their most energy-consuming behaviors, which will help with setting effective energy saving strategies at the residential building.

In addition, the analysis will indicate which appliances use the most energy, and help occupants select energy-efficient appliances.

Second, the findings will be beneficial to energy-related industries. The Occupant Behavior Prediction Model can be applied to energy sensors and energy dashboards to improve measurement and analysis strategies. The most important behavior factors will improve the strategies around residential energy monitoring sensor development and placement by providing guides on what to measure, what kind of factors should be focused on, and how to measure them. For example, the findings provide where to install energy monitoring sensors to collect critical information to analyze occupant behavior and energy consumption. Furthermore, it can be used to optimize heating, cooling, and appliance schedules.

Third, the structured list of behaviors will enhance the methodology for future building energy research areas, such as statistical analysis, case studies, energy simulation, and, etc. The model will deepen understanding of occupant behavior with regard to residential energy usage, and will improve the analysis about energy and occupant behavior. In addition, the findings will be beneficial for national energy and time-use surveys, such as the RECS and the ATUS. It can be further used to develop more meaningful energy policy.

6.4. Intellectual Merit

Most of the occupant behavior studies used occupant activities and behaviors to predict energy consumption. However, this research approached the topic in a novel way, reversing past studies'

approaches and using the concept of habit to predict occupant behaviors with energy consumption data.

This research developed the Occupant Behavior Prediction Model, which has the potential to be used for efficient energy control strategies, occupant interventions, and education for energy savings. The model showed high performance when predicting occupant activities and behaviors, which was verified with two different types of datasets: the national survey data and the sensor-measured specific household energy consumption data. The study can be scaled up for larger datasets from households throughout a city or state.

6.5. Broad Impacts

The machine learning approaches used in this study can be utilized for other diverse studies. Especially, the Occupant Behavior Prediction Model which combined the concept of habit formation and the methods of machine learning has strong potential to be applied to various research fields. In building and construction domain, this model can be extended to other energy types, occupant types, and building types in the near future. In broader domains, this model has the potential to be further integrated with research in psychology, sociology, economics, and other fields.

6.6. Limitations

There are some limitations in this study, mainly related to the datasets. First, the sensor did not identify appliances exactly. It recognized that different appliances were separate, but could not identify what a given appliance actually was – for example, whether a small appliance was a coffee

maker or a toaster. It still required manual identification of the appliance names. As a result, the appliances from the sensor data have arbitrary names such as Appliance 12, Heating element 7, etc. Since the names of the appliances are important to estimate the activities associated with them, the ambiguous appliance names are a barrier to precisely predicting activities by time. Once this limitation is resolved, activities can be predicted with other methodologies such as machine learning.

The individual datasets, the ATUS, the RECS, and the sensor-measured data have different data formats. The ATUS collects activity data from several respondents, and records a daily diary for each person, with one activity per timestamp during one specific day. It lacks other household members' activities and simultaneous activities. While the ATUS data record individual activities with state-level demographic information, the RECS data are household-level data with census division–level demographic information. In this study, the 2015 RECS and the 2015 ATUS are used for consistent data collection years, as 2015 was the most recent matching year at the moment this study started. Since then, the 2017 RECS started to record state information, which will help with a more precise regional comparison between the ATUS and the RECS. Although the ATUS is individual time-series activity data and the RECS is household yearly survey data, many studies extract the structures and important concepts from these datasets and use them together for further analysis.

6.7. Future Research

Future research will conduct more refined behavior prediction using improved data quality. Also, future research will include more sensor-measured data from residential buildings – other types of

energy, more building technologies including heating, cooling, and ventilation systems (HVAC), and indoor environmental quality (IEQ). The model and data analysis methods can be expanded to fit larger areas and other types of buildings (commercial buildings, educational buildings, etc.).

Refining Behavior Prediction

Future research will conduct more refined behavior prediction. Once the appliance names are clearly identified with more enhanced nonintrusive load monitoring (NILM) and manual detection in the sensor-measured event data, the time-series appliance data can be mapped with the time-series electricity usage. More advanced methodologies can be applied to these minute-level datasets, and with more detailed data, they can predict more precise time-series activities.

Including Gas and Water Consumption

This research concentrated on occupant activities and behaviors that affected electricity consumption and appliances. Future research will include occupant behaviors' effects on gas and water consumption, and will include other renewable energy production (such as residential solar energy production with photovoltaic panels) if applicable.

Applying to HVAC Systems and IEQ

While this research focused on appliance usage and their associated occupant behaviors, future research will apply the Occupant Behavior Prediction Model to heating, cooling, and ventilation systems and indoor environmental quality criteria including thermal comfort, lighting, noise, air quality, etc.

Expanding to Broader Area for Measured Data

Future studies will collect more measured data of occupant behaviors and energy consumption at the city or state level. Data collection will include more detailed geographical location information, which will enable more precise GIS analysis.

Expanding to Other Types of Buildings and Occupants

The Occupant Behavior Prediction Model can be expanded to occupants in other types of buildings (commercial buildings, educational buildings, facilities for the elderly, etc.), and the results can contribute to improve their energy strategies considering occupant behavior in their facilities.

APPENDICES

APPENDIX A. GIS Analysis for Main Activities: All Maps

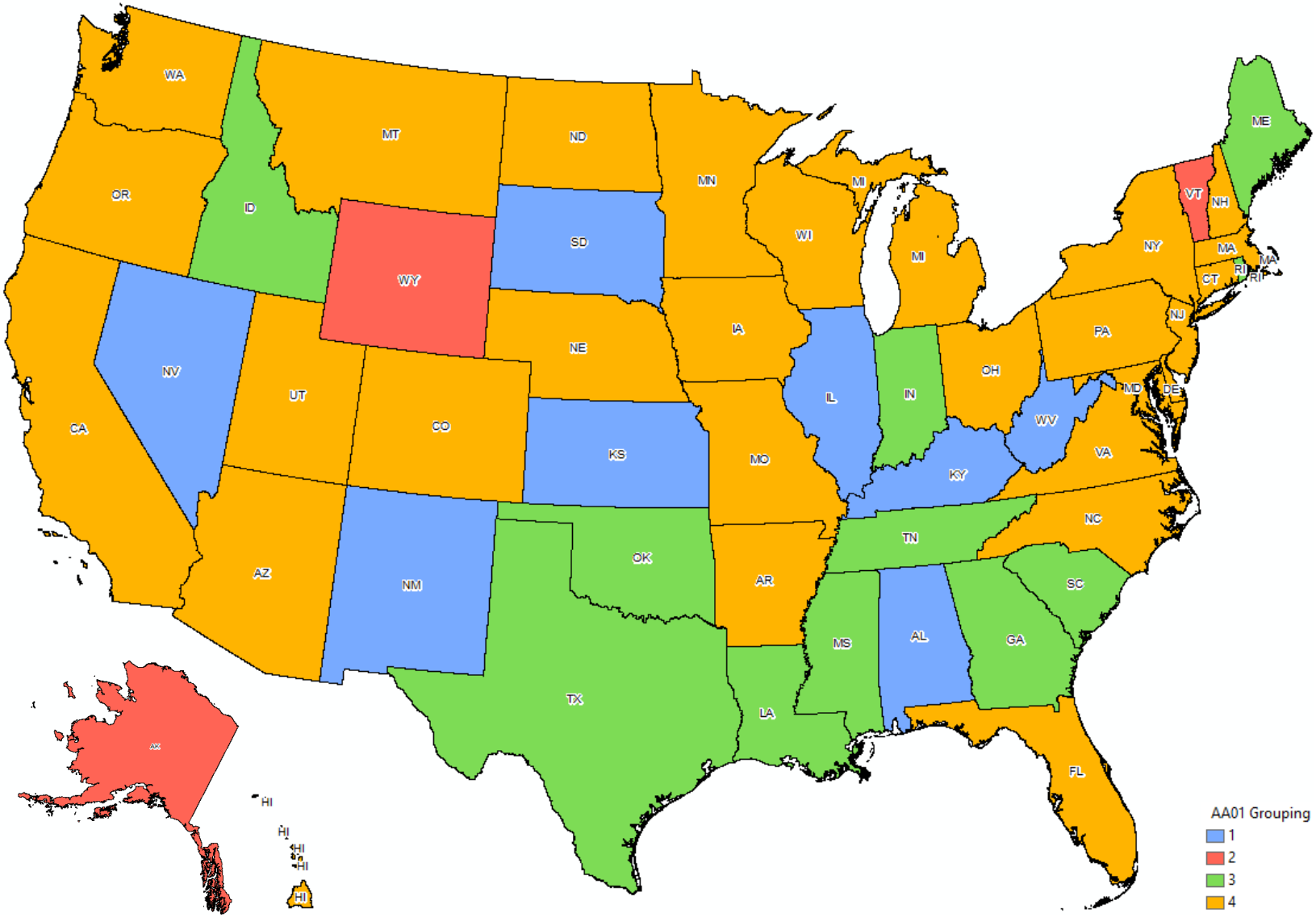


Figure A-1. AA01 State Clusters by Grouping Analysis

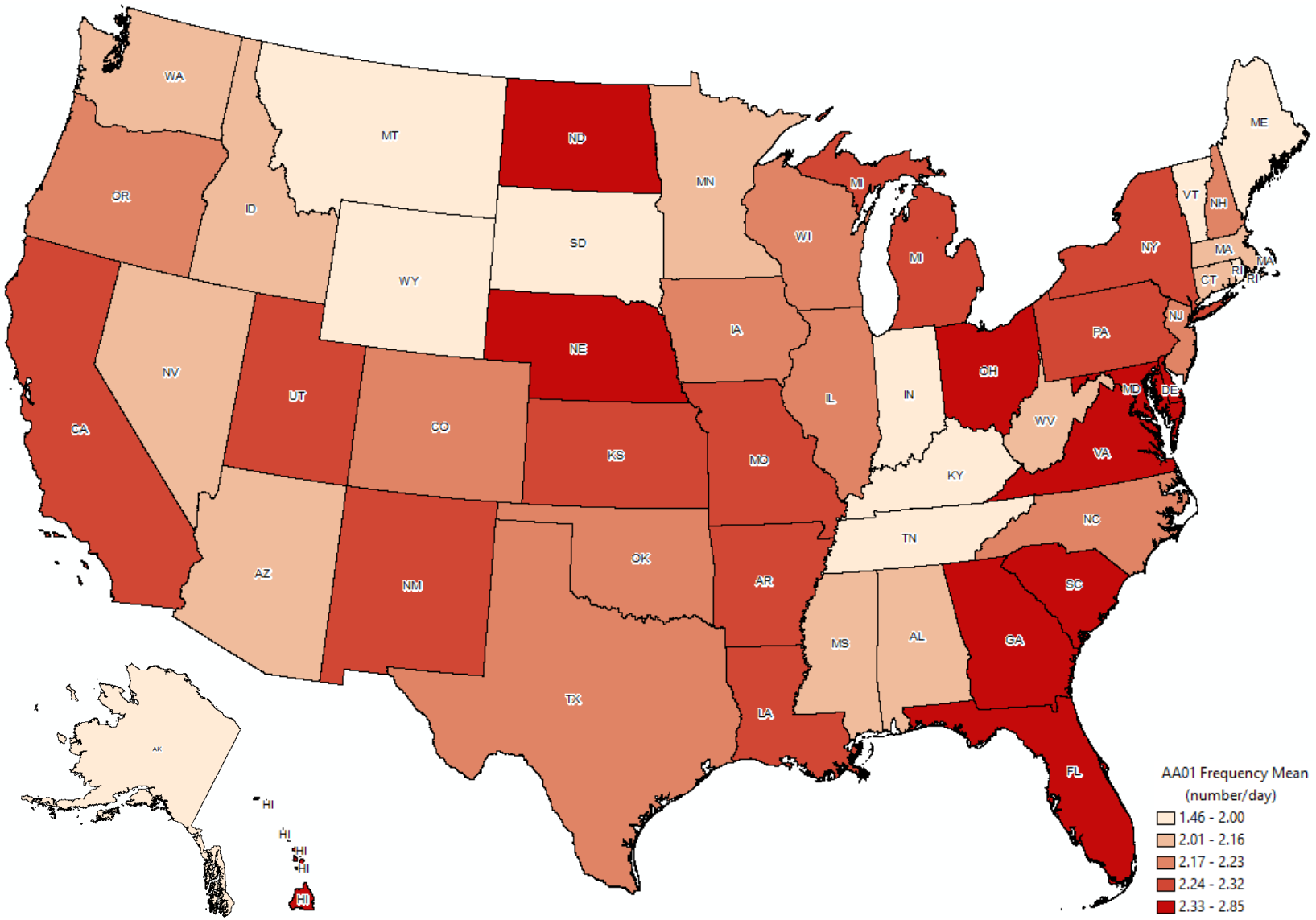


Figure A-2. AA01 Frequency by Quantiles

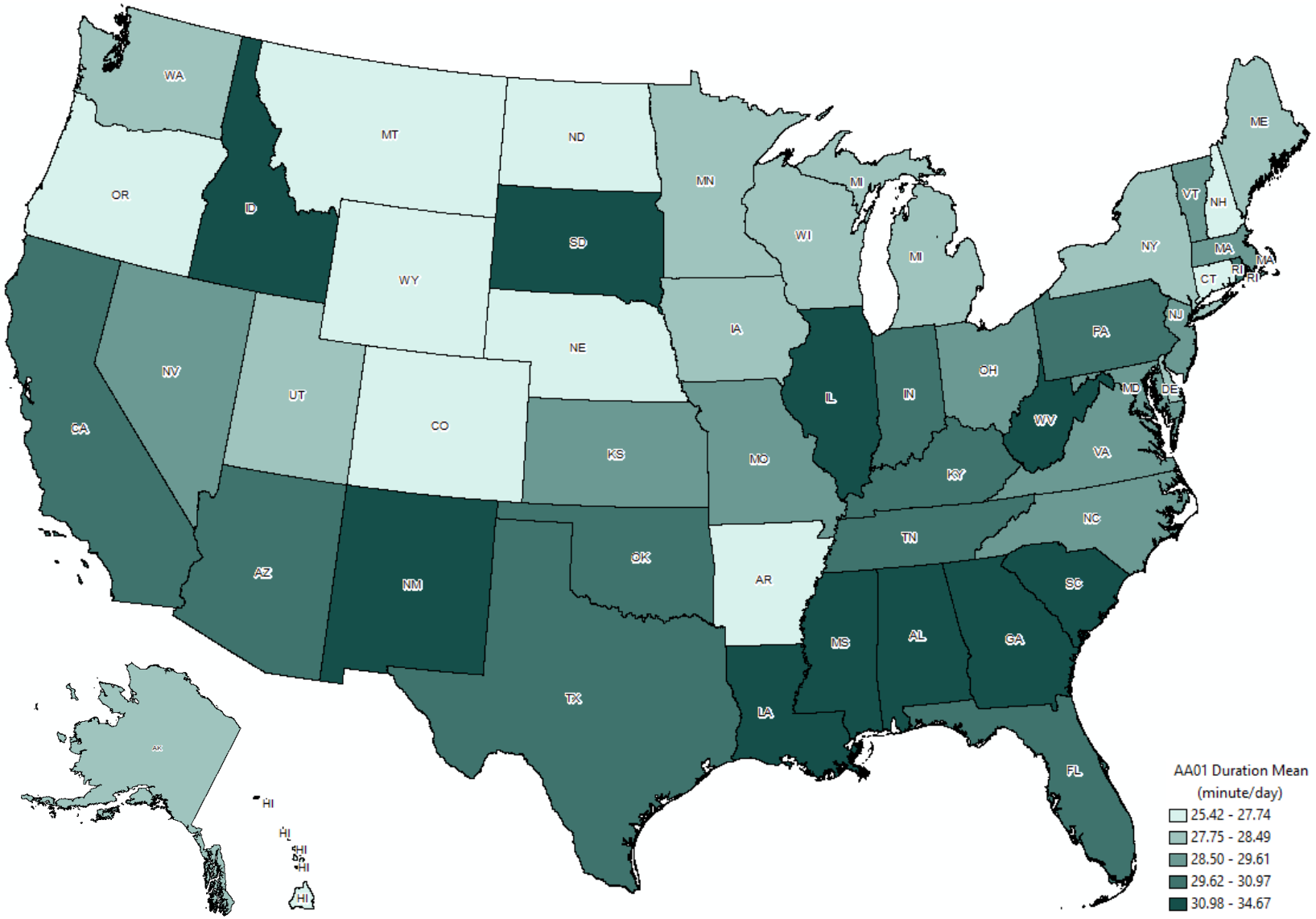


Figure A-3. AA01 Duration by Quantiles

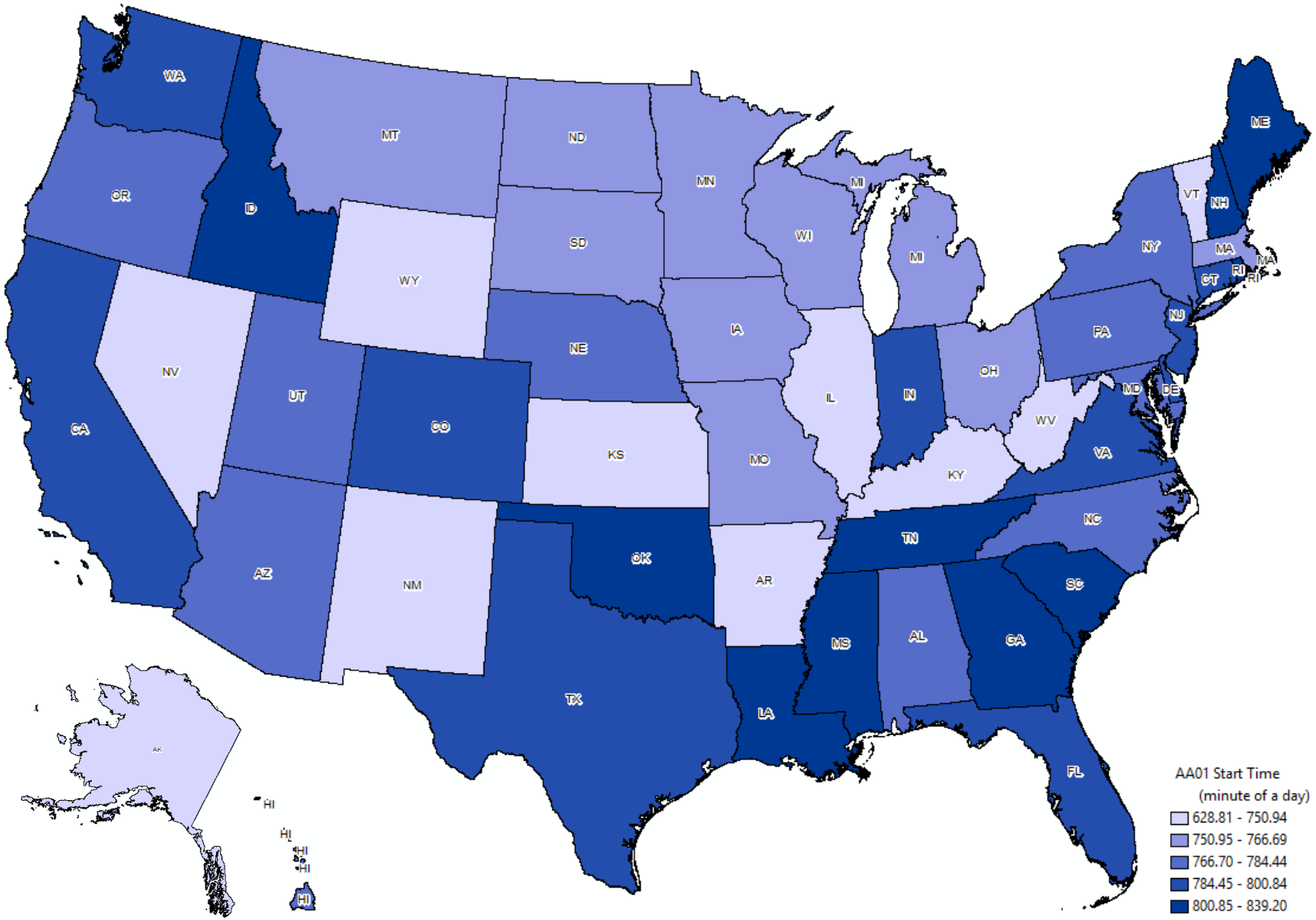


Figure A-4. AA01 Start Time by Quantiles

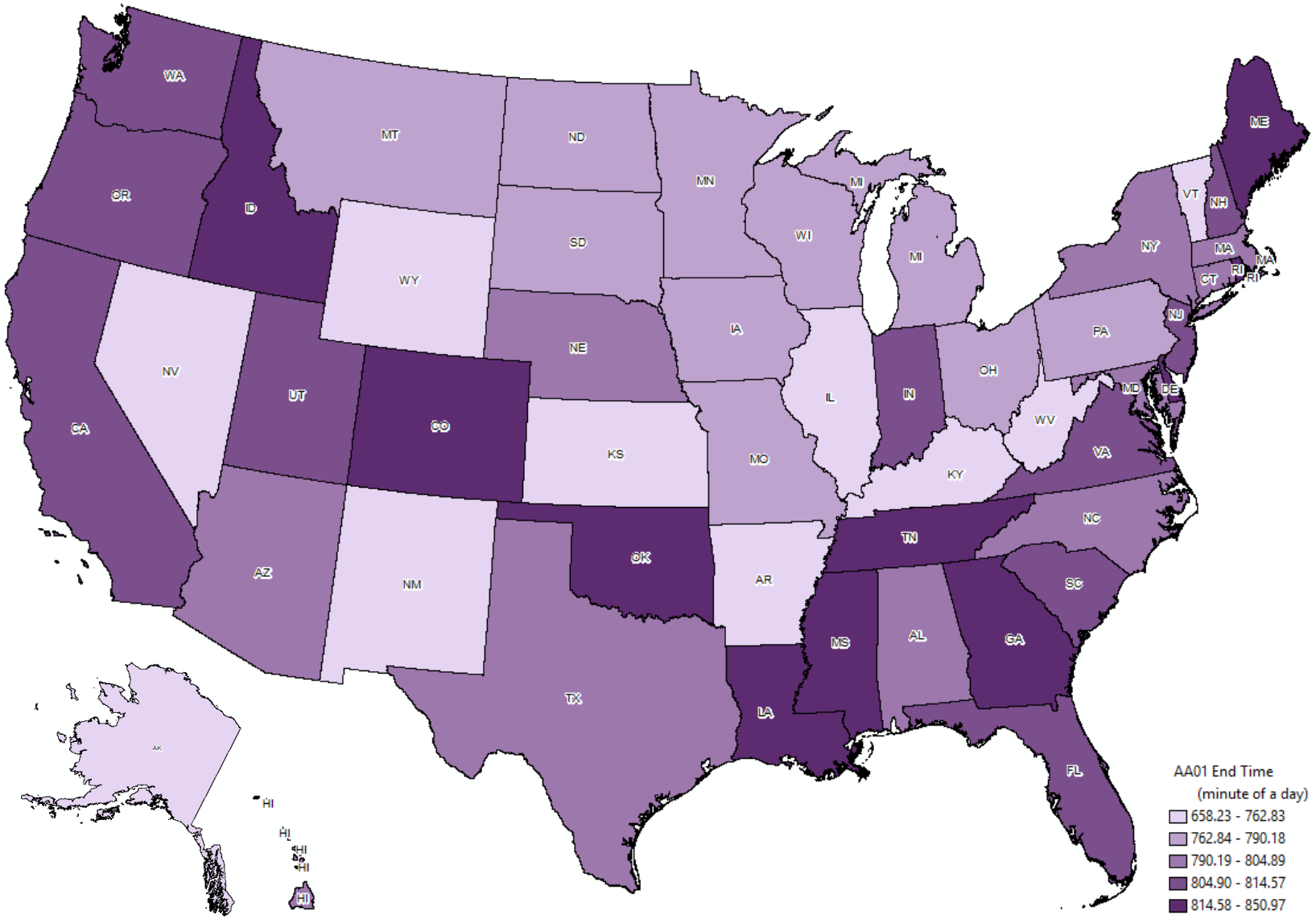


Figure A-5. AA01 End Time by Quantiles

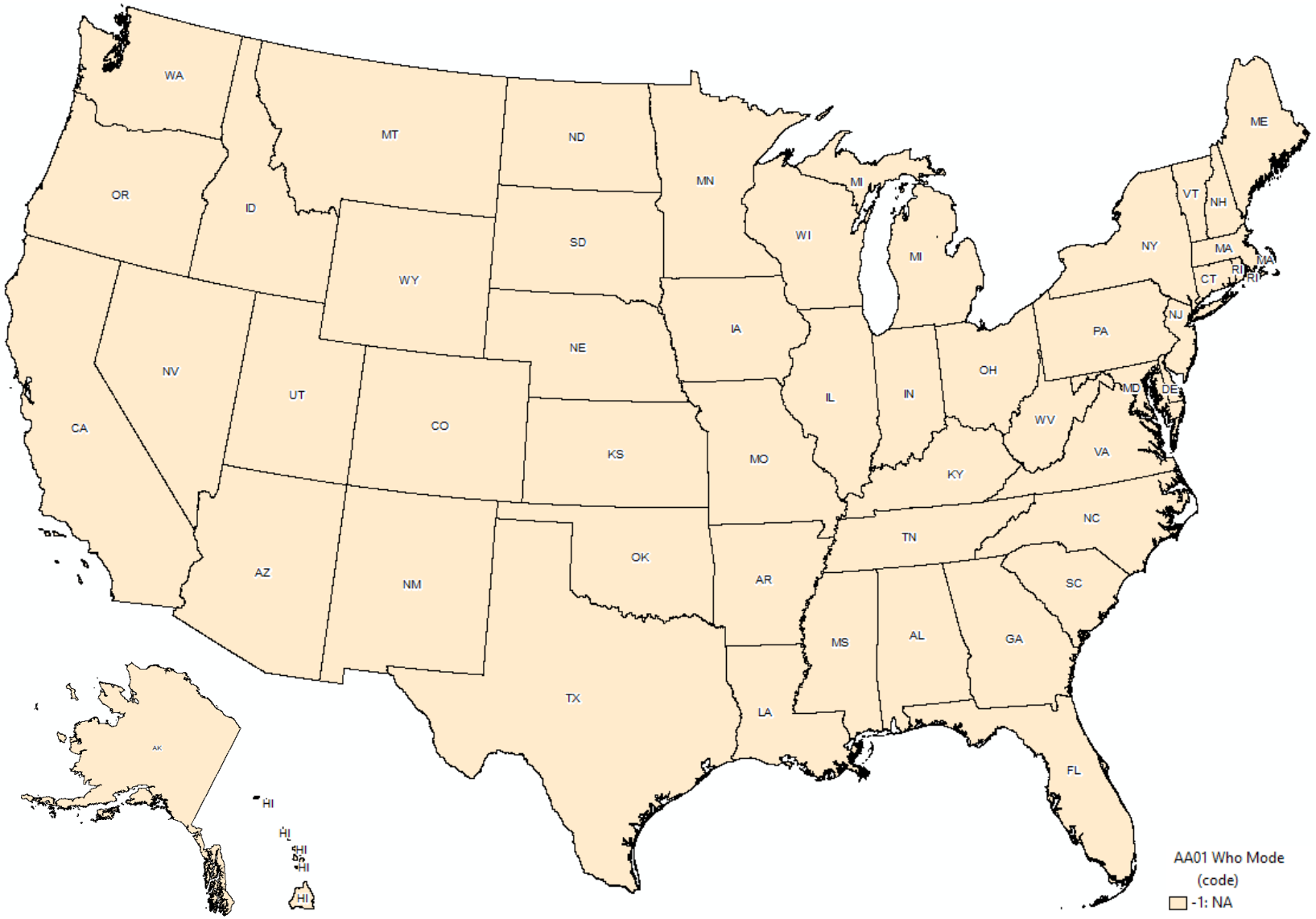


Figure A-6. AA01 Partner

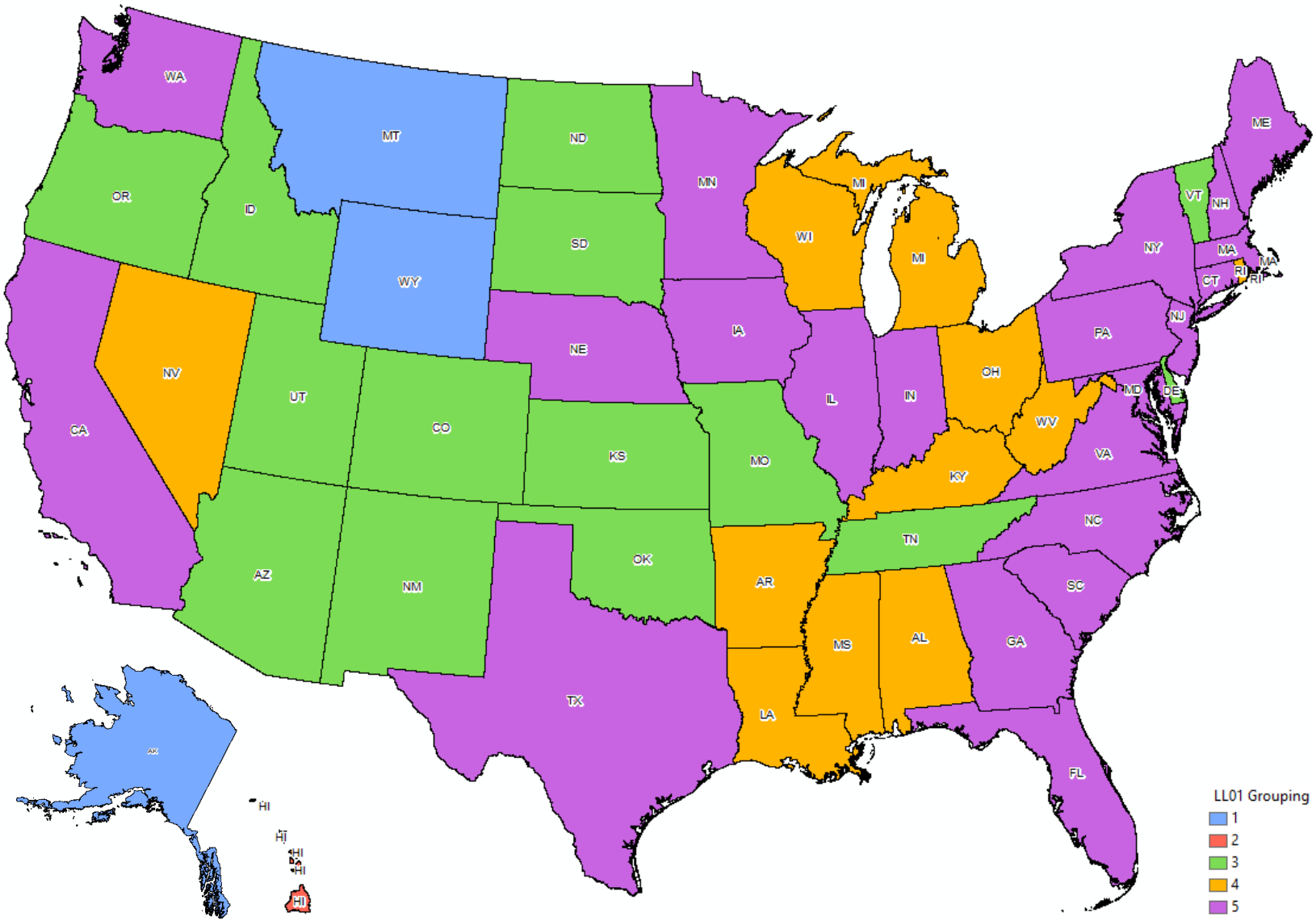


Figure A-7. LL01 State Clusters by Grouping Analysis

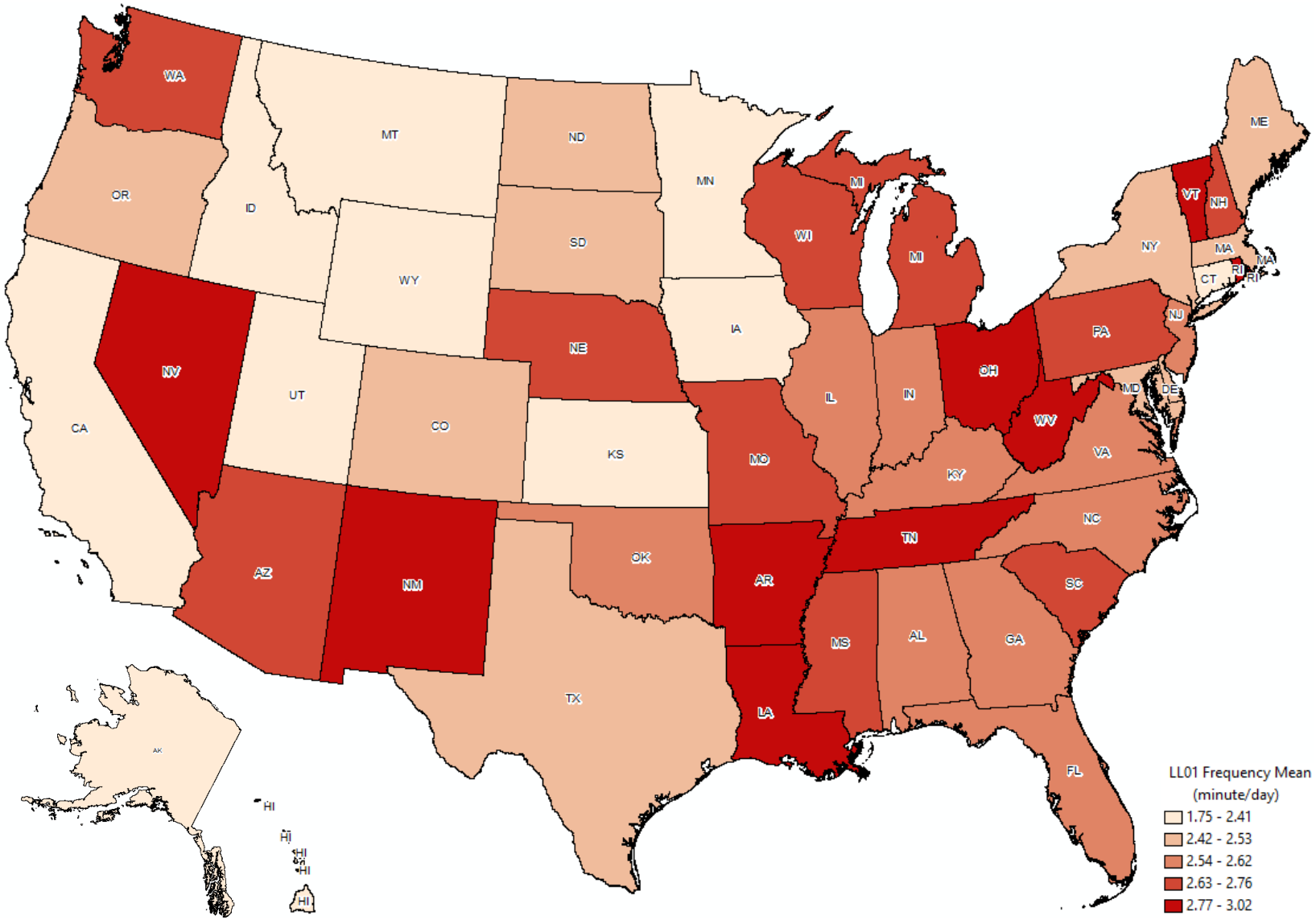


Figure A-8. LL01 Frequency by Quantiles

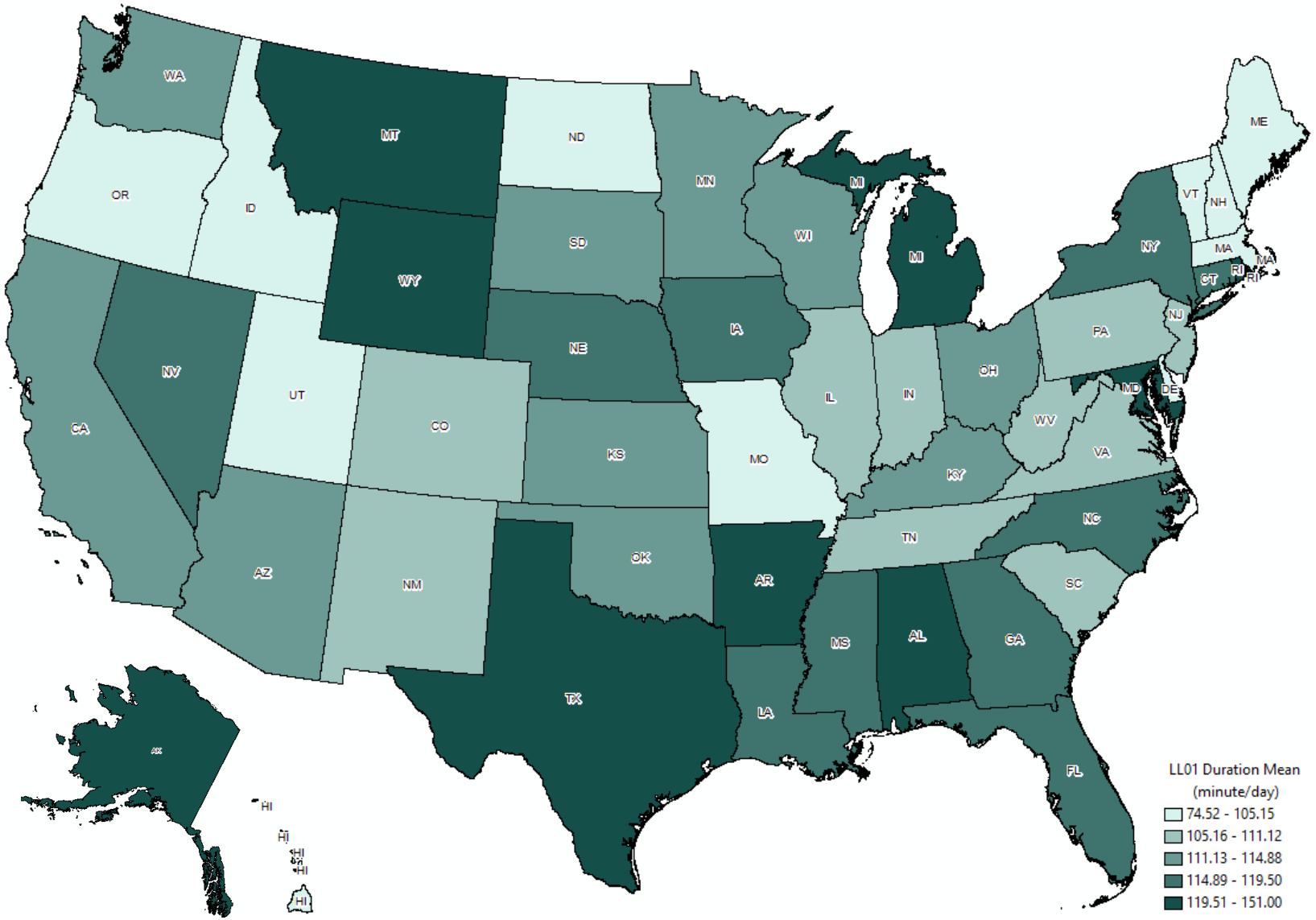


Figure A-9. LL01 Duration by Quantiles

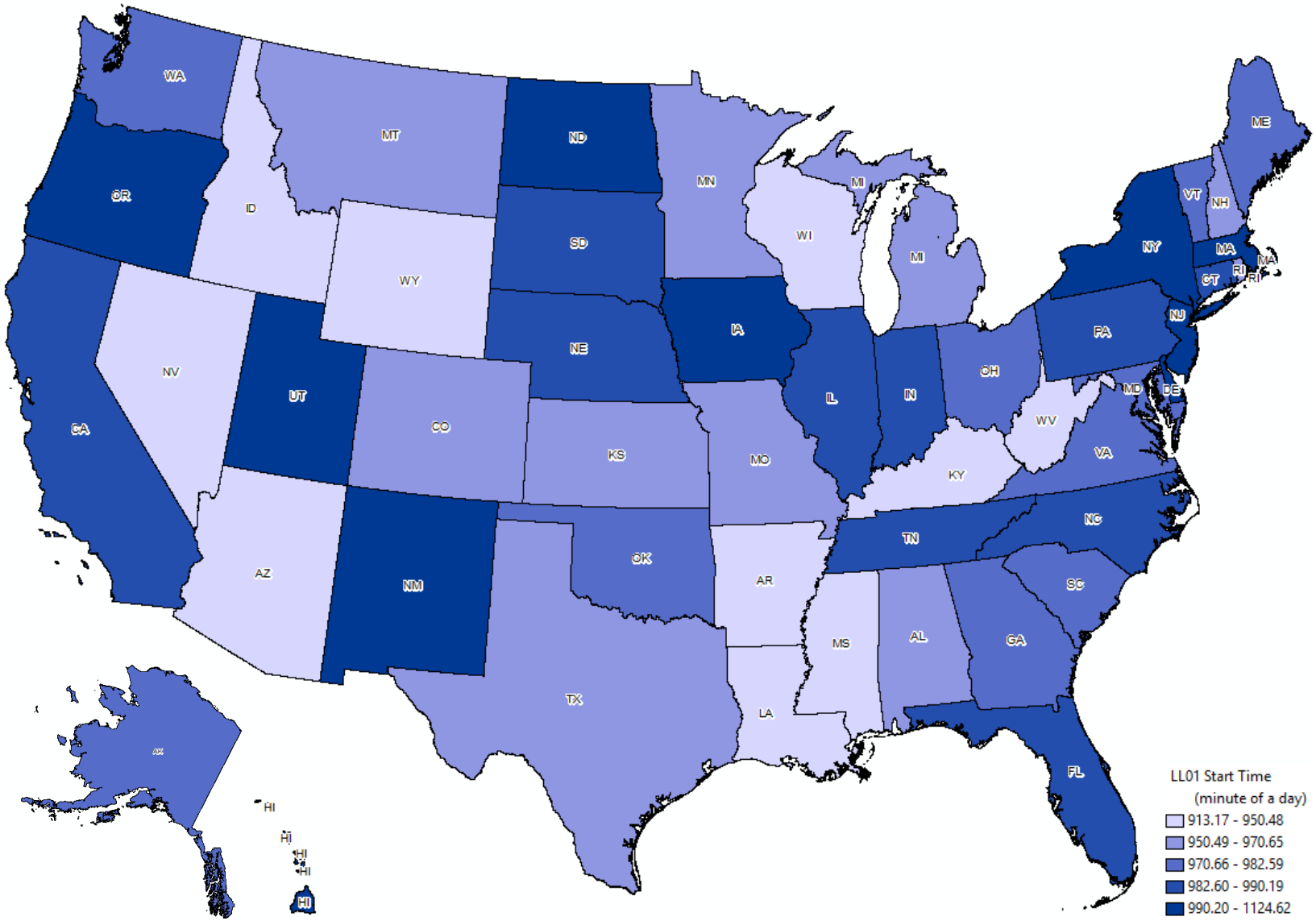


Figure A-10. LL01 Start Time by Quantiles

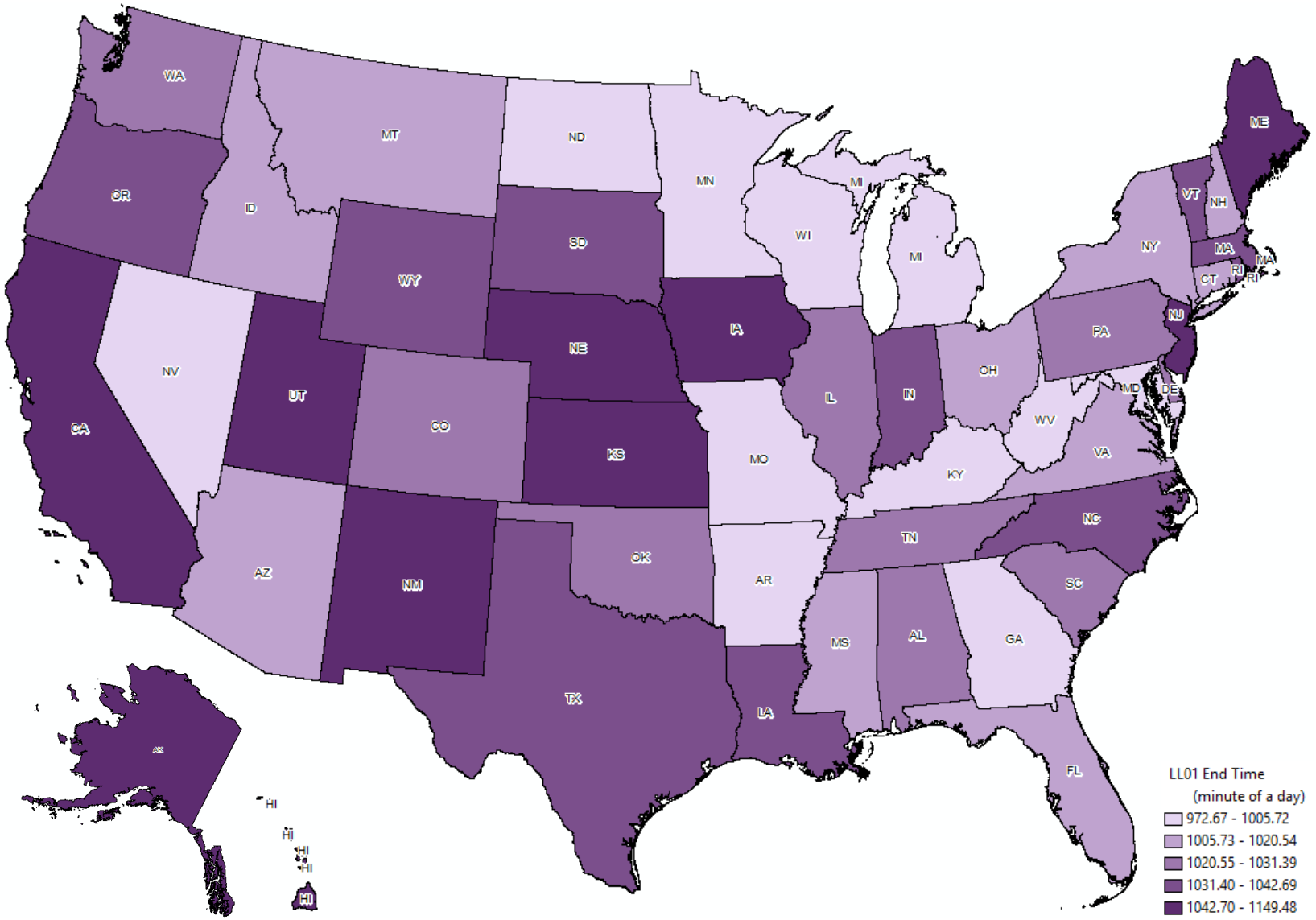


Figure A-11. LL01 End Time by Quantiles

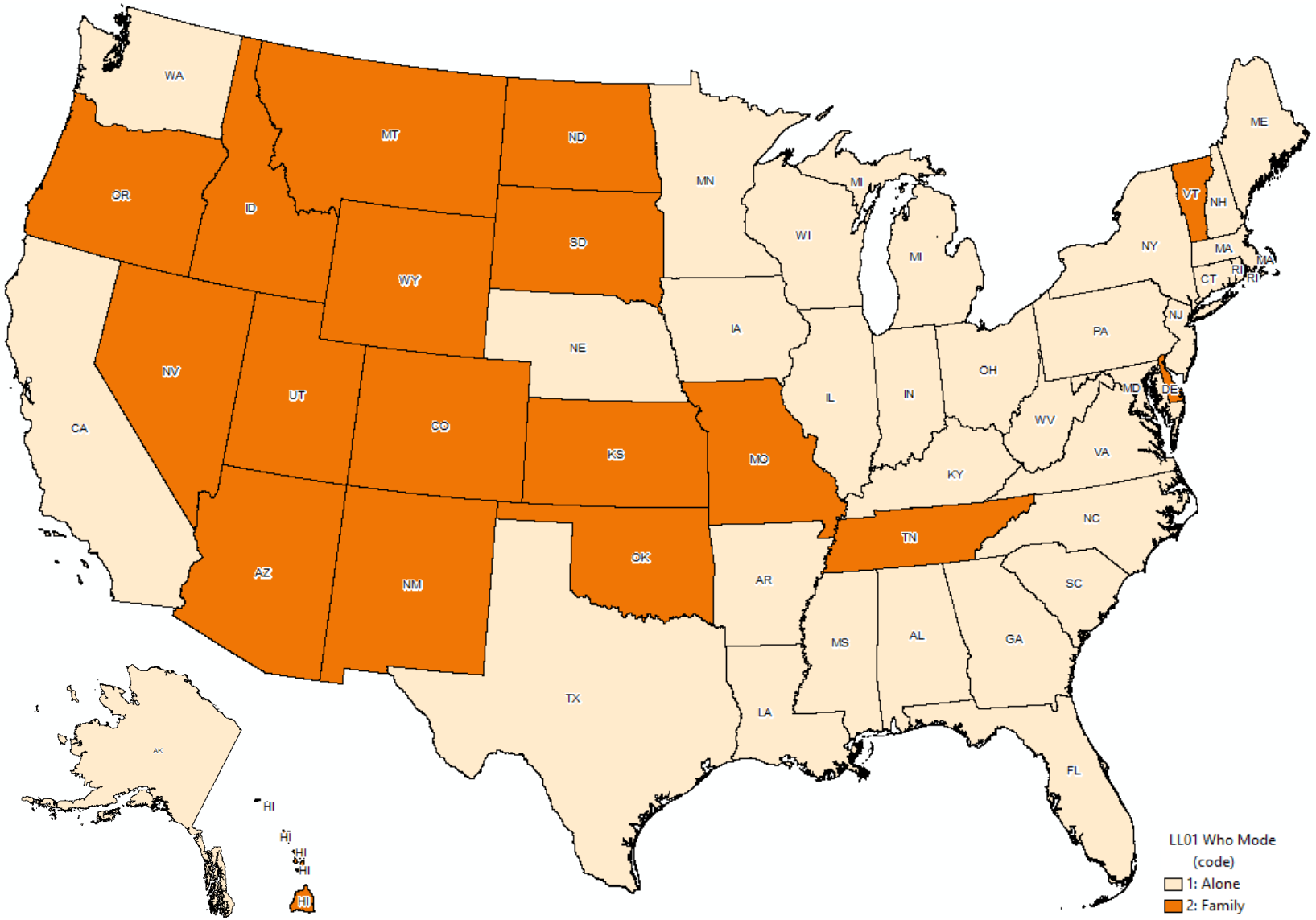


Figure A-12. LL01 Partner

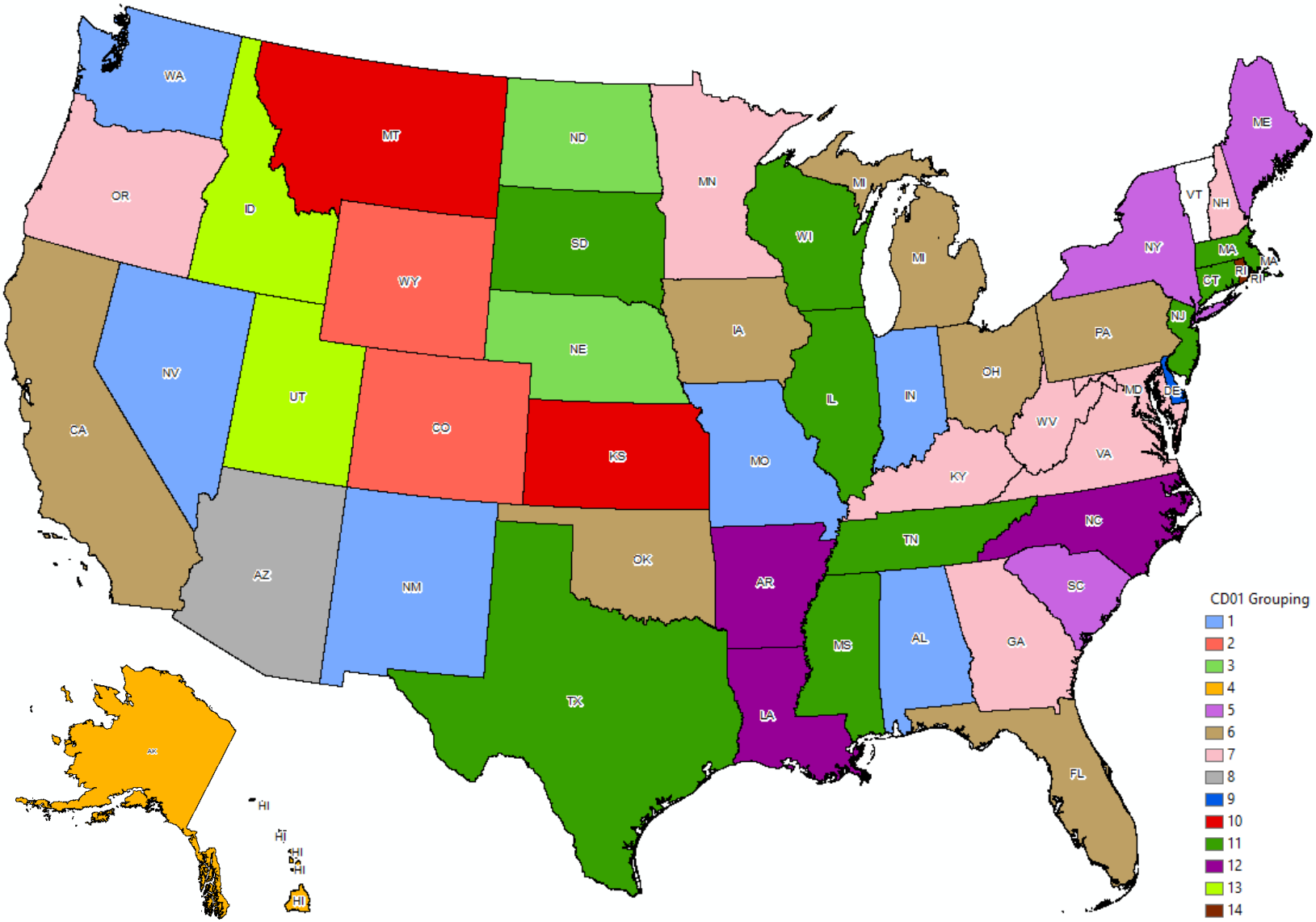


Figure A-13. CD01 State Clusters by Grouping Analysis

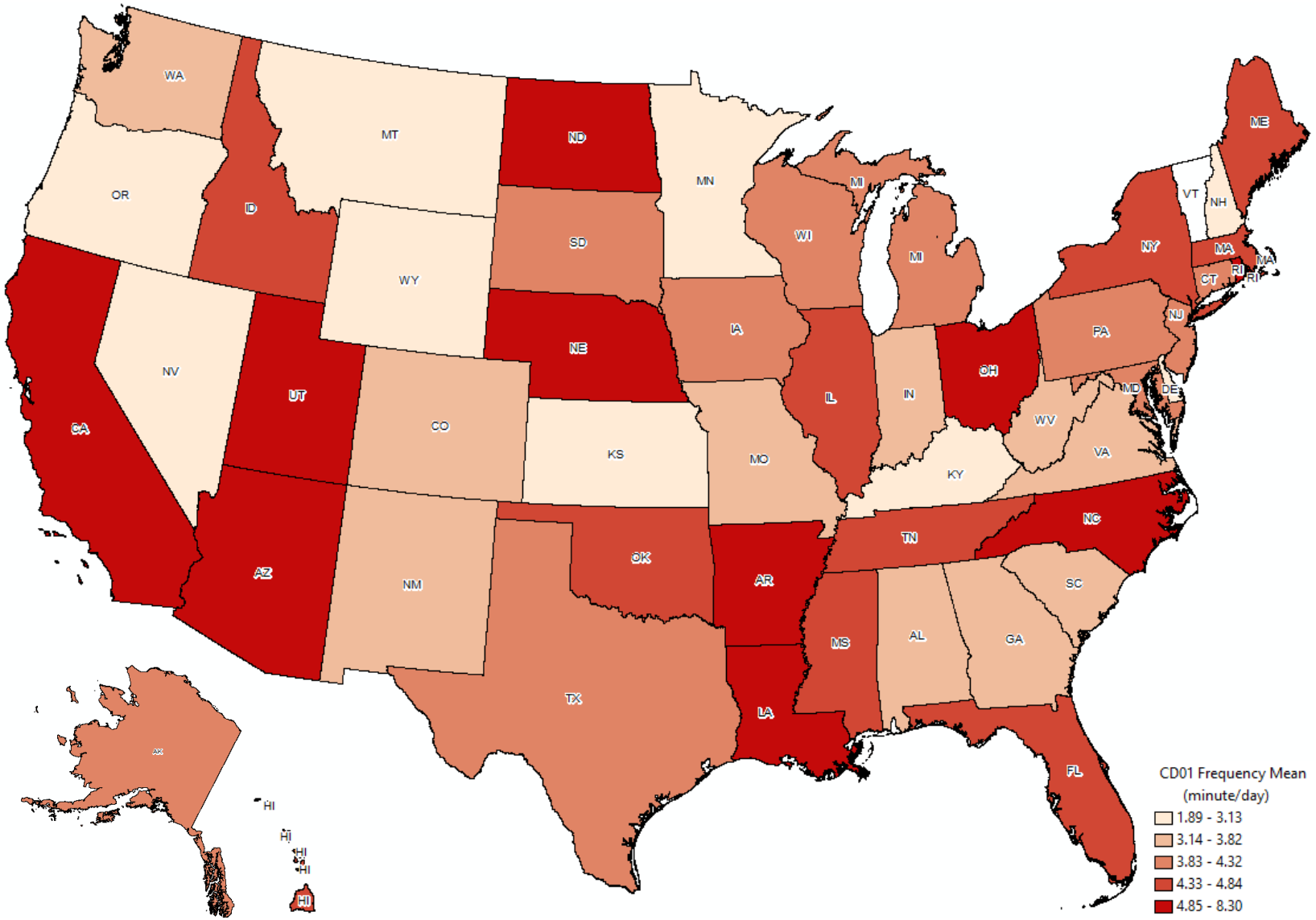


Figure A-14. CD01 Frequency by Quantiles

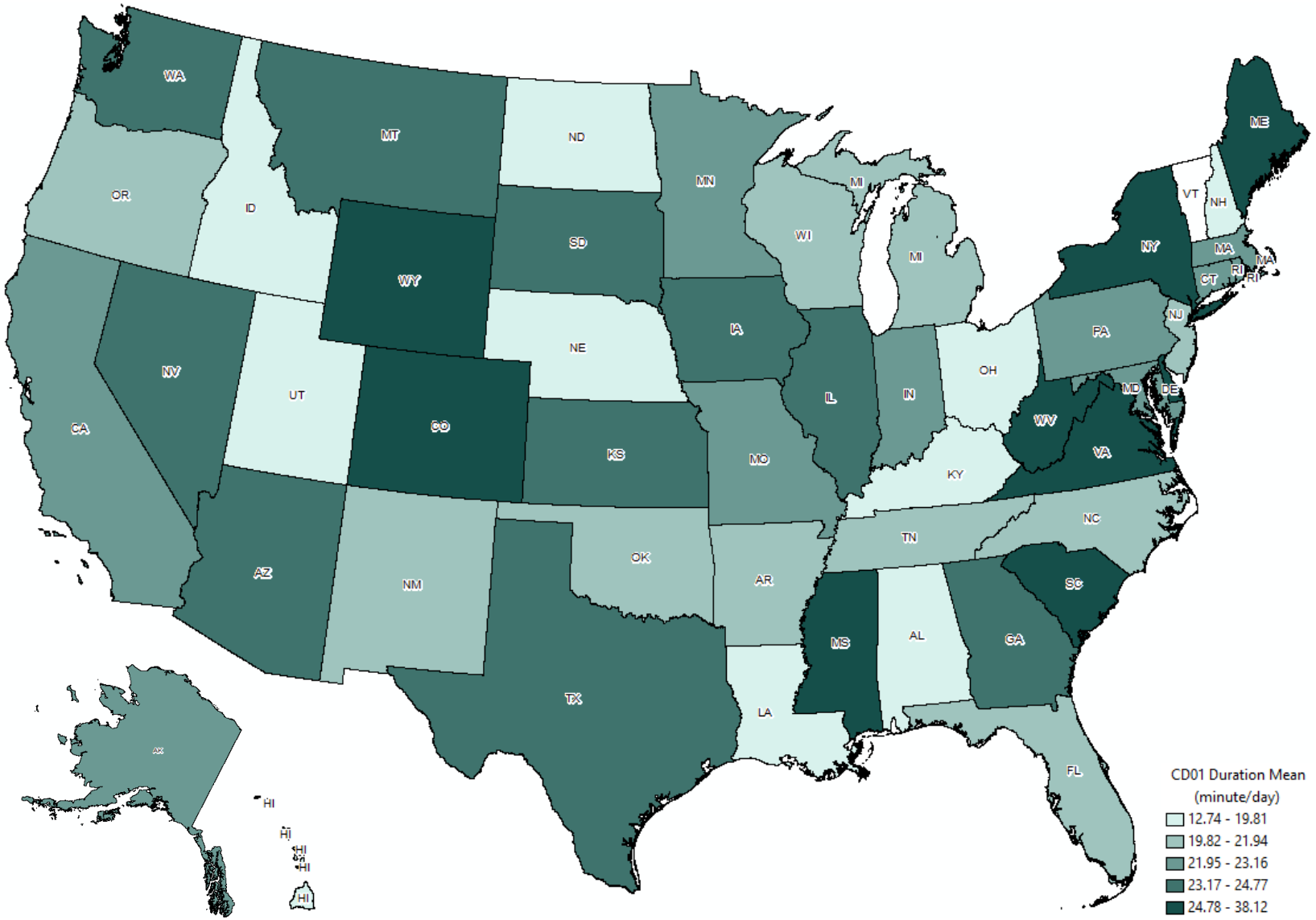


Figure A-15. CD01 Duration by Quantiles

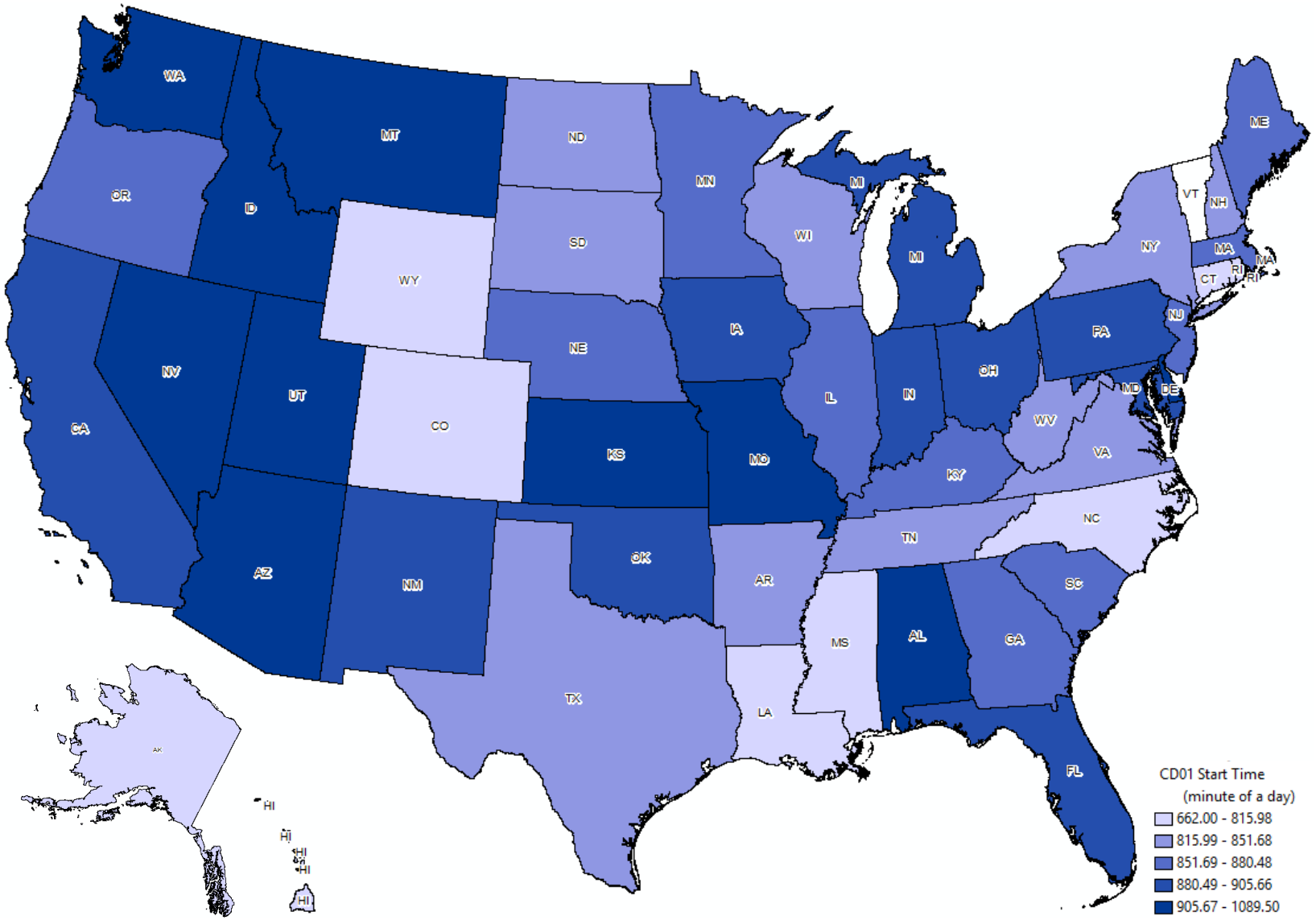


Figure A-16. CD01 Start Time by Quantiles

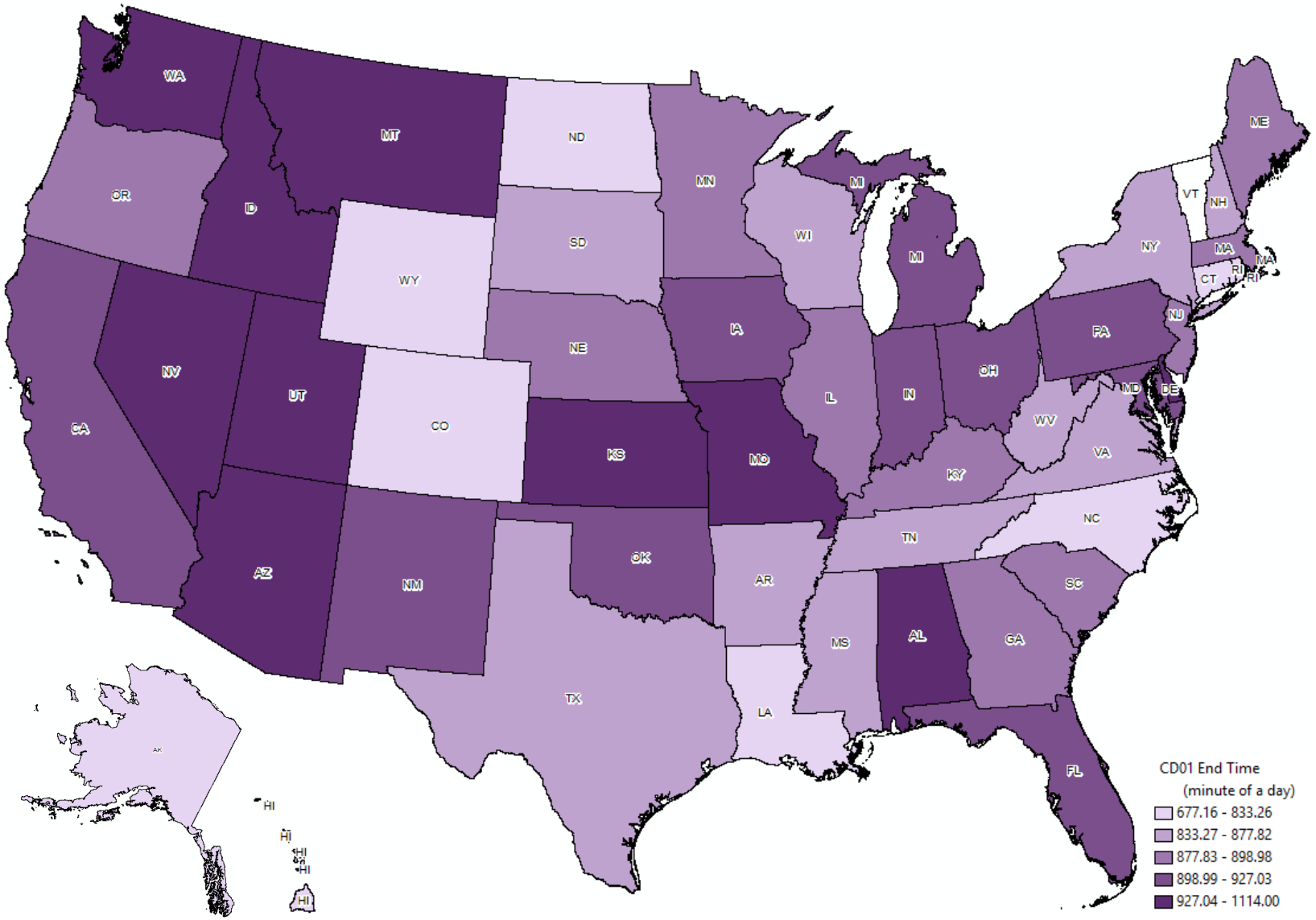


Figure A-17. CD01 End Time by Quantiles

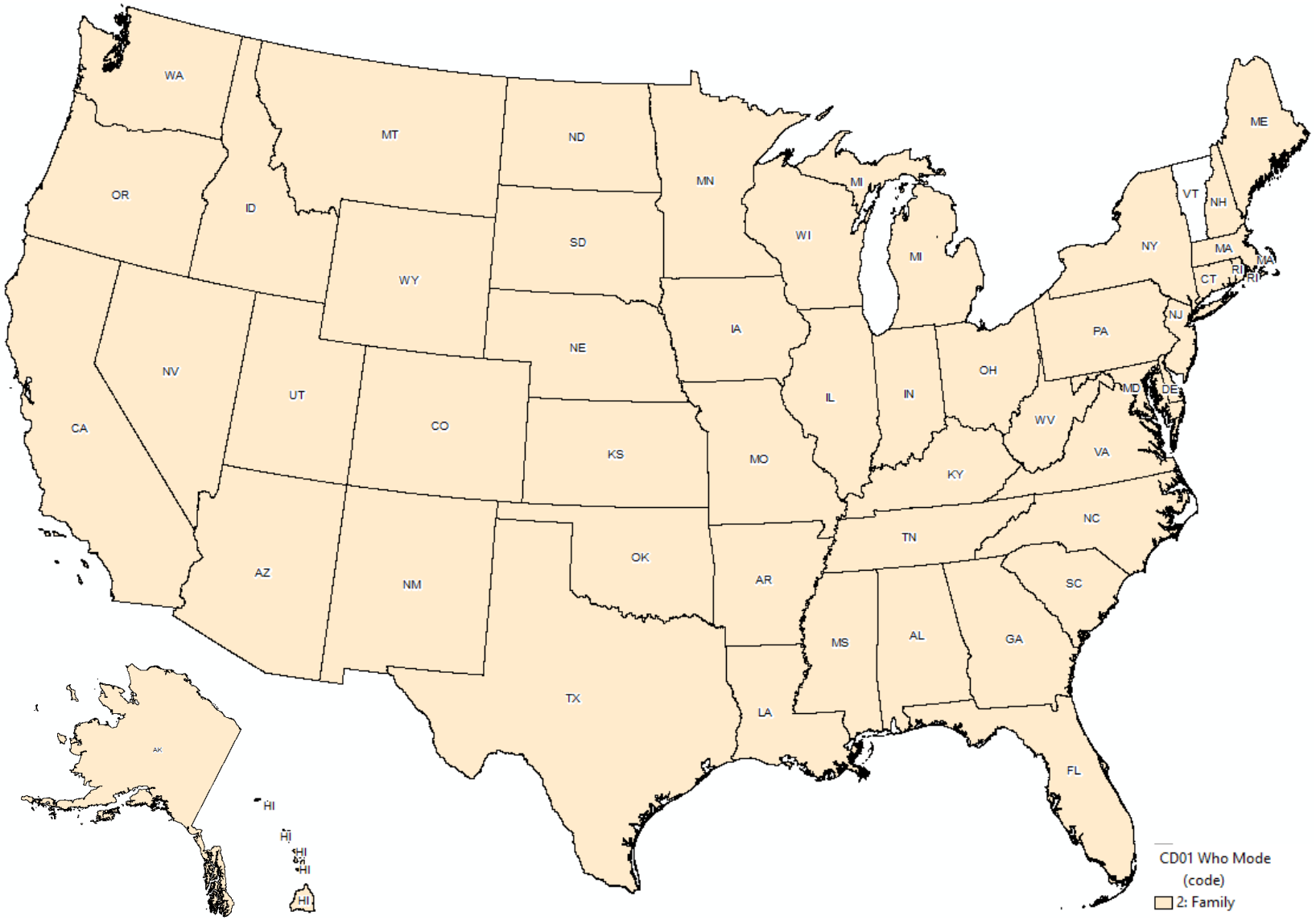


Figure A-18. CD01 Partner

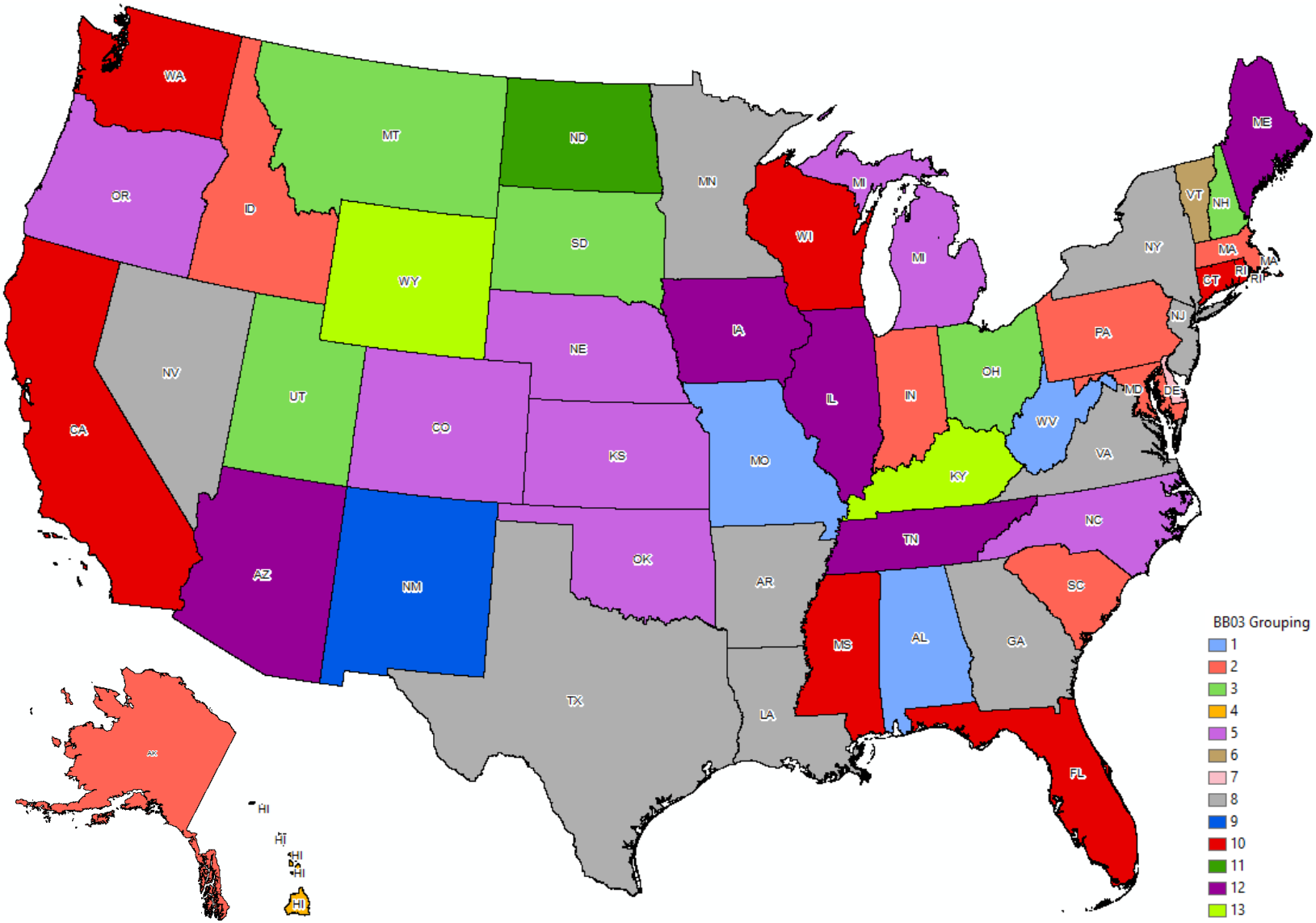


Figure A-19. BB03 State Clusters by Grouping Analysis

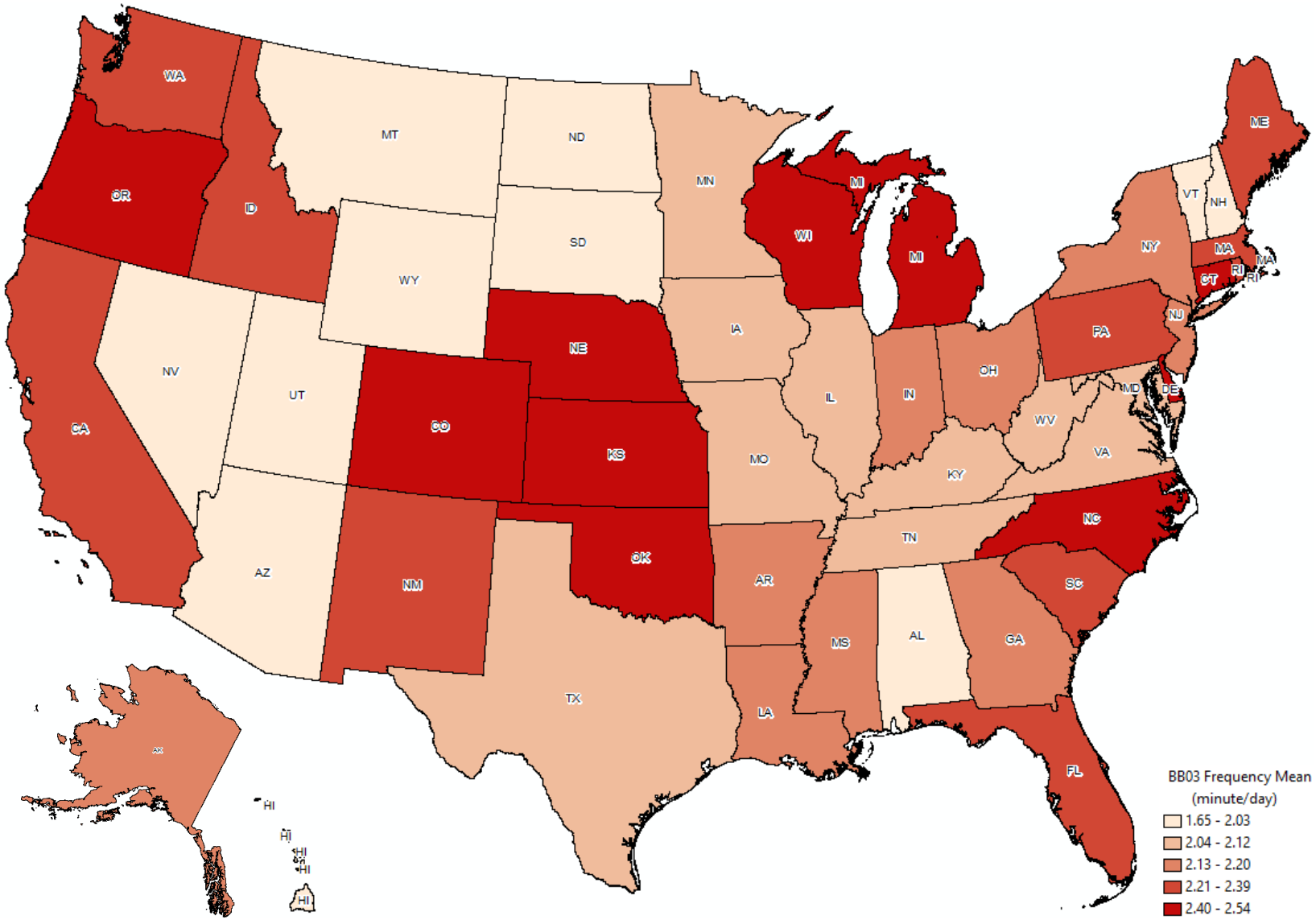


Figure A-20. BB03 Frequency by Quantiles

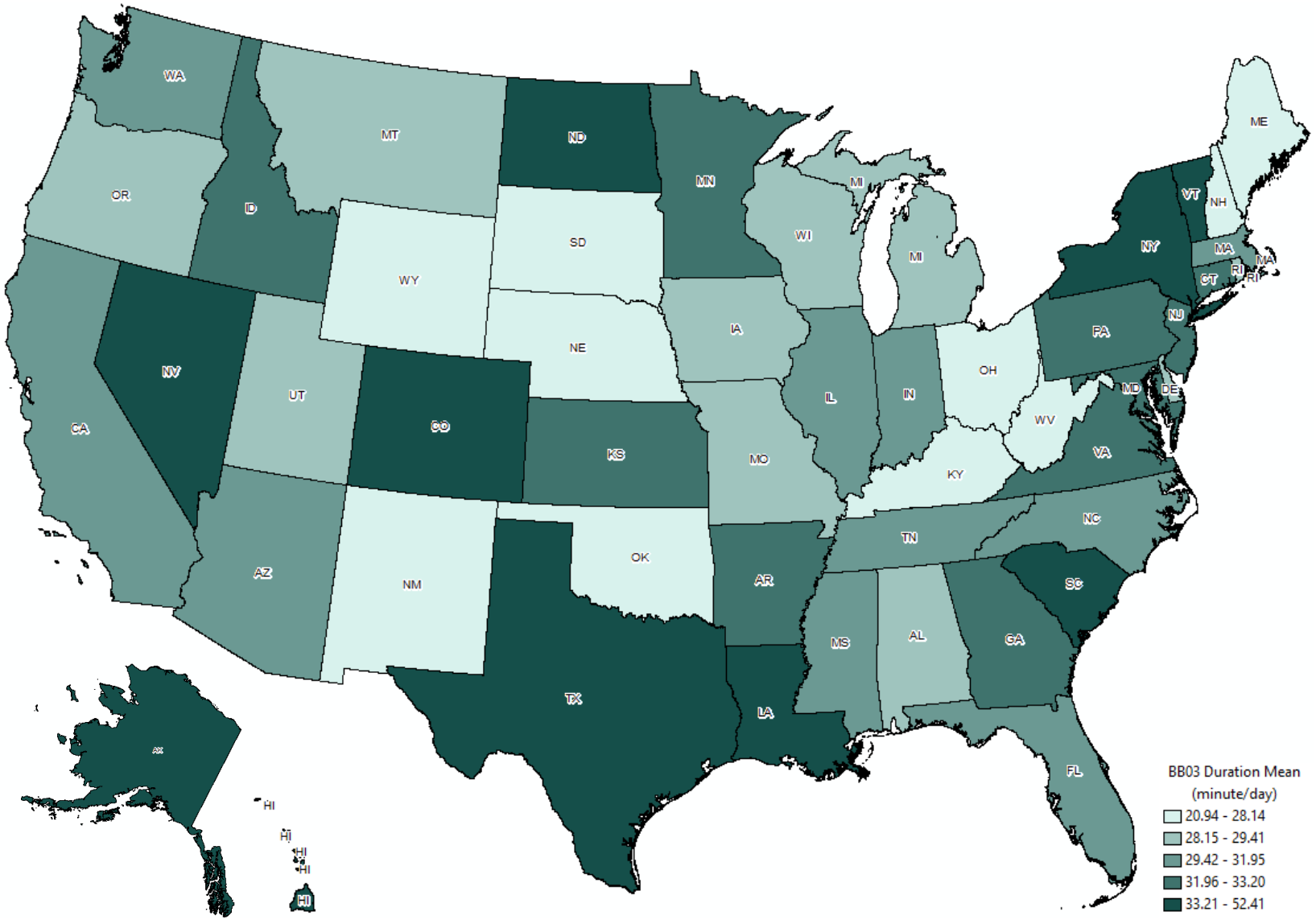


Figure A-21. BB03 Duration by Quantiles

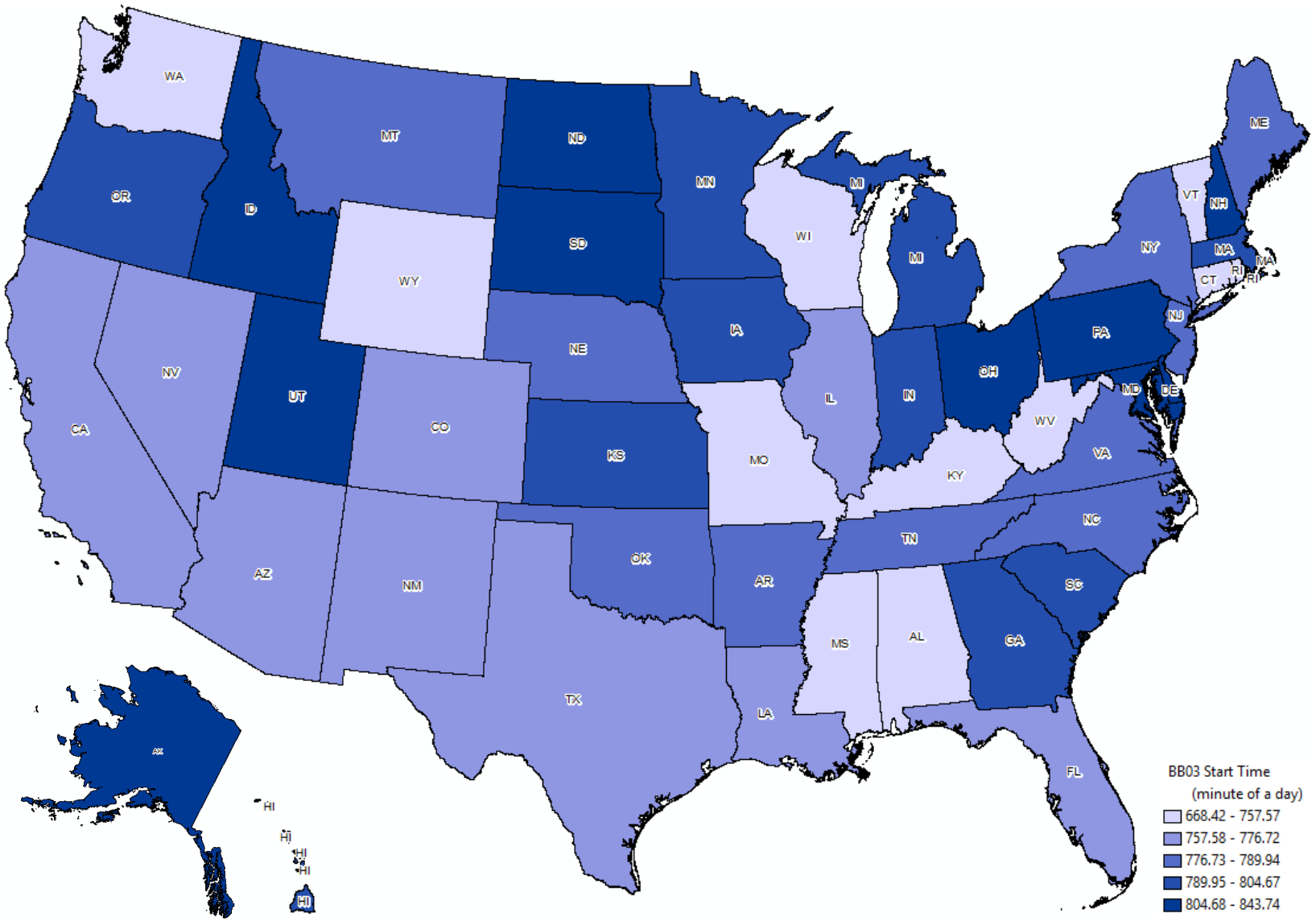


Figure A-22. BB03 Start Time by Quantiles

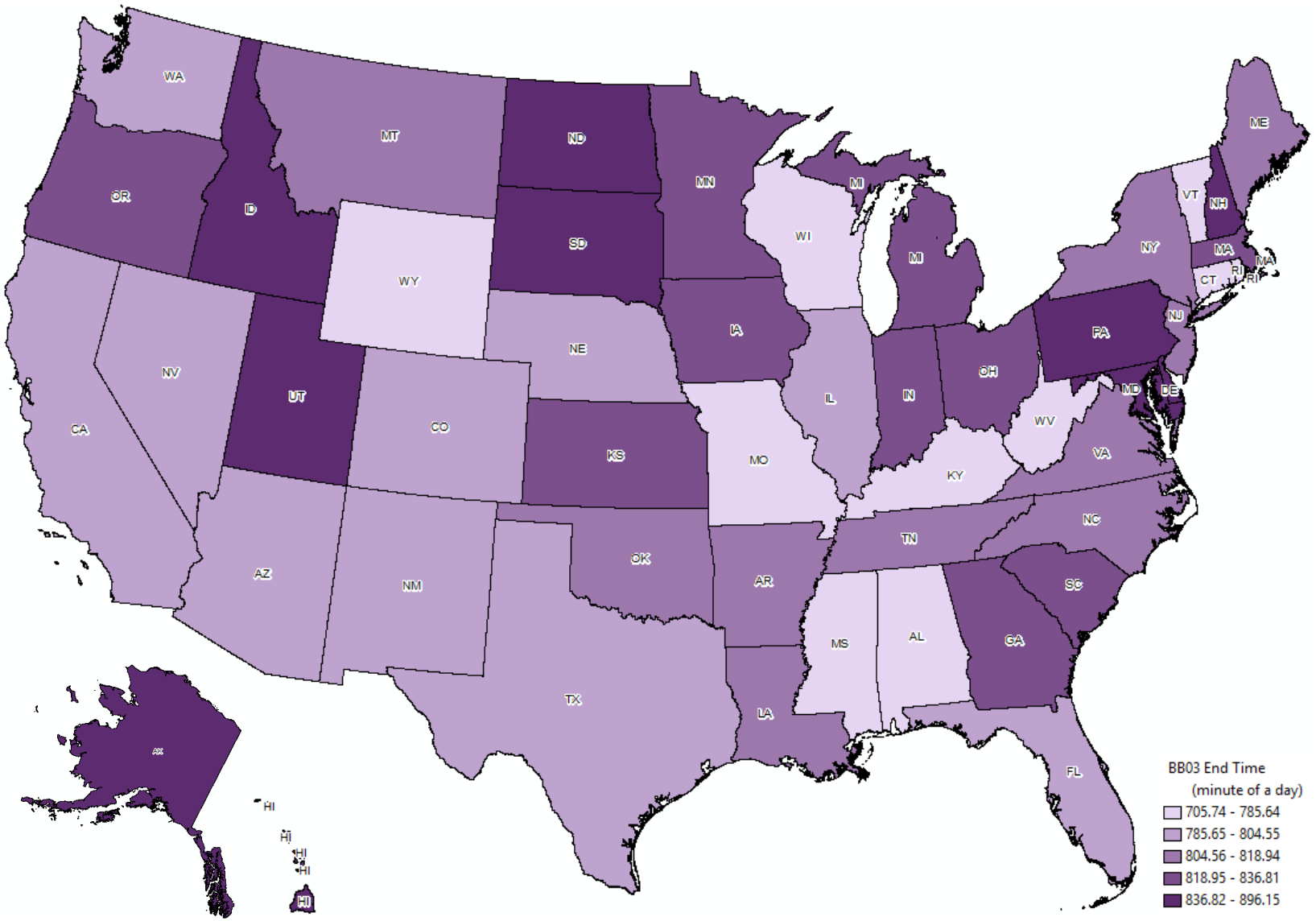


Figure A-23. BB03 End Time by Quantiles

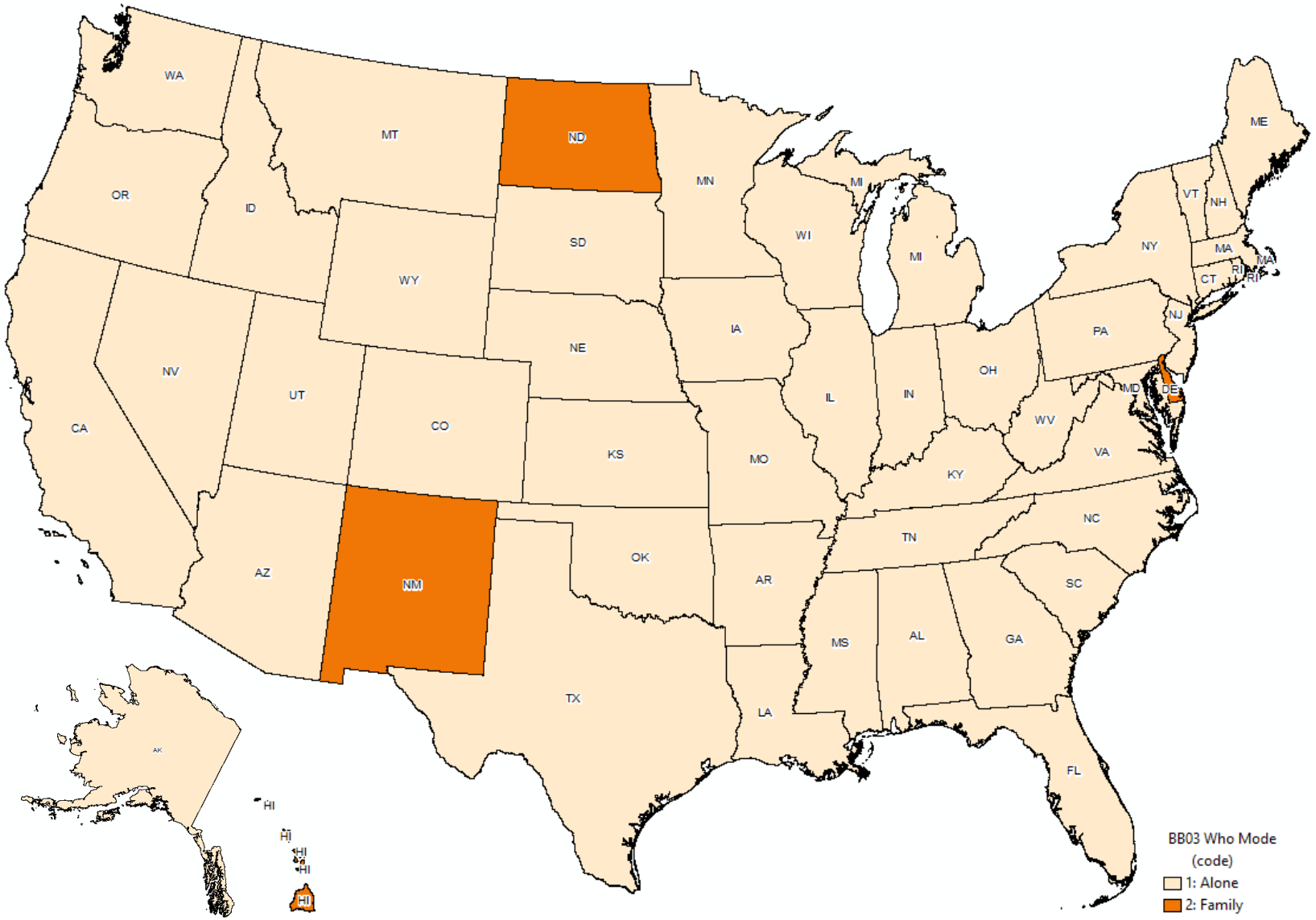


Figure A-24. BB03 Partner

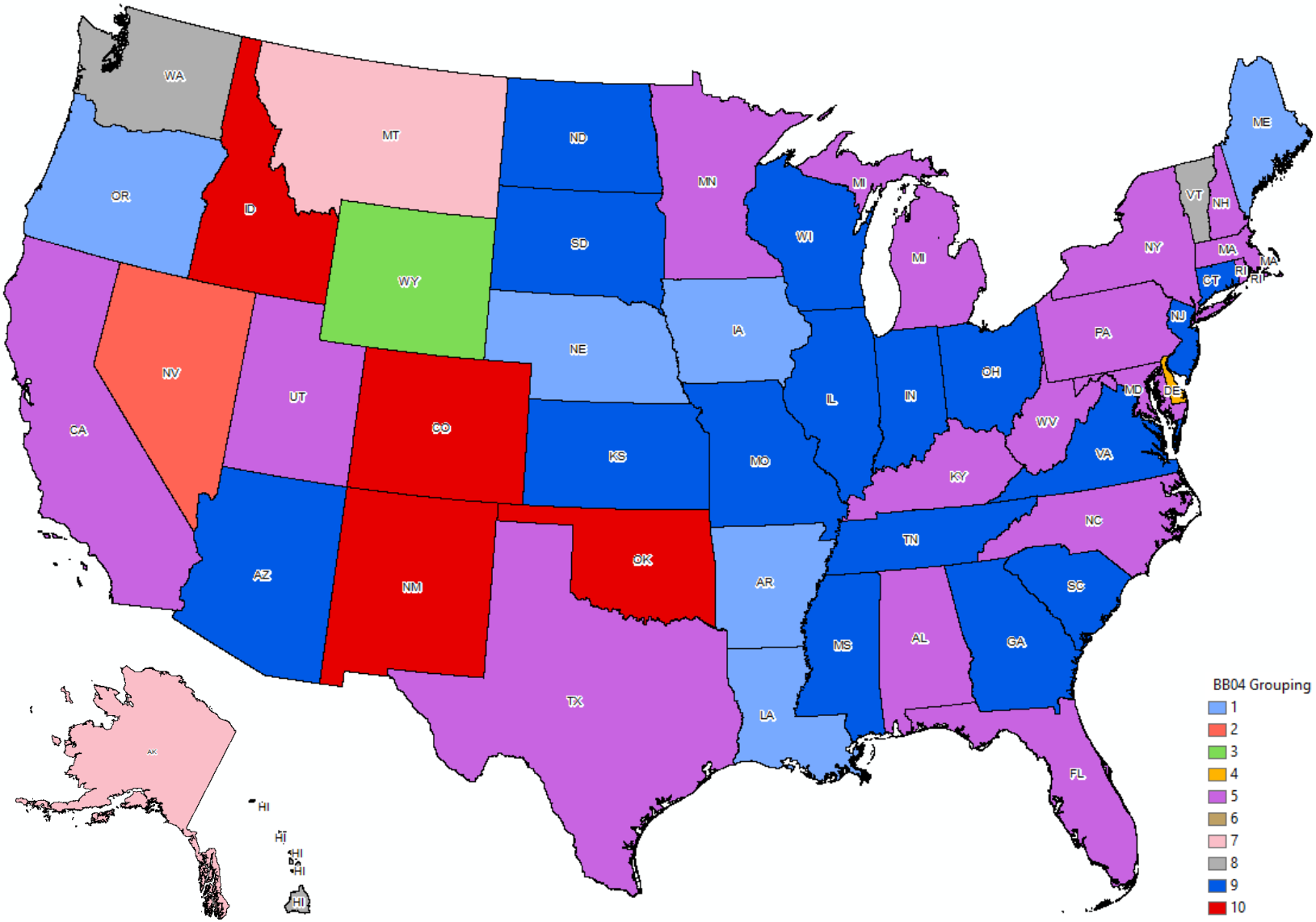


Figure A-25. BB04 State Clusters by Grouping Analysis

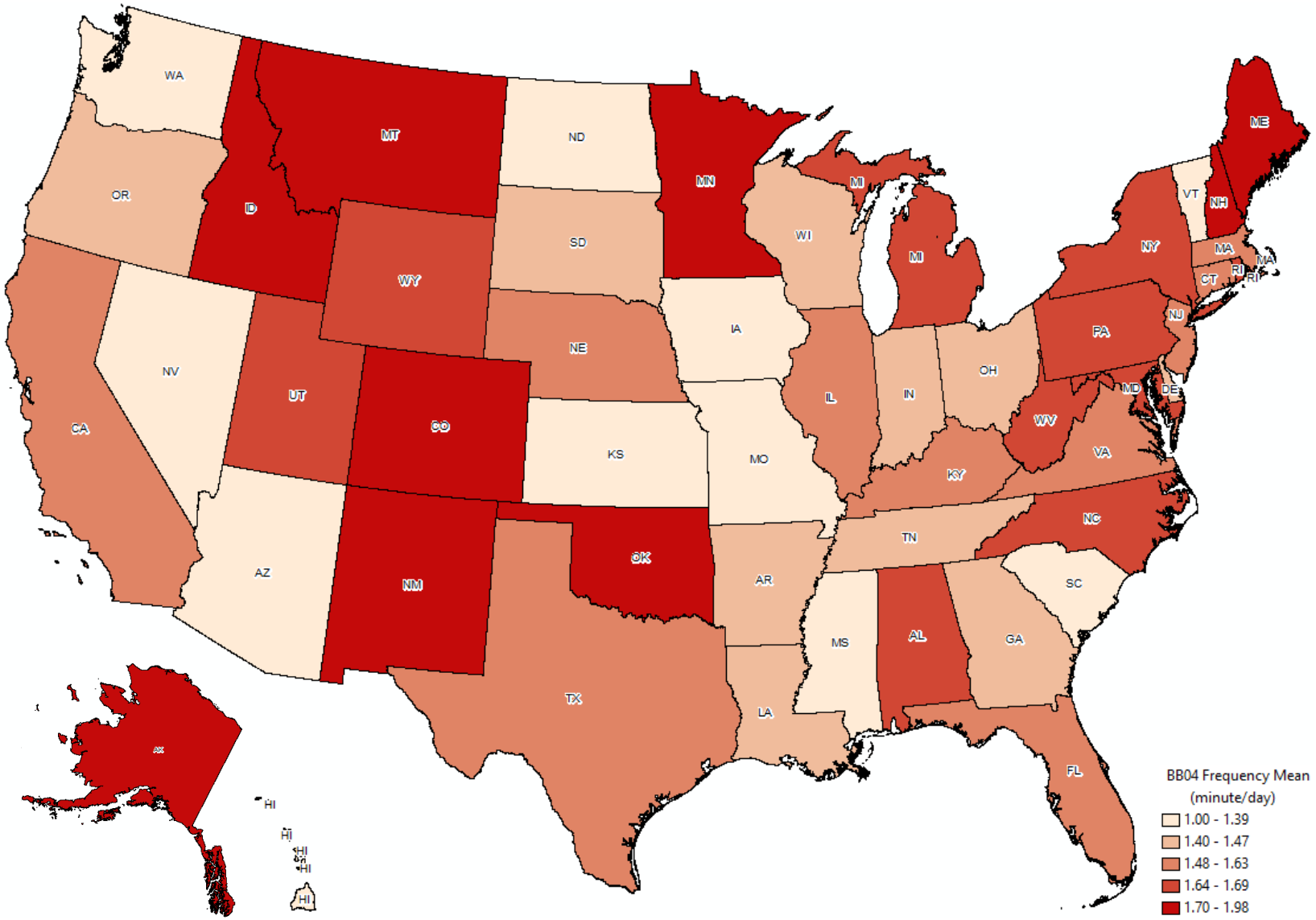


Figure A-26. BB04 Frequency by Quantiles

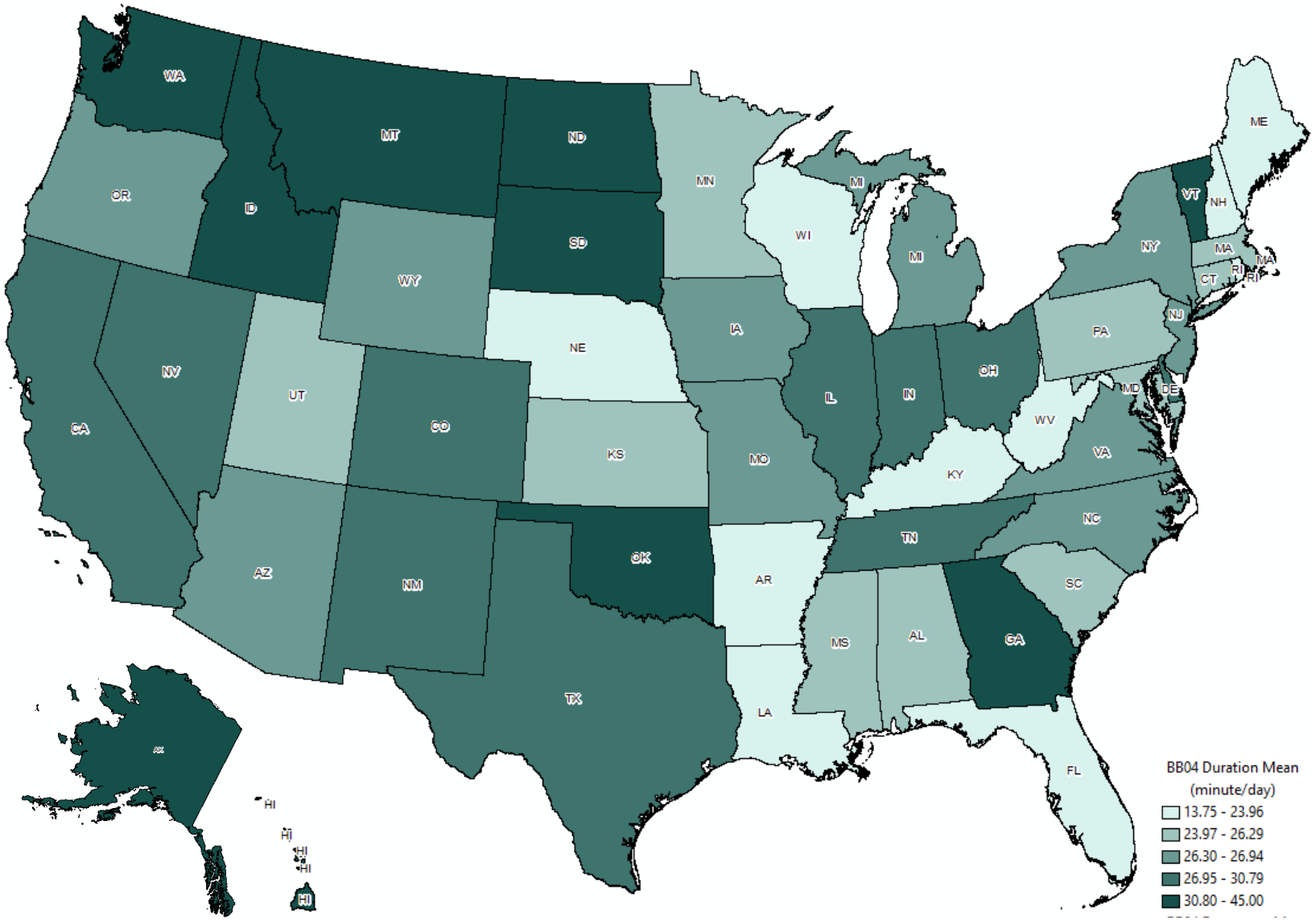


Figure A-27. BB04 Duration by Quantiles

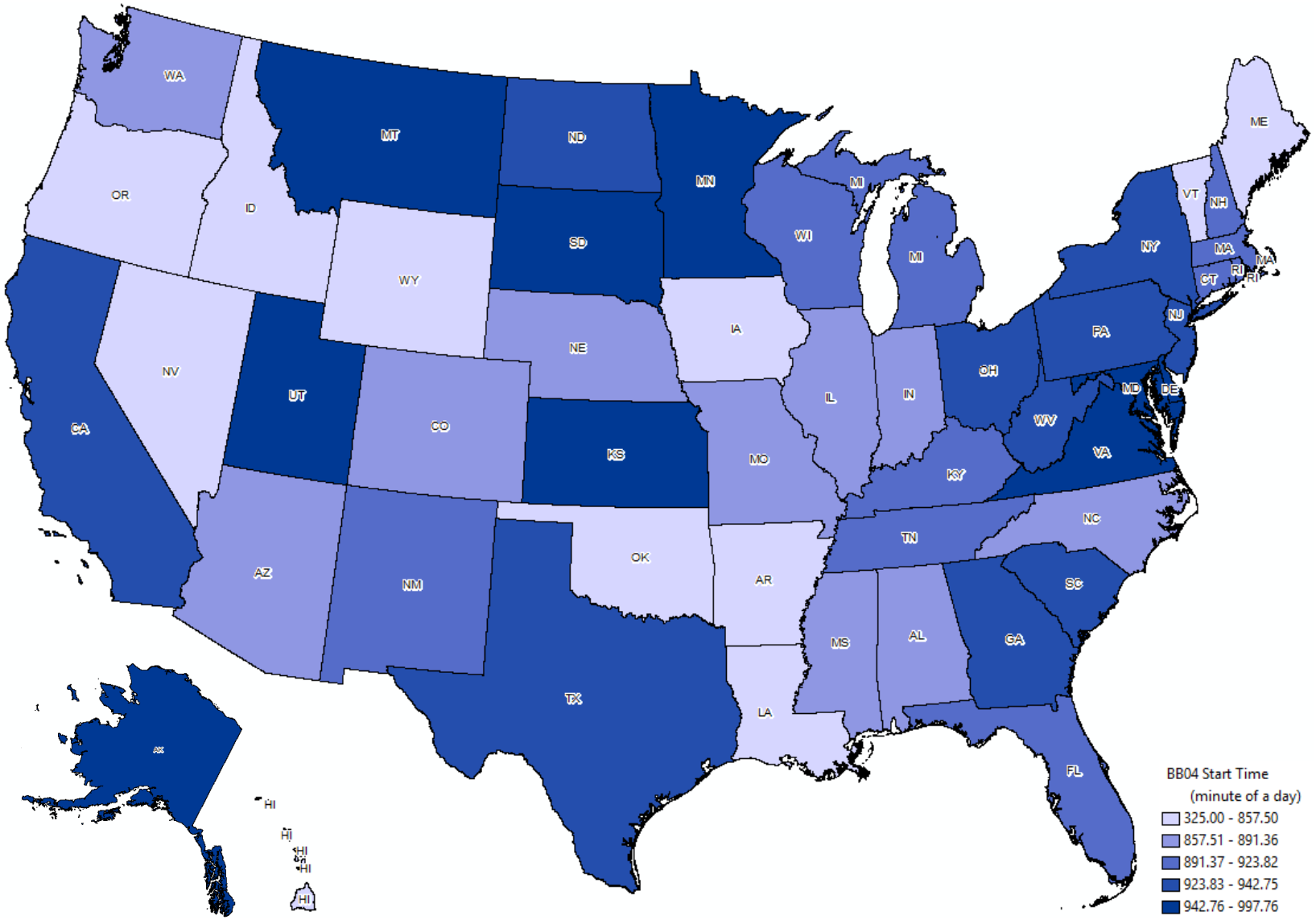


Figure A-28. BB04 Start Time by Quantiles

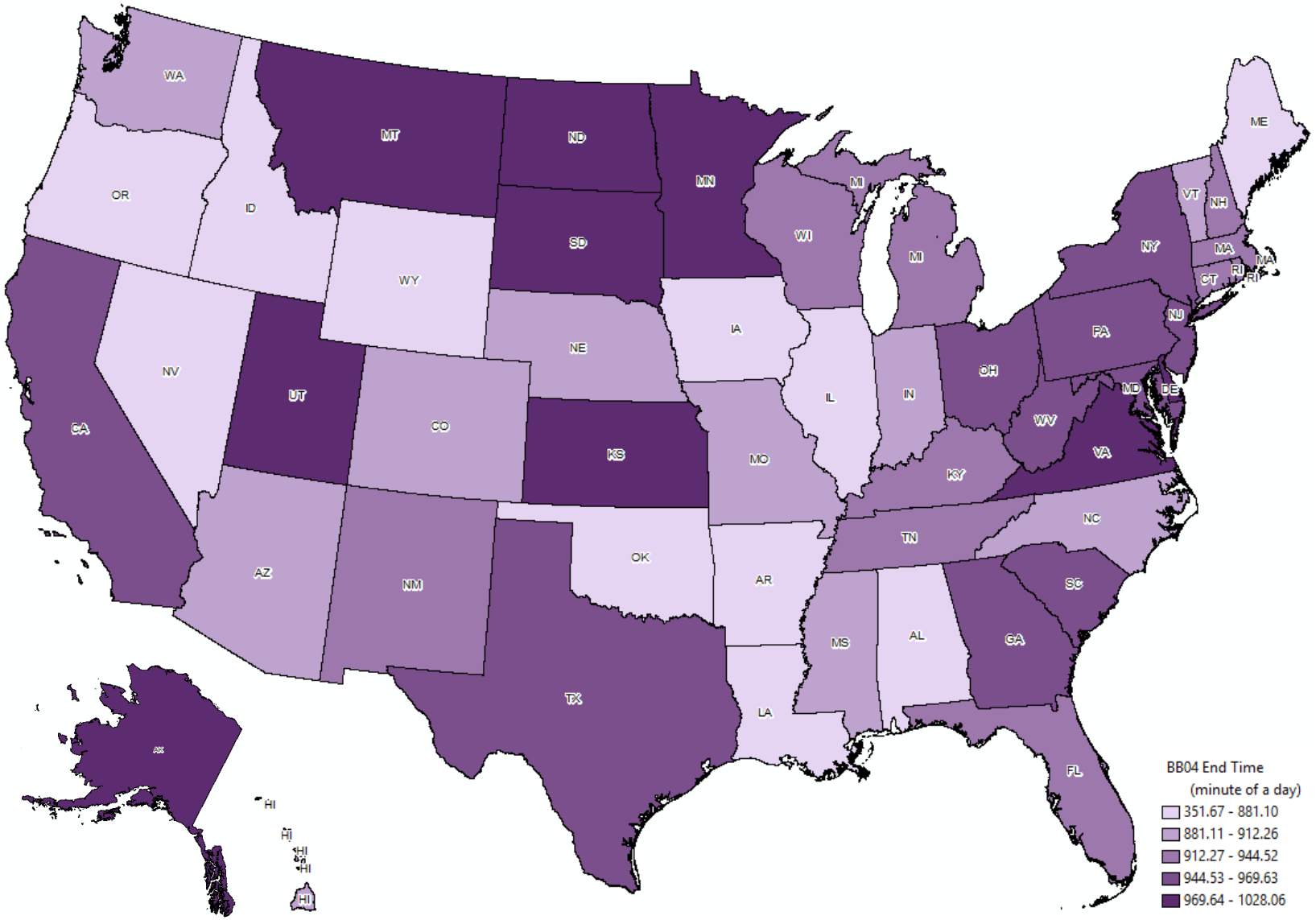


Figure A-29. BB04 End Time by Quantiles

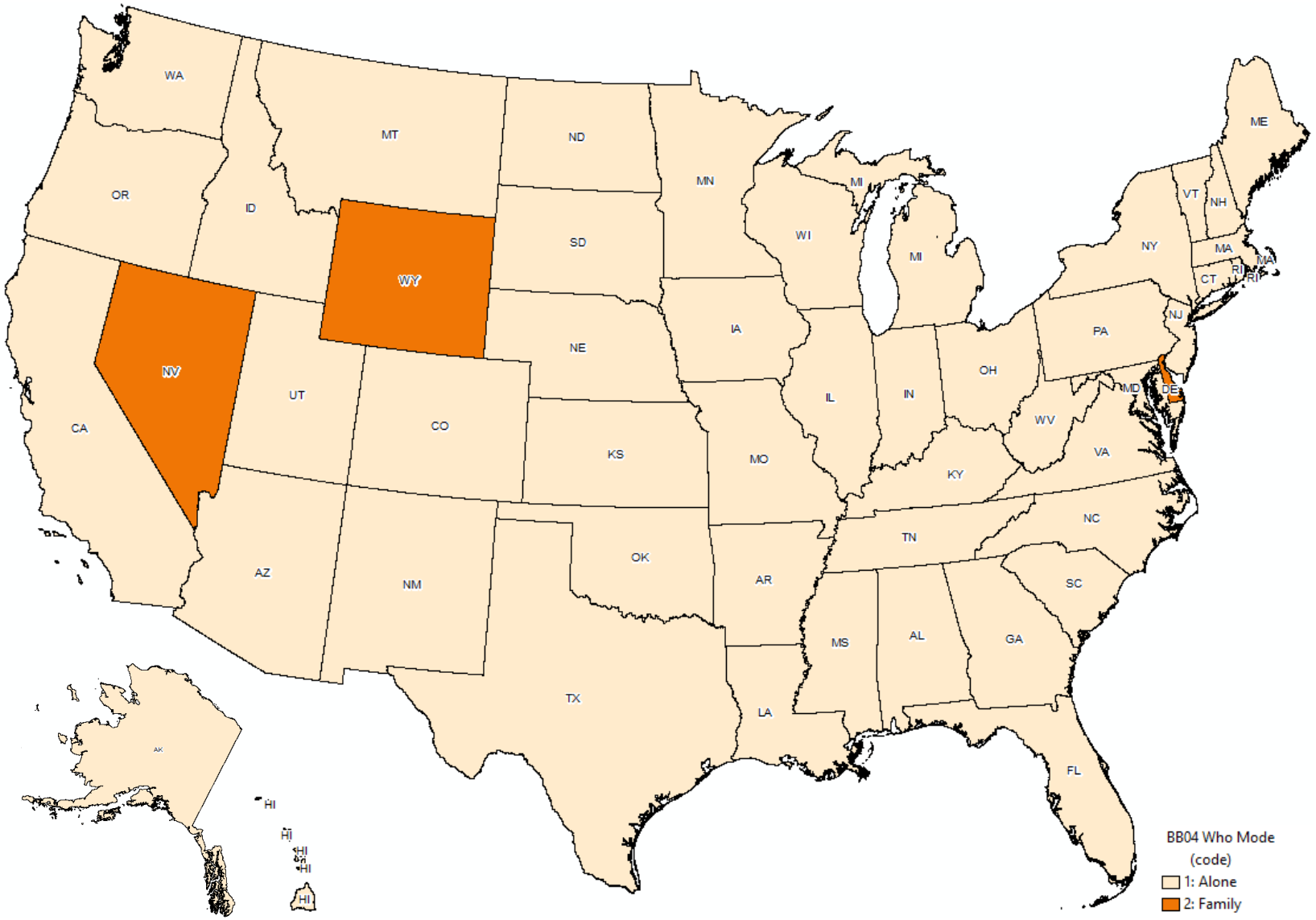


Figure A-30. BB04 Partner

APPENDIX B. Descriptive Analysis for Activities: Full Tables

Table A-1. Mean and CV of Activities by Cluster

CL	Code	Mean (Mode)						CV					
		Freq	Dur/a	Dur/d	Start	End	Partner	Freq	Dur/a	Dur/d	Start	End	Partner
1	AA01	1.88	29.90	51.51	662.09	687.09	-1	0.50	0.60	0.64	0.41	0.39	0.00
	BB01	1.13	41.71	47.57	911.07	952.78	1	0.32	0.95	1.04	0.34	0.33	0.39
	BB02	1.30	36.46	43.48	958.78	982.47	1	0.54	1.07	1.08	0.31	0.32	0.37
	BB03	1.49	27.35	37.77	819.86	847.22	1	0.53	0.77	0.80	0.36	0.36	0.39
	BB04	1.16	22.86	26.06	1019.70	1042.57	1	0.35	0.63	0.74	0.25	0.24	0.38
	BB05	1.35	13.45	19.24	858.79	872.25	1	0.45	0.58	0.83	0.40	0.39	0.44
	BB06	1.12	68.24	75.71	950.53	1018.77	1	0.36	0.94	0.92	0.24	0.24	0.38
	BB07	1.60	11.96	18.19	724.05	736.01	1	0.62	0.96	1.20	0.44	0.43	0.37
	BB08	1.11	67.53	78.05	821.42	888.95	1	0.29	0.91	1.02	0.39	0.40	0.44
	CD01	2.17	21.13	44.31	883.96	905.08	2	0.66	0.84	1.13	0.36	0.35	0.10
	CD02	1.13	23.00	27.29	684.52	707.52	2	0.30	1.23	1.55	0.58	0.57	0.18
	EF01	2.76	119.99	222.08	819.13	928.85	1	0.48	0.91	0.93	0.35	0.29	0.54
	LL01	1.52	93.41	128.16	1112.22	1171.92	2	0.50	0.66	0.66	0.18	0.22	0.35
	LL02	1.26	61.05	75.66	1072.90	1092.81	1	0.65	1.04	1.07	0.25	0.30	0.40
	LL03	1.19	54.00	71.57	1010.24	1027.30	1	0.40	1.02	1.23	0.30	0.33	0.40
2	AA01	1.62	32.83	50.13	812.98	830.36	-1	0.52	0.68	0.79	0.37	0.37	0.00
	BB01	1.23	65.17	76.46	749.46	808.13	1	0.43	0.92	0.93	0.32	0.30	0.38
	BB02	1.56	55.76	70.00	796.20	843.05	1	0.73	0.89	0.80	0.27	0.26	0.37
	BB03	1.65	34.32	52.70	812.72	844.57	1	0.53	0.94	0.92	0.27	0.27	0.41
	BB04	1.33	29.78	38.18	917.94	939.22	1	0.45	1.33	1.20	0.27	0.27	0.41
	BB05	1.13	46.31	48.19	750.13	796.44	1	0.30	1.99	1.91	0.42	0.40	0.45
	BB06	1.13	73.46	82.12	785.21	858.67	1	0.33	0.74	0.75	0.25	0.22	0.40
	BB07	1.70	15.01	23.12	816.13	824.68	1	0.61	1.11	1.16	0.37	0.37	0.34
	BB08	1.17	88.27	99.05	809.43	897.70	1	0.32	0.89	0.89	0.27	0.23	0.45
	CD01	3.05	24.88	65.85	910.43	935.31	2	0.95	0.95	0.96	0.31	0.30	0.12
	CD02	1.95	31.93	71.08	748.25	780.18	2	1.19	1.39	1.48	0.40	0.39	0.13
	EF01	1.52	86.12	123.52	867.91	893.67	1	0.60	0.95	1.13	0.31	0.34	0.40
	LL01	2.45	162.92	338.39	981.30	1011.05	1	0.47	0.70	0.53	0.17	0.30	0.40
	LL02	1.15	73.23	83.76	999.71	854.77	1	0.41	0.92	0.97	0.35	0.55	0.37
	LL03	1.33	81.54	119.81	901.68	900.09	1	0.47	0.94	0.97	0.32	0.37	0.41
3	AA01	1.82	32.24	53.33	731.64	738.04	-1	0.49	0.61	0.59	0.38	0.39	0.00
	BB01	1.08	59.92	64.11	752.67	806.92	1	0.28	0.85	0.84	0.39	0.36	0.37
	BB02	1.36	39.59	48.99	820.63	860.23	1	0.51	0.93	0.96	0.38	0.36	0.33
	BB03	1.49	27.74	37.38	792.07	802.72	1	0.55	0.97	0.91	0.37	0.38	0.41
	BB04	1.32	25.09	30.93	952.59	966.02	1	0.52	0.64	0.77	0.31	0.33	0.44
	BB05	1.00	42.14	42.14	894.71	936.86	1	0.00	1.04	1.04	0.40	0.39	0.55
	BB06	1.05	89.86	93.65	783.27	873.14	1	0.22	0.81	0.80	0.32	0.28	0.35
	BB07	1.77	14.44	24.05	768.90	774.23	1	0.89	1.10	1.11	0.40	0.40	0.23
	BB08	1.20	91.40	102.07	811.87	903.27	1	0.47	1.42	1.32	0.26	0.21	0.46
	CD01	2.06	26.04	47.72	794.07	799.39	2	0.85	1.08	1.11	0.46	0.46	0.10
	CD02	1.00	20.00	20.00	447.50	467.50	2	0.00	0.35	0.35	0.04	0.05	0.00
	EF01	2.81	133.73	264.40	925.90	955.77	1	0.49	0.68	0.87	0.28	0.33	0.47
	LL01	1.48	81.35	110.13	1020.23	819.75	1	0.52	0.65	0.67	0.36	0.61	0.41
	LL02	1.20	63.06	92.67	814.49	685.54	1	0.47	0.75	1.09	0.55	0.70	0.54
	LL03	1.28	60.11	83.12	974.37	853.55	1	0.54	0.90	1.17	0.37	0.52	0.44
4	AA01	1.75	34.01	55.12	842.77	862.88	-1	0.52	0.67	0.72	0.36	0.35	0.00
	BB01	1.34	108.83	136.88	718.53	825.17	1	0.46	0.81	0.81	0.25	0.21	0.40
	BB02	1.33	57.48	68.11	749.33	799.12	1	0.54	1.18	1.08	0.31	0.29	0.40
	BB03	1.89	35.69	61.99	757.83	791.41	1	0.62	1.07	0.95	0.29	0.28	0.41
	BB04	1.40	29.91	41.30	899.46	917.23	1	0.49	0.68	0.91	0.33	0.34	0.40

Table A-1 (cont'd)

	BB05	1.25	46.88	67.50	781.56	828.44	1	0.37	0.93	1.10	0.56	0.54	0.31
	BB06	1.18	84.85	95.63	749.34	834.20	1	0.38	0.93	0.90	0.32	0.26	0.39
	BB07	1.89	18.82	34.31	854.51	873.33	1	0.65	0.96	1.27	0.38	0.37	0.43
	BB08	1.21	118.75	138.93	707.68	826.43	1	0.35	0.84	0.91	0.29	0.23	0.51
	CD01	3.20	25.48	73.26	891.63	911.43	2	0.96	0.65	0.97	0.33	0.32	0.11
	CD02	1.71	32.49	68.93	927.99	939.91	2	0.66	0.97	1.57	0.27	0.29	0.17
	EF01	1.39	92.54	127.45	827.31	896.62	1	0.54	0.92	1.10	0.35	0.32	0.43
	LL01	1.41	74.70	100.87	1007.47	940.17	2	0.46	0.64	0.72	0.32	0.45	0.38
	LL02	1.33	58.93	73.74	915.51	974.44	1	0.59	0.81	0.78	0.36	0.34	0.41
	LL03	1.44	70.00	123.37	902.21	922.01	1	0.54	0.79	1.06	0.35	0.36	0.43
5	AA01	1.73	32.04	51.98	812.50	830.94	-1	0.52	0.97	0.91	0.36	0.35	0.00
	BB01	1.40	107.43	139.71	618.59	725.33	1	0.52	0.77	0.77	0.31	0.24	0.38
	BB02	1.47	52.86	68.02	765.39	815.66	1	0.56	1.09	0.98	0.30	0.28	0.35
	BB03	1.98	41.12	73.35	797.88	839.00	1	0.57	0.91	0.82	0.25	0.25	0.39
	BB04	1.40	29.91	40.48	904.45	934.36	1	0.46	0.73	0.79	0.27	0.27	0.38
	BB05	1.33	52.60	68.60	789.11	841.71	1	0.46	1.30	1.21	0.37	0.37	0.35
	BB06	1.22	62.11	73.53	769.94	832.06	1	0.44	0.83	0.86	0.31	0.28	0.39
	BB07	1.82	22.02	39.04	801.84	823.85	1	0.79	1.11	1.37	0.38	0.37	0.33
	BB08	1.13	69.13	78.26	780.54	849.67	1	0.30	0.92	1.01	0.26	0.24	0.42
	CD01	2.98	24.49	66.56	890.97	908.13	2	0.77	0.66	0.88	0.30	0.30	0.12
	CD02	2.14	31.03	66.64	677.58	708.61	2	0.71	1.44	1.30	0.37	0.38	0.24
	EF01	1.46	70.16	107.90	889.85	938.00	1	0.55	0.84	1.25	0.31	0.30	0.38
	LL01	1.67	89.89	138.57	1070.05	1141.48	2	0.50	0.58	0.61	0.17	0.19	0.36
	LL02	1.20	70.11	78.50	956.43	930.54	1	0.40	0.66	0.58	0.33	0.43	0.41
	LL03	1.31	62.44	97.28	895.93	938.05	1	0.45	0.87	1.04	0.31	0.32	0.40
6	AA01	1.71	30.59	48.06	851.31	870.00	-1	0.53	0.57	0.59	0.35	0.35	0.00
	BB01	1.30	68.67	81.96	747.17	810.50	1	0.40	0.83	0.79	0.32	0.29	0.38
	BB02	1.51	57.69	75.01	771.77	829.46	1	0.55	1.36	1.26	0.32	0.28	0.38
	BB03	1.82	36.63	62.14	812.24	848.86	1	0.56	0.75	0.86	0.26	0.26	0.37
	BB04	1.31	27.22	35.51	904.29	926.81	1	0.44	0.58	0.69	0.30	0.29	0.36
	BB05	1.25	45.75	47.00	856.75	902.50	2	0.40	1.36	1.31	0.37	0.29	0.38
	BB06	1.46	163.20	219.04	682.38	845.58	1	0.50	0.65	0.64	0.22	0.16	0.40
	BB07	1.62	20.66	31.79	787.73	766.04	1	0.54	1.06	1.04	0.43	0.45	0.37
	BB08	1.35	81.83	110.52	777.91	859.74	1	0.42	0.90	1.20	0.23	0.21	0.36
	CD01	2.89	29.15	65.02	907.03	936.18	2	0.95	0.97	0.91	0.32	0.30	0.09
	CD02	1.25	19.25	26.75	606.00	625.25	2	0.40	0.70	0.93	0.48	0.48	0.00
	EF01	1.59	86.62	146.83	896.04	878.30	1	0.56	1.07	1.37	0.29	0.35	0.34
	LL01	1.72	108.02	167.63	1107.38	1162.35	2	0.50	0.80	0.65	0.16	0.22	0.33
	LL02	1.18	75.82	89.64	1032.73	1108.55	1	0.51	0.53	0.68	0.25	0.22	0.34
	LL03	1.30	61.77	88.06	921.62	931.39	1	0.45	0.82	0.90	0.31	0.35	0.37
7	AA01	1.62	30.25	45.60	777.14	805.55	-1	0.53	0.65	0.68	0.39	0.37	0.00
	BB01	1.21	59.53	68.61	687.33	746.86	1	0.40	0.92	0.92	0.34	0.31	0.39
	BB02	1.44	54.23	66.47	757.27	811.49	1	0.54	0.92	0.86	0.31	0.29	0.34
	BB03	1.81	31.92	55.06	751.51	783.43	1	0.53	0.94	0.89	0.29	0.28	0.40
	BB04	1.31	27.50	35.01	856.95	884.46	1	0.44	0.86	0.92	0.29	0.28	0.39
	BB05	1.15	103.19	106.46	543.12	646.31	1	0.33	1.46	1.40	0.37	0.44	0.33
	BB06	1.12	62.74	69.94	740.51	803.25	1	0.31	0.75	0.78	0.27	0.25	0.40
	BB07	1.69	18.10	28.43	710.68	728.78	1	0.67	1.05	1.18	0.39	0.39	0.31
	BB08	1.16	82.84	93.84	778.12	860.96	1	0.32	0.86	0.96	0.26	0.24	0.53
	CD01	2.48	23.98	53.53	888.62	912.60	2	0.89	0.73	0.89	0.32	0.31	0.15
	CD02	1.58	18.91	29.85	698.83	717.74	2	0.84	1.11	1.23	0.42	0.40	0.10
	EF01	1.29	72.05	88.90	829.41	878.96	1	0.40	0.92	1.02	0.32	0.33	0.42
	LL01	2.70	156.38	362.63	839.20	977.21	1	0.50	0.76	0.60	0.19	0.18	0.39
	LL02	1.17	99.76	115.06	832.47	872.23	1	0.32	0.87	0.87	0.46	0.44	0.41
	LL03	1.29	78.79	112.29	815.67	886.70	1	0.52	0.79	0.94	0.36	0.34	0.40

Table A-1 (cont'd)

8	AA01	1.90	37.83	64.28	721.85	756.88	-1	0.53	0.63	0.60	0.34	0.31	0.00
	BB01	1.21	50.52	57.13	723.70	766.40	1	0.41	1.24	1.13	0.38	0.36	0.41
	BB02	1.49	44.13	56.60	767.04	809.62	1	0.56	0.94	0.87	0.32	0.31	0.36
	BB03	1.72	33.14	52.65	757.66	790.80	1	0.57	1.00	0.93	0.31	0.30	0.41
	BB04	1.34	27.38	35.08	881.29	901.36	1	0.45	0.72	0.79	0.31	0.31	0.41
	BB05	2.25	15.67	35.25	1026.94	1042.61	1	0.67	0.76	0.74	0.28	0.27	0.00
	BB06	1.18	64.97	71.73	797.45	862.42	1	0.45	0.75	0.73	0.29	0.27	0.43
	BB07	1.79	16.84	27.75	793.66	810.49	1	0.77	1.46	1.47	0.41	0.40	0.38
	BB08	1.21	75.18	83.57	789.57	864.75	1	0.35	0.92	0.82	0.30	0.29	0.45
	CD01	2.51	23.46	54.76	891.21	914.67	2	0.67	0.61	0.91	0.32	0.31	0.12
	CD02	1.31	26.19	32.19	812.98	839.17	2	0.60	1.11	0.97	0.42	0.42	0.24
	EF01	1.55	71.54	115.44	899.44	905.53	1	0.63	0.84	1.14	0.37	0.40	0.44
	LL01	1.73	109.51	168.79	1066.48	1124.03	1	0.52	0.78	0.69	0.18	0.25	0.36
	LL02	1.36	79.08	117.64	893.50	972.58	1	0.48	0.81	1.08	0.40	0.37	0.43
	LL03	1.23	60.07	78.71	920.12	889.11	1	0.43	1.06	1.03	0.35	0.42	0.40
9	AA01	1.71	33.91	52.88	819.67	845.85	-1	0.54	0.73	0.73	0.36	0.35	0.00
	BB01	1.17	64.04	73.22	775.45	783.14	1	0.41	0.89	0.91	0.35	0.37	0.38
	BB02	1.31	53.53	63.65	830.51	867.30	1	0.50	0.92	0.96	0.33	0.33	0.39
	BB03	1.62	34.74	52.58	852.22	886.96	1	0.56	0.94	0.93	0.26	0.25	0.41
	BB04	1.28	34.33	41.47	909.04	928.53	1	0.42	1.04	0.91	0.32	0.32	0.37
	BB05	1.33	24.17	36.67	692.00	716.17	1	0.43	0.70	0.97	0.72	0.72	0.43
	BB06	1.11	80.91	88.16	685.50	766.41	1	0.35	0.70	0.75	0.29	0.27	0.44
	BB07	1.55	22.65	30.91	680.78	703.43	1	0.76	1.64	1.36	0.41	0.40	0.38
	BB08	1.38	63.44	91.25	780.75	844.19	1	0.38	0.64	0.83	0.23	0.17	0.54
	CD01	2.86	32.26	78.90	824.91	857.17	2	0.85	0.91	0.97	0.36	0.34	0.18
	CD02	1.00	34.60	34.60	518.00	552.60	2	0.00	1.27	1.27	0.57	0.55	0.00
	EF01	1.67	100.74	132.19	885.78	949.22	1	0.61	0.83	0.76	0.35	0.34	0.46
	LL01	1.61	88.39	132.87	946.89	955.86	1	0.60	0.74	0.75	0.28	0.33	0.40
	LL02	1.40	70.97	112.17	879.96	950.92	1	0.55	0.78	1.18	0.32	0.28	0.18
	LL03	1.47	94.96	147.07	928.30	918.05	1	0.61	0.81	0.92	0.31	0.37	0.35
10	AA01	1.76	32.55	53.87	809.71	829.12	-1	0.49	0.60	0.63	0.36	0.35	0.00
	BB01	1.21	58.48	68.49	777.70	828.56	1	0.40	0.91	0.88	0.36	0.34	0.42
	BB02	1.55	51.07	68.81	813.61	860.54	1	0.60	0.78	0.77	0.33	0.31	0.39
	BB03	1.64	32.21	49.40	785.47	817.69	1	0.50	0.78	0.83	0.30	0.29	0.39
	BB04	1.32	27.36	35.85	924.81	952.18	1	0.47	0.73	1.06	0.29	0.28	0.41
	BB05	1.00	5.00	5.00	1351.00	1356.00	2	NA	NA	NA	NA	NA	NA
	BB06	1.15	70.90	79.58	815.09	886.00	1	0.48	0.73	0.70	0.29	0.28	0.45
	BB07	1.67	18.83	35.08	820.59	803.14	1	0.74	1.27	1.97	0.36	0.38	0.39
	BB08	1.29	85.75	96.57	848.32	934.07	1	0.36	0.88	0.83	0.25	0.24	0.35
	CD01	2.62	25.82	64.01	931.35	957.17	2	0.77	0.68	0.94	0.28	0.27	0.06
	CD02	1.70	27.65	32.90	742.78	770.43	2	0.74	1.24	1.11	0.52	0.52	0.15
	EF01	1.44	69.37	108.30	903.48	932.85	1	0.58	0.91	1.14	0.33	0.34	0.36
	LL01	1.55	93.18	132.27	1092.85	1120.86	2	0.50	0.62	0.64	0.21	0.29	0.33
	LL02	1.29	69.88	96.07	965.32	983.77	1	0.36	0.64	0.74	0.33	0.35	0.35
	LL03	1.32	68.29	109.24	913.08	949.46	1	0.53	0.84	0.99	0.32	0.33	0.41

Table A-2. Mean and CV of Activities by Region

Rgn	Code	Mean (Mode)						CV					
		Freq	Dur/a	Dur/d	Start	End	Partner	Freq	Dur/a	Dur/d	Start	End	Partner
1	AA01	1.74	31.09	50.16	762.97	779.15	-1	0.52	0.61	0.63	0.39	0.38	0.00
	BB01	1.28	69.75	86.10	721.91	788.12	1	0.46	0.96	0.97	0.34	0.31	0.37
	BB02	1.43	52.17	64.36	798.69	847.65	1	0.59	1.15	1.00	0.31	0.30	0.33
	BB03	1.72	35.01	56.01	806.22	839.76	1	0.57	0.87	0.88	0.29	0.28	0.40
	BB04	1.34	26.55	34.04	938.78	963.83	1	0.45	0.80	0.86	0.27	0.27	0.38
	BB05	1.40	25.17	42.00	812.09	837.25	1	0.46	1.13	1.45	0.47	0.45	0.42
	BB06	1.24	98.45	122.33	759.48	857.92	1	0.41	0.82	0.85	0.27	0.22	0.36
	BB07	1.78	16.77	26.93	798.02	807.51	1	0.68	1.20	1.23	0.38	0.38	0.35
	BB08	1.04	56.17	56.17	754.46	810.63	1	0.20	1.13	1.13	0.35	0.34	0.34
	CD01	2.66	24.85	62.00	868.03	892.88	2	0.86	0.87	1.08	0.34	0.33	0.13
	CD02	1.54	24.10	52.50	729.59	743.40	2	0.70	1.02	1.55	0.43	0.43	0.16
	EF01	2.14	94.10	160.29	846.52	912.99	1	0.58	1.00	1.14	0.36	0.33	0.49
	LL01	1.94	115.34	209.09	1026.86	1062.05	1	0.58	0.78	0.80	0.22	0.29	0.39
	LL02	1.27	87.43	109.86	982.55	939.08	1	0.47	0.89	0.97	0.29	0.41	0.40
	LL03	1.29	71.23	100.22	891.45	906.00	1	0.47	0.98	1.01	0.35	0.39	0.40
2	AA01	1.74	31.94	50.63	732.07	756.79	-1	0.53	0.75	0.71	0.40	0.38	0.00
	BB01	1.25	76.72	92.24	735.04	799.74	1	0.43	0.97	0.94	0.35	0.31	0.40
	BB02	1.48	46.55	59.50	804.29	844.11	1	0.64	0.93	0.89	0.32	0.31	0.35
	BB03	1.70	31.46	49.87	803.18	833.25	1	0.55	1.01	0.98	0.30	0.30	0.40
	BB04	1.28	27.63	34.76	905.89	928.86	1	0.43	0.81	0.90	0.29	0.28	0.40
	BB05	1.33	62.74	70.58	747.64	810.37	1	0.57	1.85	1.63	0.48	0.44	0.35
	BB06	1.25	96.98	119.92	793.86	890.84	1	0.46	0.78	0.93	0.26	0.22	0.39
	BB07	1.71	15.55	26.13	744.87	760.42	1	0.69	1.29	1.54	0.43	0.43	0.36
	BB08	1.28	108.18	134.06	795.76	903.94	1	0.42	0.94	1.01	0.23	0.22	0.42
	CD01	2.52	24.09	53.16	894.20	915.49	2	0.85	0.83	0.87	0.34	0.33	0.11
	CD02	1.31	18.91	24.25	734.20	753.11	2	0.59	1.11	1.17	0.44	0.44	0.15
	EF01	2.19	104.71	180.76	871.51	933.24	1	0.59	0.91	1.04	0.32	0.32	0.47
	LL01	1.99	120.83	221.60	1008.25	1054.79	1	0.57	0.80	0.81	0.23	0.29	0.37
	LL02	1.31	70.54	98.45	863.88	853.13	1	0.56	0.74	1.08	0.44	0.48	0.39
	LL03	1.31	69.99	104.62	907.73	920.12	1	0.48	0.95	1.10	0.33	0.35	0.41
3	AA01	1.78	33.09	54.47	774.07	796.02	-1	0.51	0.68	0.71	0.37	0.37	0.00
	BB01	1.24	73.58	88.90	733.69	803.17	1	0.43	0.91	0.98	0.35	0.31	0.39
	BB02	1.42	52.72	65.44	807.68	855.72	1	0.56	1.04	1.00	0.33	0.31	0.39
	BB03	1.67	33.28	51.53	792.82	824.52	1	0.56	0.94	0.92	0.31	0.31	0.41
	BB04	1.31	26.94	34.57	928.72	950.03	1	0.45	0.76	0.89	0.28	0.28	0.40
	BB05	1.22	41.44	44.50	868.56	910.00	1	0.35	1.27	1.17	0.32	0.32	0.47
	BB06	1.20	103.93	121.24	765.92	869.85	1	0.43	0.95	0.94	0.29	0.25	0.41
	BB07	1.64	16.72	25.97	782.12	796.05	1	0.67	1.17	1.31	0.38	0.38	0.35
	BB08	1.23	79.95	92.39	800.01	879.96	1	0.35	0.91	0.86	0.28	0.25	0.48
	CD01	2.65	24.42	59.50	861.34	883.79	2	0.83	0.88	1.03	0.35	0.34	0.11
	CD02	1.83	31.86	57.47	701.49	733.35	2	1.07	1.37	1.56	0.47	0.45	0.14
	EF01	2.18	109.99	189.03	880.99	931.10	1	0.63	0.87	1.02	0.32	0.33	0.48
	LL01	2.01	123.36	230.85	1005.78	1047.22	1	0.56	0.82	0.81	0.22	0.30	0.38
	LL02	1.21	68.21	84.99	967.68	986.23	1	0.44	0.88	0.97	0.34	0.36	0.41
	LL03	1.29	71.37	101.53	931.78	932.77	1	0.49	0.84	0.99	0.33	0.37	0.41
4	AA01	1.74	31.11	50.70	761.52	785.51	-1	0.52	0.63	0.71	0.39	0.38	0.00
	BB01	1.25	72.40	87.81	730.23	800.29	1	0.45	0.98	0.99	0.35	0.32	0.38
	BB02	1.47	50.54	63.52	807.31	850.02	1	0.64	0.93	0.91	0.32	0.30	0.37
	BB03	1.71	31.85	51.98	784.08	814.31	1	0.56	0.81	0.90	0.32	0.31	0.38
	BB04	1.31	30.07	38.36	914.14	940.34	1	0.46	1.21	1.13	0.30	0.30	0.39
	BB05	1.10	57.74	59.29	718.88	776.62	1	0.27	1.62	1.57	0.41	0.42	0.36
	BB06	1.26	87.25	114.19	744.11	831.36	1	0.47	0.90	1.05	0.30	0.26	0.42

Table A-2 (cont'd)

	BB07	1.72	16.87	28.49	749.97	747.64	1	0.72	1.16	1.54	0.42	0.43	0.37
	BB08	1.16	76.30	87.67	797.03	873.33	1	0.32	0.80	0.93	0.28	0.24	0.42
	CD01	2.55	24.12	56.86	922.73	940.88	2	0.86	0.69	1.01	0.32	0.32	0.11
	CD02	1.53	28.39	44.66	784.17	812.55	2	0.80	1.07	1.40	0.41	0.42	0.21
	EF01	2.12	99.44	178.62	848.30	914.95	1	0.60	0.86	1.10	0.32	0.30	0.45
	LL01	1.90	116.50	207.95	1015.14	1071.83	2	0.56	0.72	0.80	0.22	0.27	0.36
	LL02	1.19	72.31	83.98	933.17	918.74	1	0.46	0.85	0.83	0.38	0.45	0.39
	LL03	1.30	66.10	102.89	924.03	941.82	1	0.52	0.94	1.11	0.33	0.36	0.39

Table A-3. Mean and CV of Activities by Day

Day	Code	Mean (Mode)						CV					
		Freq	Dur/a	Dur/d	Start	End	Partner	Freq	Dur/a	Dur/d	Start	End	Partner
WD	AA01	1.80	30.91	51.15	728.65	750.85	-1	0.51	0.70	0.70	0.40	0.39	0.00
	BB01	1.26	66.52	81.50	729.77	794.41	1	0.45	0.97	1.00	0.36	0.32	0.38
	BB02	1.41	45.23	56.12	824.48	860.96	1	0.57	1.04	0.98	0.33	0.32	0.35
	BB03	1.71	30.03	48.14	804.55	833.06	1	0.58	0.85	0.89	0.32	0.31	0.40
	BB04	1.30	26.75	33.87	937.48	962.35	1	0.45	0.78	0.84	0.28	0.27	0.39
	BB05	1.26	47.57	56.60	764.81	812.38	1	0.43	1.91	1.68	0.46	0.44	0.39
	BB06	1.22	87.94	109.68	797.75	885.69	1	0.44	0.91	1.02	0.28	0.24	0.38
	BB07	1.69	15.99	25.61	753.25	762.12	1	0.66	1.20	1.47	0.40	0.40	0.35
	BB08	1.20	79.27	95.55	774.71	853.98	1	0.36	1.00	1.11	0.29	0.27	0.39
	CD01	2.61	23.05	56.61	829.08	851.80	2	0.84	0.88	1.08	0.37	0.36	0.11
	CD02	1.54	23.70	43.25	666.13	686.18	2	1.07	1.00	1.71	0.48	0.47	0.16
	EF01	2.52	107.13	197.43	856.13	930.26	1	0.54	0.93	1.04	0.33	0.31	0.50
	LL01	1.93	110.28	200.81	1033.02	1073.30	1	0.57	0.79	0.85	0.22	0.28	0.38
	LL02	1.25	74.18	94.19	939.10	943.88	1	0.51	0.92	1.04	0.38	0.42	0.38
	LL03	1.32	65.19	97.11	936.91	951.24	1	0.49	0.90	1.02	0.32	0.35	0.40
WE	AA01	1.71	33.35	53.06	794.15	816.12	-1	0.52	0.65	0.70	0.36	0.36	0.00
	BB01	1.25	79.52	95.62	732.37	803.23	1	0.44	0.92	0.94	0.33	0.30	0.39
	BB02	1.48	55.49	69.89	787.57	840.42	1	0.63	1.00	0.93	0.31	0.29	0.38
	BB03	1.68	35.97	56.31	785.22	819.65	1	0.54	0.96	0.93	0.29	0.29	0.40
	BB04	1.31	28.84	37.06	903.68	925.45	1	0.44	1.03	1.06	0.29	0.29	0.41
	BB05	1.29	45.50	52.36	801.98	847.49	1	0.49	1.56	1.42	0.40	0.38	0.42
	BB06	1.24	105.76	128.19	738.64	844.39	1	0.45	0.85	0.88	0.28	0.24	0.41
	BB07	1.69	17.11	28.09	789.06	801.07	1	0.72	1.19	1.34	0.40	0.40	0.36
	BB08	1.21	85.46	97.60	805.53	890.99	1	0.36	0.91	0.93	0.27	0.23	0.45
	CD01	2.57	25.82	58.79	950.40	970.47	2	0.86	0.77	0.91	0.29	0.29	0.12
	CD02	1.67	30.52	50.77	801.11	831.63	2	0.77	1.39	1.46	0.39	0.38	0.17
	EF01	1.71	98.86	157.45	875.67	917.35	1	0.63	0.86	1.07	0.32	0.33	0.44
	LL01	2.01	129.81	239.92	990.39	1039.78	1	0.56	0.78	0.76	0.23	0.29	0.37
	LL02	1.23	73.99	91.55	940.60	910.96	1	0.45	0.78	0.89	0.34	0.42	0.41
	LL03	1.28	74.36	107.77	896.74	902.75	1	0.49	0.93	1.07	0.34	0.38	0.40

Table A-4. Mean and CV of Activities by Gender

Sex	Code	Mean (Mode)						CV					
		Freq	Dur/a	Dur/d	Start	End	Partner	Freq	Dur/a	Dur/d	Start	End	Partner
M	AA01	1.68	27.99	43.50	742.09	762.11	-1	0.52	0.62	0.66	0.41	0.39	0.00
	BB01	1.18	71.82	83.16	768.12	832.97	1	0.38	0.95	0.97	0.35	0.32	0.38
	BB02	1.31	50.98	58.79	831.62	879.36	1	0.61	0.92	0.90	0.33	0.31	0.33
	BB03	1.49	31.68	44.83	780.19	809.33	1	0.52	1.03	1.02	0.35	0.35	0.40
	BB04	1.24	27.42	32.83	926.24	953.66	1	0.41	0.84	0.87	0.30	0.29	0.38
	BB05	1.27	66.33	73.29	831.18	897.52	1	0.42	1.49	1.36	0.36	0.33	0.41
	BB06	1.26	112.15	138.56	755.71	867.86	1	0.46	0.82	0.86	0.28	0.23	0.39
	BB07	1.61	17.67	26.90	777.90	792.26	1	0.76	1.18	1.46	0.42	0.41	0.36
	BB08	1.24	90.43	107.93	785.88	876.31	1	0.37	0.90	0.96	0.26	0.23	0.45
	CD01	1.93	24.53	44.14	931.85	950.74	2	0.68	0.87	0.97	0.35	0.34	0.09
	CD02	1.50	33.93	63.02	723.11	751.28	2	0.69	1.33	1.51	0.43	0.43	0.16
	EF01	2.19	101.67	177.66	857.27	914.77	1	0.57	0.91	1.10	0.33	0.32	0.45
	LL01	2.00	128.62	238.84	999.04	1049.63	1	0.56	0.78	0.80	0.23	0.29	0.37
	LL02	1.26	80.61	104.82	931.17	955.06	1	0.49	0.86	1.00	0.36	0.38	0.40
	LL03	1.34	79.60	120.42	909.89	925.88	1	0.50	0.88	1.02	0.33	0.37	0.40
F	AA01	1.81	35.02	58.24	772.52	796.10	-1	0.51	0.68	0.69	0.37	0.36	0.00
	BB01	1.28	73.89	90.91	718.51	787.47	1	0.46	0.95	0.97	0.34	0.31	0.39
	BB02	1.48	50.51	64.58	797.91	842.25	1	0.60	1.05	0.97	0.32	0.30	0.38
	BB03	1.79	33.38	55.52	803.03	835.40	1	0.56	0.87	0.87	0.28	0.28	0.40
	BB04	1.33	27.78	36.04	920.81	943.13	1	0.45	0.93	0.97	0.28	0.28	0.40
	BB05	1.28	17.89	27.19	715.40	733.29	1	0.52	1.34	1.66	0.53	0.52	0.38
	BB06	1.20	79.53	96.48	779.80	859.33	1	0.42	0.94	1.02	0.29	0.25	0.42
	BB07	1.73	15.88	26.57	764.15	772.32	1	0.65	1.20	1.39	0.40	0.39	0.35
	BB08	1.09	58.78	61.06	814.49	873.27	1	0.26	1.01	0.98	0.32	0.30	0.41
	CD01	2.88	24.24	63.37	865.04	887.65	2	0.84	0.81	0.98	0.33	0.32	0.12
	CD02	1.66	23.56	38.94	734.80	758.37	2	1.00	1.13	1.54	0.44	0.44	0.17
	EF01	2.13	105.18	181.49	872.30	934.06	1	0.64	0.90	1.03	0.33	0.32	0.49
	LL01	1.95	113.08	205.37	1022.01	1062.16	1	0.57	0.80	0.80	0.22	0.28	0.38
	LL02	1.21	66.01	78.38	950.43	897.34	1	0.48	0.84	0.85	0.37	0.47	0.40
	LL03	1.27	62.35	88.91	922.43	928.33	1	0.49	0.93	1.04	0.33	0.37	0.40

BIBLIOGRAPHY

BIBLIOGRAPHY

Aarts, H., Paulussen, T., & Schaalma, H. (1997). Physical exercise habit: on the conceptualization and formation of habitual health behaviours. *Health Education Research, 12*(3), 363-374.

Abdi, H. (2010). Coefficient of variation. *Encyclopedia of research design, 1*, 169-171.

Ajzen, I. (1991). The Theory of Planned Behavior. Organizational Behavior and Decision Processes: University of Massachusetts at Amherst: Academic Press. Inc.

Aksanli, B., Akyurek, A. S., & Rosing, T. S. (2016). *User behavior modeling for estimating residential energy consumption.* Paper presented at the Smart City 360°.

APA. (2018). American Psychological Association (APA) Dictionary of Psychology. Retrieved from https://dictionary.apa.org/behavior

Bartusch, C., Odlare, M., Wallin, F., & Wester, L. (2012). Exploring variance in residential electricity consumption: Household features and building properties. *Applied Energy, 92*, 637-643.

Belsley, D. A., Kuh, E., & Welsch, R. E. (2005). *Regression diagnostics: Identifying influential data and sources of collinearity* (Vol. 571): John Wiley & Sons.

Bernard, M., McBride, J., Desmond, D., & Collings, N. (1988). *Events--The third variable in daily household energy consumption.* Paper presented at the Proceedings of the 1988 ACEEE Summer Study on Energy Efficiency in Buildings.

Bishop, C. M. (2006). *Pattern recognition and machine learning*: springer.

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

Caliński, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics-Theory and Methods, 3*(1), 1-27.

Chen, Yang, W., Yoshino, H., Levine, M. D., Newhouse, K., & Hinge, A. (2015). Definition of occupant behavior in residential buildings and its application to behavior analysis in case studies. *Energy and Buildings, 104*, 1-13.

Chen, C.-F., & Chao, W.-H. (2011). Habitual or reasoned? Using the theory of planned behavior, technology acceptance model, and habit to examine switching intentions toward public transit. *Transportation Research Part F: Traffic Psychology and Behaviour, 14*(2), 128-137.

Chen, J., Taylor, J. E., & Wei, H.-H. (2012). Modeling building occupant network energy consumption decision-making: The interplay between network structure and conservation. *Energy and Buildings, 47*, 515-524.

DaftLogic. (2018). List of the Power Consumption of Typical Household Appliances. Retrieved from https://www.daftlogic.com/information-appliance-power-consumption.htm

Danner, U. N., Aarts, H., & Vries, N. K. (2008). Habit vs. intention in the prediction of future behaviour: The role of frequency, context stability and mental accessibility of past behaviour. *British Journal of Social Psychology, 47*(2), 245-265.

Darby, S. (2006). The effectiveness of feedback on energy consumption. *A Review for DEFRA of the Literature on Metering, Billing and direct Displays, 486*(2006).

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*.

Dong, B., Cao, C., & Lee, S. E. (2005). Applying support vector machines to predict building energy consumption in tropical region. *Energy and Buildings, 37*(5), 545-553.

Dubrawski, A. (2015). *95792: Data Mining (Lecture Note)*. Retrieved from Pittsburgh, PA:

EIA. (2018). Residnetial Energy Consumption Survey (RECS). Retrieved from https://www.eia.gov/consumption/residential/about.php

ESRI. (2018). Data classification methods. *ArcGIS Pro.* Retrieved from http://pro.arcgis.com/en/pro-app/help/mapping/layer-properties/data-classification-methods.htm

Farid, D. M., Zhang, L., Rahman, C. M., Hossain, M. A., & Strachan, R. (2014). Hybrid decision tree and naïve Bayes classifiers for multi-class classification tasks. *Expert Systems with Applications, 41*(4), 1937-1946.

Fonseca, J. A., & Schlueter, A. (2015). Integrated model for characterization of spatiotemporal building energy consumption patterns in neighborhoods and city districts. *Applied Energy, 142*, 247-265.

Friedrichs, F., & Igel, C. (2005). Evolutionary tuning of multiple SVM parameters. *Neurocomputing, 64*, 107-117.

Generac. (2018). Estimating Power Needs: Portable Generators. Retrieved from https://www.lowes.com/projects/pdfs/portable-generator-wattage-chart.pdf

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research, 3*(Mar), 1157-1182.

Hall, M. A. (1998). Correlation-based feature subset selection for machine learning. *Thesis submitted in partial fulfillment of the requirements of the degree of Doctor of Philosophy at the University of Waikato*.

Heiple, S., & Sailor, D. J. (2008). Using building energy simulation and geospatial modeling techniques to determine high resolution building sector energy consumption profiles. *Energy and Buildings, 40*(8), 1426-1436.

HES. (2018). Home Energy Saver & Score: Engineering Documentation. Retrieved from http://hes-documentation.lbl.gov/calculation-methodology/calculation-of-energy-consumption/major-appliances/miscellaneous-equipment-energy-consumption/default-energy-consumption-of-mels

Higashino, M., Fujimoto, T., Yamaguchi, Y., & Shimoda, Y. (2014). *Simulation of Home Appliance Use and Electricity Consumption to Quantify Residential Energy Management Resources.* Paper presented at the Proceedings of the 2nd Asia Conference of International Building Performance Simulation Association, Nagoya, Japan.

Holmes, G., Hall, M., & Prank, E. (1999). *Generating rule sets from model trees.* Paper presented at the Australasian Joint Conference on Artificial Intelligence.

Hong, T., Taylor-Lange, S. C., D'Oca, S., Yan, D., & Corgnati, S. P. (2016). Advances in research and applications of energy-related occupant behavior in buildings. *Energy and Buildings, 116*, 694-702.

Howard, B., Parshall, L., Thompson, J., Hammer, S., Dickinson, J., & Modi, V. (2012). Spatial distribution of urban building energy consumption by end use. *Energy and Buildings, 45*, 141-151.

Huang, C.-L., & Wang, C.-J. (2006). A GA-based feature selection and parameters optimizationfor support vector machines. *Expert Systems with Applications, 31*(2), 231-240.

Huang, Z. (1998). Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data Mining and Knowledge Discovery, 2*(3), 283-304.

ICC. (2009). 2009 International Energy Conservation Code. Washington, DC: International Code Council.

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters, 31*(8), 651-666.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112): Springer.

Johnson, B. J., Starke, M. R., Abdelaziz, O. A., Jackson, R. K., & Tolbert, L. M. (2014). *A method for modeling household occupant behavior to simulate residential energy consumption.* Paper presented at the Innovative Smart Grid Technologies Conference (ISGT), 2014 IEEE PES.

Kahneman, D., Krueger, A. B., Schkade, D. A., Schwarz, N., & Stone, A. A. (2004). A survey method for characterizing daily life experience: The day reconstruction method. *Science, 306*(5702), 1776-1780.

Kavousian, A., Rajagopal, R., & Fischer, M. (2013). Determinants of residential electricity consumption: Using smart meter data to examine the effect of climate, building characteristics, appliance stock, and occupants' behavior. *Energy, 55*, 184-194.

Kodinariya, T. M., & Makwana, P. R. (2013). Review on determining number of Cluster in K-Means Clustering. *International Journal, 1*(6), 90-95.

Kolesnikov, A., & Trichina, E. (2012). *Determining the number of clusters with rate-distortion curve modeling.* Paper presented at the International Conference Image Analysis and Recognition.

Kolter, J. Z., & Ferreira Jr, J. (2011). *A Large-Scale Study on Predicting and Contextualizing Building Energy Usage.* Paper presented at the AAAI.

Lagakos, S. W. (2006). The challenge of subgroup analyses—reporting without distorting. *New England Journal of Medicine, 354*(16), 1667-1669.

Li, Z., & Jiang, Y. (2006). Characteristics of cooling load and energy consumption of air conditioning in residential buildings in Beijing. *Heating Ventilating & Air Conditioning, 36*(8), 1-6.

Lovie, P. (2005). Coefficient of variation. *Encyclopedia of statistics in behavioral science*.

Lutzenhiser, L. (1993). Social and behavioral aspects of energy use. *Annual Review of Energy and the Environment, 18*(1), 247-289.

Ma, J., & Cheng, J. C. (2016). Estimation of the building energy use intensity in the urban scale by integrating GIS and big data technology. *Applied Energy, 183*, 182-192.

Madhulatha, T. S. (2012). An overview on clustering methods. *arXiv preprint arXiv:1205.1117*.

Mayfield, E., Adamson, D., & Rose, C. (2014). *LightSide Researcher's Workbench User Manual*. Pittsburgh, PA: Carnegie Mellon University.

Mayfield, E., & Rose, C. P. (2010). *An interactive tool for supporting error analysis for text mining.* Paper presented at the Proceedings of the NAACL HLT 2010 Demonstration Session.

Mitchell, T. M. (1997). Machine learning. 1997. *Burr Ridge, IL: McGraw Hill, 45*(37), 870-877.

Mo, Y. (2018). *Occupant Behavior Prediction Model Based on Energy Consumption Using Machine Learning Approaches.* (Doctoral Dissertation), Michigan State University, East Lansing, MI.

Ouellette, J. A., & Wood, W. (1998). Habit and intention in everyday life: The multiple processes by which past behavior predicts future behavior. *Psychological Bulletin, 124*(1), 54.

Ouyang, J., & Hokao, K. (2009). Energy-saving potential by improving occupants' behavior in urban residential sector in Hangzhou City, China. *Energy and Buildings, 41*(7), 711-720.

Piedmont, R. L. (2014). *Latent Variables. In Encyclopedia of Quality of Life and Well-Being Research*. Netherlands: Springer.

Quinlan, J. R. (1992). *Learning with continuous classes.* Paper presented at the 5th Australian joint conference on artificial intelligence.

Ramsey, S. A., Klemm, S. L., Zak, D. E., Kennedy, K. A., Thorsson, V., Li, B., . . . Litvak, V. (2008). Uncovering a macrophage transcriptional program by integrating evidence from motif scanning and expression dynamics. *Plos Computational Biology, 4*(3), e1000021.

Reinhart, C. F., & Davila, C. C. (2016). Urban building energy modeling–A review of a nascent field. *Building and Environment, 97*, 196-202.

Sanquist, T. F., Orr, H., Shui, B., & Bittner, A. C. (2012). Lifestyle factors in US residential electricity consumption. *Energy Policy, 42*, 354-364.

Santin, O. G. (2011). Behavioural patterns and user profiles related to energy consumption for heating. *Energy and Buildings, 43*(10), 2662-2672.

Santin, O. G., Itard, L., & Visscher, H. (2009). The effect of occupancy and building characteristics on energy use for space and water heating in Dutch residential stock. *Energy and Buildings, 41*(11), 1223-1232.

Scikit-Learn. (2017). sklearn.preprocessing.StandardScaler. Retrieved from http://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html

Shermis, M. D., & Burstein, J. (2013). *Handbook of automated essay evaluation: Current applications and new directions*: Routledge.

Smappee. (2018). How does Smappee's appliance recognition technology work? Retrieved from https://www.smappee.com/us/blog/smappee-appliance-recognition/

Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 63*(2), 411-423.

Triandis, H. C. (1979). *Values, attitudes, and interpersonal behavior.* Paper presented at the Nebraska symposium on motivation.

U.S.BLS. (2018). American Time Use Survey. Retrieved from https://www.bls.gov/tus/overview.htm

U.S.DOE. (2015). Building Efficiency. Retrieved from http://www.energy.gov/eere/efficiency/buildings

U.S.DOE. (2017). How much energy is consumed in U.S. residential and commercial buildings? Retrieved from https://www.eia.gov/tools/faqs/faq.php?id=86&t=1

Van Raaij, W. F., & Verhallen, T. M. (1983). A behavioral model of residential energy use. *Journal of Economic Psychology, 3*(1), 39-63.

Vassileva, I., Wallin, F., & Dahlquist, E. (2012a). Analytical comparison between electricity consumption and behavioral characteristics of Swedish households in rented apartments. *Applied Energy, 90*(1), 182-188.

Vassileva, I., Wallin, F., & Dahlquist, E. (2012b). Understanding energy consumption behavior for future demand response strategy development. *Energy, 46*(1), 94-100.

Wang, Y., & Witten, I. H. (1996). Induction of model trees for predicting continuous classes.

WholesaleSolar. (2018). How Much Power Do Your Appliances Use? Retrieved from https://www.wholesalesolar.com/solar-information/how-to-save-energy/power-table

Widén, J., & Wäckelgård, E. (2010). A high-resolution stochastic model of domestic activity patterns and electricity demand. *Applied Energy, 87*(6), 1880-1892.

Wilkinson, L., Engelman, L., Corter, J., & Coward, M. (2004). Cluster Analysis *Systat*

(Vol. 11, pp. 65-124): University of Illinois Urbana-Champaign.

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). *Data Mining: Practical machine learning tools and techniques*: Morgan Kaufmann.

Wood, G., & Newborough, M. (2003). Dynamic energy-consumption indicators for domestic appliances: environment, behaviour and design. *Energy and Buildings, 35*(8), 821-841.

Wood, W., Quinn, J. M., & Kashy, D. A. (2002). Habits in everyday life: Thought, emotion, and action. *Journal of personality and social psychology, 83*(6), 1281.

Wood, W., Tam, L., & Witt, M. G. (2005). Changing circumstances, disrupting habits. *Journal of personality and social psychology, 88*(6), 918.

Yoshino, H., Hong, T., & Nord, N. (2017). IEA EBC annex 53: Total energy use in buildings—Analysis and evaluation methods. *Energy and Buildings, 152*, 124-136.

Yu, Z., Fung, B. C., Haghighat, F., Yoshino, H., & Morofsky, E. (2011). A systematic procedure to study the influence of occupant behavior on building energy consumption. *Energy and Buildings, 43*(6), 1409-1417.

Zhao, D., McCoy, A. P., Du, J., Agee, P., & Lu, Y. (2017). Interaction effects of building technology and resident behavior on energy consumption in residential buildings. *Energy and Buildings, 134*, 223-233.

Zhou, J. (2016). *Machine Learning - Regression: Bias and Variance (Lecture Note)*. Retrieved from East Lansing, MI: