SYNTHESIZING POPULATION FOR TRAVEL ACTIVITY ANALYSIS

by

Hong Zhuo

A Dissertation Submitted in

Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

in Geography

at

The University of Wisconsin-Milwaukee

August 2017

ABSTRACT

SYNTHESIZING POPULATION FOR TRAVEL ACTIVITY ANALYSIS

by

Hong Zhuo

The University of Wisconsin-Milwaukee, 2017
Under the Supervision of Professor Zengwang Xu

Population synthesis is a fundamental procedure for individual-based modeling in transportation research. The population synthesis generates anonymized individuals with selected social-demographic variables that have similar statistical distributions as that of the samples from the real population. Previous studies on population synthesis focused on generating general-purpose population by fitting the joint distributions of multiple variables to their sampled distributions. In addition to fitting the joint distributions, this study focuses on generating population for travel activity analysis by considering individuals' travel activity patterns and associated social, economic, and demographic characteristics.

A person's daily movement is a time-sequence of activities connected by travel behaviors. It can be described as vectors that include important transportation attributes such as travel distance, travel mode, activity type, activity time, and activity sequence. A multidimensional pattern vector method is used in this study to represent an individual's daily travel activities. This method is based on the combination of time-geography, sequence alignment, and pattern vector. Using the 2001 and 2009 National Household Travel Survey (NHTS), the travel distance and

activity sequence of individuals are normalized, compared, and integrated into a dissimilarity matrix. Major travel activity patterns are then examined by cluster analysis. The random forest model is applied to examine the prominent socio-demographic characteristics that correlate to the activity patterns. The prominent socio-demographic characteristics are then used to synthesize population microdata. Since the algorithm complexity of population synthesis grows exponentially with the number of attributes, the methodology used in this study can effectively reduce the computational intensity by focusing on the most important variables for travel activity analysis.

This study also addresses another issue in traditional population synthesis algorithms, i.e., the probability distributions at the individual and household levels cannot be fitted simultaneously. In this study, Iterative Proportional Fitting (IPF) algorithm is used to consider the distributions at different scales and to generate synthetic population microdata with the prominent socio-demographic characteristics. The performance of the algorithm that generates synthesized population is evaluated by scatter plot and Normalized Root Mean Square Error (NRMSE) analysis. In addition, the distributions of socio-demographic attributes in the synthesized data are compared with that of variables in the observed sample dataset. The verification result indicates that the new method can produce a better population microdata.

This dissertation describes how to generate a synthetic population for Milwaukee County, WI with prominent socio-demographic variables for travel activity analysis. By critically selecting the prominent socio-demographic factors, the computational intensity of population synthesis is reduced. It is also found that, by aggregating the IPF-generated weights of individuals and using them to the household level, the overall goodness-of-fit can be managed at a reasonable level and

the distributions of socio-demographic factors at the individual and household levels can be fitted.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter One: Introduction

1.1 Significance

Areal aggregate data have been the major data source of many traditional transportation analyses (McNally, 2008). One of the major limits of the areal aggregate data is the ignorance of characteristics and behaviors at the individual level. The limitation of the traditional transportation models has prompted the emergence and development of individual-based transportation models. The individual-based transportation modeling has been widely used in contemporary travel activity analysis, such as the activity-based travel demand models (Pritchard & Miller, 2012). The individual-based modeling is a state-of-the-art technology in transportation modeling and has several advantages over the traditional transportation modeling. Firstly, it can predict the state in the future by simulating individuals' decisions over a period of time (Pritchard & Miller, 2012), providing more details on how individuals travel and participate in activities (Müller & Axhausen, 2010). Secondly, it is more sensitive to the difference between individuals' complex travel activity patterns (Recker, 1995). For these reasons, the individual-based transportation models have been widely used in travel activity analysis.

To develop individual-based transportation models, information describing how individuals behave is required, which is what microdata can offer. Microdata refer to the information at the individual level. In transportation research, there exists extensive literature on the application of microdata about travel behaviors and traveler's socio-economic characteristics. For example, Ballas et al. (1999) pointed out that microdata are necessary for the transportation policy analysis in which they can be used to micro-simulate an individual's social and economic activities.

Moreover, activity-based transportation models, such as the well-known transportation modeling system TRansportationANalysisSIMulation System (TRANSIMS), use population microdata generated by the Beckman's (1996) algorithm to simulate individuals' daily travel activities.

However, obtaining data at the micro level is time-consuming, expensive and even cost prohibitive (Mohammadian et al., 2010). Hermes and Poulsen (2012) acknowledged that micro-level social surveys are important sources of data for transportation research, but the time and cost often constrain those surveys to small sample sizes. These difficulties prompt us to try using limited population samples to synthesize a full size of population, and therefore, the synthetic population is developed as a major microdata, which can serve as the basis of individual-based transportation modeling. The synthetic population is computer-generated microdata in which each individual has the simulated social, economic, and demographic attributes, and these attributes have the similar distribution as that of the real population (Auld, Mohammadian, & Wies, 2009).

A major drawback of the synthetic population is that its computational intensity and complexity increases exponentially with the number of the attributes (Pritchard & Miller, 2012). It would be ideal but not possible to generate a universal-purpose synthetic population that has all the characteristics of the real-world human beings. Therefore, it is critical to select the most important socio-demographic variables for the generation of the synthetic population. In general, the selection of variables depends on research topics. For example, if a research question is about poverty status, it is reasonable to include the household income and employment status into the list of variables of interest. If a study focuses on the dissemination of a disease, the age and gender might be deemed to be variables necessary to be synthesized. This study aims to generate a synthetic population that can be effectively used in travel activity analysis. Because

there is no current consensus on which socio-demographic characteristics should be selected for travel activity analysis, one emphasis of this research will be placed on exploring the statistically significant socio-demographic variables to be selected for the generation of synthetic population.

This study will analyze the travel activity patterns based on the 2001 and the 2009 National Household Travel Survey (NHTS), explore the prominent social, economic, and demographic characteristics that correlate to the major travel activity patterns, and generate a synthetic population with these characteristics. A method is designed to integrate the different aspects of the human daily travel activities to explore the major patterns. An individual's daily travel activity is described as a chain of timed activities linked by travel behaviors. Characteristics of travel and activity are both considered when individual's daily travel activities are compared and clustered to discover the major patterns.

In addition, it would be important to know who are the people performing the specific daily travel activities. It is well recognized that people with certain socio-demographic characteristics are related to particular travel and activity behaviors (Hanson, 1982; Pas, 1984). Especially, socio-demographic characteristics of individuals and households affect the travel and activities a person or a family performs (Hanson & Hanson, 1981). Given that human travel activities may be classified into several discrete groups, the random forest model is used to examine the relationship between different kinds of travel activities and the individuals' social, economic, and demographic characteristics.

As a result, this study generates a synthetic population that is endowed with the prominent socio-demographic characteristics for travel activity analysis. Population synthesis refers to the process that generates the synthetic population microdata. The most commonly used method in

the population synthesis is the Iterative Proportional Fitting (IPF) algorithm invented by Deming and Stephan (1940). The IPF algorithm is a procedure that iteratively adjusts cell values in a contingency table until it matches with the predefined marginal totals (Fienberg, 1970; Wong, 1992; Ye et al., 2009). The two-step IPF algorithm developed by Beckman et al. (1996) is the most conventional and well-known method used to synthesize population. However, Beckman's method can only match the distribution between the reproduced and real microdata at the household level. It is unable to fit the statistical distributions of the individual and household characteristics simultaneously (Pritchard & Miller, 2012; Guo & Bhat, 2007). In this study, a method is needed to fit the distributions at the individual and household levels simultaneously. The aggregated household weights are used to improve the IPF algorithm to ensure the distributions at the individual and household levels can be fitted. After the synthetic population is generated, a verification process will be conducted to evaluate the performance. There are many different measures to verify the synthetic population. In this study, the Normalized Root Mean Square Error (NRMSE) goodness-of-fit is used for the verification purpose. In addition, the distributions of simulated socio-demographic variables are compared with that of variables in the observed data to see whether their distributions are consistent.

1.2 Objectives

This dissertation has three objectives:

1. To examine the major daily travel activity patterns based on the 2001 and 2009 National Household Travel Survey (NHTS). Examining the major travel activity patterns requires not only an appropriate method to represent individuals' travel activities, but also an

effective clustering analysis method to find out the major patterns. This dissertation will review previous literature on representing individuals' travel activities, design a proper method to represent them, explore the major travel activity patterns, and find out the differences between the 2001 and 2009 NHTS.

2. To investigate the prominent socio-demographic variables that correlate to the major travel activity patterns. To explore the prominent socio-demographic variables, this research will review previous literature that examined the relationship between travel activity patterns and socio-demographic variables, and summarize the important socio-demographic variables related to travel activity patterns. In addition, this research will also discuss the previous methods that were used to investigate their relationship, and propose an appropriate method to explore the prominent socio-demographic variables.
3. And, to generate the synthetic population that are endowed with those prominent socio-demographic attributes for travel activity analysis. This dissertation will outline the significance of generating synthetic population, introduce the most commonly used algorithms, discuss the limitations of existing algorithms, and propose a new method to overcome the problems. After the population synthesis is completed, this dissertation will evaluate the performance to measure the degree of accuracy.

1.3 Dissertation Organization

The rest of this dissertation is organized as follows. Chapter Two reviews the previous literature on travel activity analysis, the correlation between socio-demographic variables and travel activity pattern, and population synthesis respectively. In particular, the commonly used methods representing human travel activities, which include time-geography, sequence

alignment, and pattern vectors, are examined. The literature review also discusses the socio-demographic factors that are related to human activity patterns, the approaches to analyze their relationship, the methods to generate the synthetic population, and the technique to verify the synthetic population.

Chapter Three introduces the data sources and methodology employed in this research to study human travel activity patterns, its relationship with socio-demographic characteristics, and population synthesis. Specifically, it contains (a) the method to represent human daily travel activities in the form of pattern vectors, (b) the approach to derive dissimilarity matrices for travel activities and activity sequence, (c) the technique to normalize and integrate dissimilarity matrices for cluster analysis, (d) the classification model to be used to examine the relationship between human activity patterns and socio-demographic characteristics, (e) the method to utilize the IPF-based algorithm to synthesize the population, and (f) the goodness-of-fit statistics to validate the synthetic population.

In Chapter Four, this dissertation summarizes the results, including the major travel activity patterns in the 2001 and 2009 NHTS, the prominent socio-demographic variables, and overall goodness-of-fit of the synthesized population microdata. To be specific, it first discusses the analysis results on the 2001 and 2009 NHTS, presents the major travel activity patterns, and provides a summary description for each activity pattern. Then, it presents the classification model's accuracy and the importance of socio-demographic variables. Lastly, this chapter evaluates the performance of the population synthesis, and compares the distributions of the socio-demographic factors in the synthesized population against that of variables in the observed data.

In Chapter Five, the conclusions and the direction of future research are included. It summarizes the major conclusions and proposes the direction of future research.

Chapter Two: Literature Review

2.1 Representation of Individual's Travel Activities

Travel activity analysis is an important part of transportation research. Appropriately representing an individual's travel activities is the first step of travel activity analysis. Human travel activities can be represented in different ways. For example, Roseman (2010) proposed a reciprocal movement cycle to represent individual or household movement over a time span of one week. As another example, an individual's daily travel activities were described as a space-time prism as illustrated in Kwan (2004), Kwan and Lee (2004), and Kwan and Ren (2008). In general, the representation of human daily travel activities is mainly based on three methods: time-geography method, sequence alignment, and pattern vectors.

2.1.1 Time-Geography Method

Time-geography method is one of the earliest approaches in representing human travel and activities in space (Kwan, 2004). The time-geography concept was first proposed by Hägerstrand (1970) at the 9th European Congress of the Regional Science Association. A core viewpoint in Hägerstrand's time-geography conceptual framework is that time is a very important variable, and it should be given more consideration in geography. Therefore, Hägerstrand (1970) believed that both space and time are important properties in the research on individuals' travel and activity behaviors. He considered that a person's daily movement could be described as a space-time path in three-dimensional space as illustrated in Figure 2.1. In this three-dimensional representation, (x, y) coordinates denote the activity locations, and the vertical

z-axis represents the time axis. In this notation, an individual's daily travel and activities can be described as a space-time path. For example, in Figure 2.1, a space-time path is used to represent an individual's daily activities; the individual leaves his/her home in the morning for work, attends social recreation after work, and then returns home in the evening.

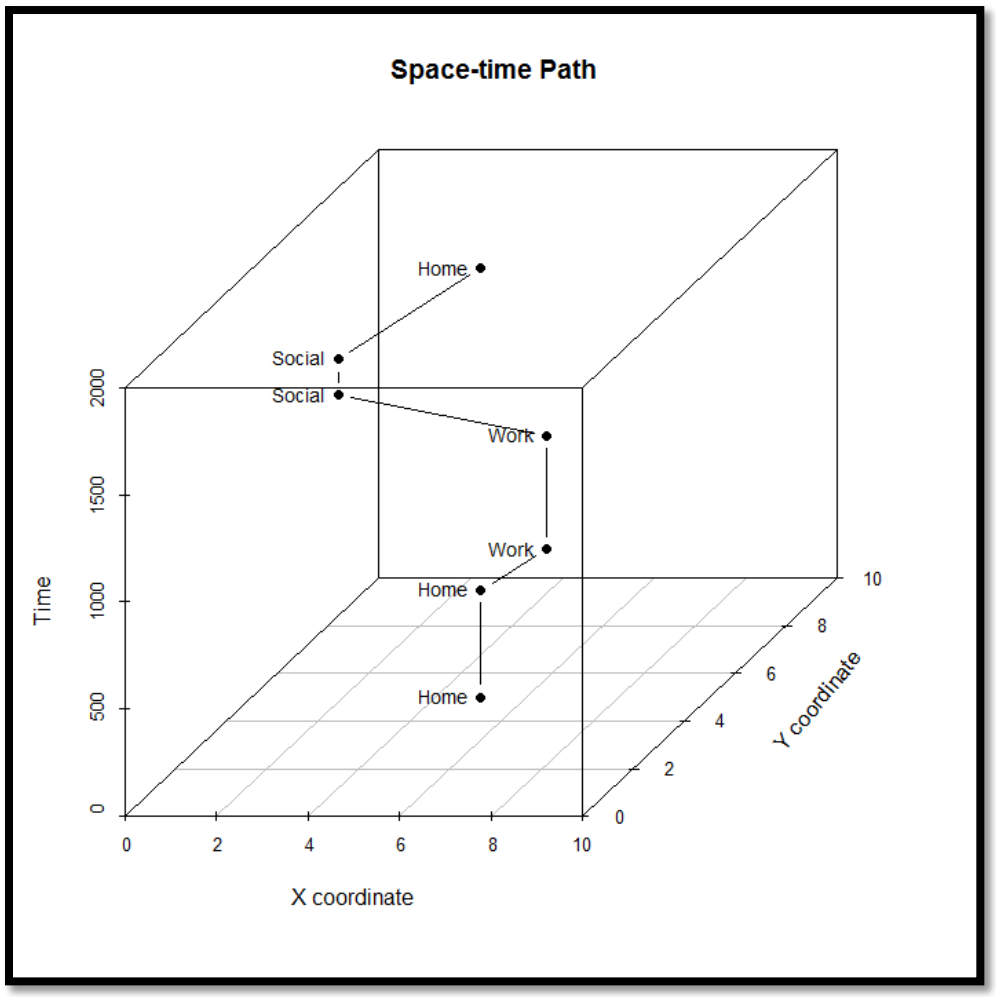

Figure 2.1 An example of an individual's space-time path

Another fundamental concept in time-geography is the spatial and temporal constraints. Due to the spatial and temporal constraints, people cannot travel to anywhere at any time (Hägerstrand,

1970). For example, assuming that the individual in Figure 2.1 spends one hour on social recreation after his work at one location, the individual is "constrained" in that location during the period, and the individual should not be considered being at any other locations. As another example, employed people usually go to work in the morning, and then they go home after work in the evening. Their daily movements are "constrained" by the work location and the work time. These two examples suggest that the spatial and temporal constraints are important components of individuals' travel and activity behaviors.

Since its introduction, time-geography has been the fundamental concept for many geographical studies during the past four decades. It has many applications in geography, and a major example is the accessibility research. In transportation, accessibility is an indicator that measures the degree of easiness to which a location can be accessed by people (Kwan, 1998). Kwan (1998, 1999) pointed out that most of conventional accessibility research methods are only based on location proximity, which are less sensitive to the personal differences in travel and activities. In today's accessibility studies, the time-geography concept has been constantly based by many researchers like Lenntorp (1976), Miller (1991), Miller (1999), Kwan (1998), and Kwan (1999). For example, Lenntorp (1976) proposed Potential Path Space (PPS) and Potential Path Area (PPA) in his accessibility study. The PPS, also known as space-time prism (Kwan, 2004), is a person's potentially accessible section under spatial and temporal constraints, and its projection on a two-dimensional plane is the PPA (Miller, 1991; Kwan, 1999). Kwan (1998, 1999) also brought the space-time concept into her accessibility research, and she argued that the time element should not be excluded from the accessibility study because the space-time measurement of accessibility outperforms the conventional method since it is more sensitive to personal differences.

In addition to the accessibility research, Hägerstrand's time-geography has also found its applications in the latest travel activity analyses, and it represents a major effort among geographers who study human travel activities (Kwan, 2000; Kwan, 2004; Kwan & Ren, 2008). For instance, Kwan (2004) pointed out that, built on time-geography conceptual framework, the geo-visualization is a useful method to examine human travel activities. For instance, Kwan (2000) described the geo-visualization methods under the GIS environment used to analyze human activity pattern in both spatial and temporal dimensions for the Portland metropolitan area, and she concluded that geo-visualization methods provide a better understanding of human travel and activity behaviors although they are computationally intensive. As another example, in a recent study, Ding (2009) also incorporated the time-geography method in his human activity pattern study. In his research, he sampled individuals from the travel survey conducted in Portland in 1994, and described a person's daily movement as a space-time trajectory in a three-dimension space where the vertical axis is the time. Each space-time trajectory is composed of a sequence of activities chained by travel behaviors in 24 hours. He used the multidimensional sequence alignment analysis and clustering analysis to partition the data set into several groups and then derived some common activity patterns. In Ding's study, the time-geography plays an important role in describing an individual's daily travel and activities.

Despite these developments, the application of time-geography method remains challenging. Firstly, the application of time-geography method is restricted by the data availability. The time-geography method requires data at the micro level, but not enough individual level data are available due to its collection cost and privacy issue. Spatially explicit individual data can be obtained with the rapid development of location-tracking technology such as GPS and smartphone triangulation, but they are very costly or not accessible. Secondly, it is difficult to

implement the time-geography in current GIS environment because of its complexity (Miller, 1991; Miller, 1999). Many researchers pointed out that algorithms based on the time-geography concept are time-consuming and computationally intensive (Kwan, 2000; Kwan, 2004; Kwan & Lee, 2004). For example, Ding's (2009) algorithm based on the time-geography concept took up an enormous amount of computational resources. Last but not least, the time-geography method is mainly used to visualize travel activity patterns (Kwan & Lee, 2004). It is difficult to integrate the time-geography concepts in current GIS platform which is mainly based on two-dimensional cartography. Hence, the time-geography method provides a useful way to geo-visualize human daily travel and activities, but not a very efficient analytical method (Kwan & Lee, 2004).

2.1.2 Sequence Alignment Method

Another method is to represent human travel and activities as an activity sequence. An individual's travel activities can be viewed as a sequence of activities linked by travel (Root & Recker, 1981). Many of these travel behaviors and activities are what people perform on a daily basis. As such, an individual's daily travel activities have a high similarity to many sequences encountered in other disciplines, such as a DNA sequence in biology. In biology, a popular method commonly used to analyze a DNA sequence is sequence alignment method. For example, Needleman and Wunsch (1970) invented a global pairwise sequence alignment algorithm to search for the maximum matches between any two amino acid sequences. Sequence alignment method has been applied to the social science by Abbott and Forrest (1986) and Abbott and Hrycak (1990). Thrift (2002) considered the sequence alignment method one of the new quantitative methods in geography.

The sequence alignment method can incorporate the different patterns of activity sequence into the travel activity analysis. It has shown great potential in analyzing human activity behaviors and has been used by many researchers, for example Wilson (1998), Joh et al. (2001), Joh et al. (2002), Ding (2009), and Shoval and Isaacson (2007). Wilson (1998) considered that the best alignment method is the one that is able to maximize the similarity while reducing the difference to the lowest extent, which is very similar to the basic principle of clustering analysis that is to minimize the difference within clusters and maximize the difference between clusters. The process of sequence alignment can be carried out on the element-wise operations (such as inserting gaps, deleting gaps, and substitution) of sequences (Wilson, 1998). Joh et al. (2002) proposed a new multi-dimensional alignment method to compare the similarity and difference between activity sequences. Their model can take into consideration both the activity sequence and the interdependency between different attributes. They believed that their method could be used to conduct a goodness-of-fit test for simulated activity-based models. Ding (2009) considered that the basic idea of the sequence alignment methods is to find out the best alignment between any two comparable sequences with the least dissimilarity. He treated human travel activities as multidimensional sequences in which time, travel distance, activity type, and transportation mode are all incorporated, and then undertook trajectory-based activity analysis in a GIS environment. He combined event-based with time-sliced methods to encode human travel activities into multidimensional sequences, and then utilized pairwise similarity as a measurement of the difference between sequences. In this dissertation, the sequence alignment method will also be utilized to examine the similarity or dissimilarity between two activity sequences.

2.1.3 Pattern Vectors

Root and Recker (1981) considered that an individual's daily travel and activities is a sequence of activities chained by time and linked by travel behaviors. In addition, the individual's travel and activities can also be a sequence of other features of travel and activities, such as travel distance, transportation mode, activity time, activity duration, the number of activities performed, and activity locations, etc. An individual's daily travel and activities is multidimensional. Therefore, the multidimensional pattern vector method can be used to represent and analyze human travel activities. Recker et al. (1981) represented human travel activity as a multi-dimensional pattern vector, $C = \delta(X)$, where the C denotes the activity pattern category, and the X represents the multidimensional pattern vector which is composed of representative transportation attributes. In order to reduce the complexity of multidimensional human travel activities, one commonly decomposes the multidimensional vectors into simplified forms. For example, Recker et al. (1981) decomposed multidimensional pattern vectors by projecting them onto the space-time and activity-time axes, on which they referred the space and activity axes as distance to home and activity type respectively.

Recker et al. (1985) deemed the activity pattern analysis as seeking solutions to activity pattern similarity problem or classification problem. They described an individual's daily travel activities as a pattern vector, and the components of the vector were the representative variables about the travel or activities, such as travel mode, activity purpose, and activity duration. The human travel activities in their work were represented as 3D pattern vectors, and then decomposed into two corresponding 2D spaces. Walsh-Hadamard's transformation technique was then applied to reduce the complexity, and activity pattern analysis was conducted afterward (Recker, McNally, & Root, 1985).

In summary, although the time-geography method is a useful approach to visualize individuals' daily activity, it has not yet well developed as an analytical tool because it is complicated and takes up an enormous amount of computation resources (Kwan, 2004). The sequence alignment method has not been well applied in travel activities studies because of its limitations in dealing with a large dataset and non-categorical data (Joh, Arentze, Hofman, & Timmermans, 2002). In this study, a new method is designed to integrate the different aspects of the human daily travel activities to explore the major patterns in the National Household Travel Survey (NHTS). The method is based on the concepts of time-geography, and takes advantages of the advances of the sequence alignment and pattern vector methods.

2.2 Socio-Demographic Characteristics Related to Travel Activities

2.2.1 Socio-Demographic Characteristics of Individuals

In many transportation studies, individuals' travel activity patterns are considered being closely correlated to socio-demographic characteristics (Hanson, 1982; Pas, 1984; Joh, Arentze, Hofman, & Timmermans, 2002). For instance, Levinson and Kumar (1995) found that gender and employment status are important factors that impact individual's travel behaviors, and the increasing participation rate of females in job markets affects the time allocated at home and non-work travel. Pucher and Renne (2004) used the 2001 NHTS dataset to examine the differences in travel activity patterns between urban and rural areas, and they found that residential location is a critical factor, due to the different transportation modes between the urban and rural areas. Compared with urban areas, people who live in rural areas own a higher percentage of private vehicle (Pucher & Renne, 2004). Tal and Handy (2010) applied

multivariate analysis to analyze the 2001 NHTS to examine whether immigrants' travel patterns are influenced by socio-demographic traits, such as age, race, gender, household life cycle, vehicle ownership and so forth, and they found that the arrival time to the United States, place of birth, and motherland's culture impact immigrants' activity patterns.

Hanson and Hanson (1981) considered that a few major socio-demographic factors are essential to human travel activity patterns, and these factors include: the socio-economic status (such as income, occupation, and education level), social roles (such as age, gender, and life cycle), and vehicle availability for each household. Among these essential socio-demographic factors, four of them are discussed here: race, gender, age, and household life cycle. Firstly, race has been an important socio-demographic variable that influences individuals' travel activity patterns (Giuliano, 2003). For example, Tal and Handy (2005) analyzed the 2001 NHTS to extract the impact factors on immigrants' travel behaviors, and they found that race/ethnicity is a crucial factor. More specifically, they found that immigrants, compared with the native US citizens, prefer to use the public transit at the early period of their arrival in the US, and white immigrants make more trips per day and have a higher percentage of driving private vehicle than any other races. Moreover, race/ethnicity is considered a highly correlated factor that will influence the perception of transportation services such as infrastructure condition and traffic condition, which further impact their travel behaviors (Tal & Handy, 2005).

Secondly, gender is also recognized as a critical socio-demographic attribute that affects a person's travel and activities (Levinson & Kumar, 1995). Many have examined the relationships between gender and travel. For example, Hanson and Hanson (1980) analyzed the travel survey data collected in Uppsala, Sweden in 1971 and examined whether gender will affect the travel activity patterns. They compared travel activity data about men and women with full-time

employment status to see if there is a gender effect in their activity behaviors. They found that there is a huge difference between working men and working women on many aspects of a person's activity behavior, such as travel distance, transportation means, trip purpose, etc. In another example, Kwan (1999) studied the difference of accessibility to urban opportunities between men and women, and she found that women have relatively fewer opportunities than men. These are just two simple examples concluding that gender is a critical variable.

The third socio-demographic characteristic that impacts individuals' travel and activities is age. The association of age and travel activities can also be found in many existing literature, such as Collia et al. (2003) and Mohammadian and Bekhor (2008). Collia et al. (2003) analyzed the 2001 NHTS to compare the travel patterns distinction between the young (19-64 years old) and the old (65+ years old) in the United States. They found that, when compared with the senior adults, the young adults make more trips, travel longer distances, and travel longer time. As in another example, Mohammadian and Bekhor (2008) reviewed the previous literature about the effect of socio-demographic factors on certain groups of population, and they found the baby-boomers' travel and activities are different from the younger and older generation.

The last socio-demographic variable discussed here is household life cycle. The household life cycle is determined by age, marital status, and the number of children (Pas, 1984). It has been found by many researchers that the household life cycle is a non-negligible factor that shapes human travel activity patterns (Hanson & Hanson, 1981; Hanson, 1982). For example, Mohammadian and Bekhor (2008) pointed out that the 2001 NHTS analysis result shows that the household life cycle affects individuals' travel and activities, and found that the life cycle of retired elderly people makes them prefer to travel during the traffic peak hours, but they are less inclined to use public transit as the transportation mode. As another example, Pas (1984) found

that whether a couple has children or not will influence the travel and activity behaviors of a family. Assuming that a couple has a ten-year-old child, then dropping off and picking up the child at the elementary school will become a routine activity. As such, the couple's travel and activities have been spatially and temporally affected by their child's routine. This is a simple example of how household life cycle plays a role in influencing individuals' travel activity patterns.

In summary, from the previous literature, we can see that socio-demographic variables indeed relate to an individual's travel activity patterns. In this dissertation, these socio-demographic factors will be used to explore their correlation with travel activity patterns.

2.2.2 Methods to Explore the Relationship Between Socio-Demographic Characteristics and Travel Activity Patterns

Various statistical methods have been used to explore the relationships between activity patterns and socio-demographic descriptors. A useful method is the Principal Component Analysis (PCA), which is a statistical technique that converts a large number of variables into a few components. Hanson and Hanson (1981) applied PCA to the individual level travel survey data to reduce a large number of travel activity measures into several major factors. Then, they used a linear multivariate regression model to carry out a stepwise regression analysis to examine the relationship between socio-demographic descriptors and the major factors reduced by the PCA. This methodology that combines PCA with stepwise regression analysis was also used by Hanson (1982).

Another method is non-parametric grouping/clustering methods employed by Janelle and Goodchild (1983) and Janelle et al. (1988). For example, Janelle et al. (1988) utilized several grouping methods, which include dichotomous variable grouping, four/six-variable role group aggregation and cluster analysis based on Ward's algorithm, to partition space-time diary data, and then evaluated the performances of all methods by discriminant analysis and graphic measures.

A similar technique to explore the relationship between travel activity patterns and socio-demographic characteristics is classification model, such as linear discriminant analysis and multinomial logit regression. The linear discriminant analysis uses a linear combination of explanatory variables to divide a dataset into several categories. Recker (1995) presented a method to partition human travel activities into several categories, and then used the linear discriminant analysis to correlate the activity patterns with socio-demographic factors. The socio-demographic factors employed in his research include household characteristics (such as employment status and gender) and urban form indicators (such as the density of household units and the density of the employment in an urban environment). The logit regression analysis can be used to explore the correlation between socio-demographic factors with travel activity patterns. The logit regression model is usually used when the dependent variable is categorical data. According to the number of categories for the dependent variable, the logit regression model can be divided into two types: binary logit model (two categories) and multinomial logit model (more than two categories). The generalized logit regression can be formulated as below:

$$ln\frac{p(Y=i)}{p(Y=K)} = \beta_0 + \beta_1 * x_1 + \beta_2 * x_2 + \cdots + \beta_n * x_n \qquad \text{(Equation 2.1)}$$

where $i = 1, 2, \ldots, K-1$.

On the left hand of the Equation 2.1, the letter Y denotes the dependent variable and $p(Y = K)$ is the probability of the Kth category. On the right side, $(\beta_0,\ \beta_1,\ \beta_2, \dots, \beta_n)$ are the regression coefficients, and $(x_1,\ x_2, \dots, x_n)$ are the explanatory variables. The logit regression analysis is often used in social science (Pereira & Itami, 1991; Bhat, 1995; Franklin, 1995; Kahn, 2005), and it is considered as an appropriate method to explore the correlated socio-demographic factors (Pas, 1984). For instance, by using the linear logit regression model and parametric maximum-likelihood estimation (MLE), Pas (1984) found that individual activity pattern is a function of socio-demographic indicators, such as life cycle and role in a family.

From previous literature, we can find that classification model is a commonly used technique to examine the relevant socio-demographic attributes. Moreover, in this study, individual's travel activities are partitioned into a few discrete groups, which serve as the categorical dependent variable. For these reasons, classification model is an appropriate choice to explore the relationship between the different groups of activity patterns and its corresponding socio-demographic characteristics. In particular, a classification model called random forest is used to find statistically significant socio-demographic characteristics. More details about why the random forest classification model is chosen will be discussed in Chapter 3.2.5.

2.3 Population Synthesis

2.3.1 The Significance of Population Synthesis

Aggregate data on Traffic Analysis Zones (TAZs) or census tracts are often used in the traditional transportation research. For instance, the well-known four-step travel demand model uses the population data at the TAZs level. The four-step model refers to trip generation, trip

distribution, mode choices and traffic/route assignment (McNally, 2008). However, as time goes by, the limitations of traditional transportation research that are based on aggregate level data began to be noticed. For example, many researchers pointed out that using aggregate level data would cause ecological fallacy and modifiable areal unit problem (MAUP) (Openshaw, 1984; Wong, 1992; Norman, 1999). Moreover, although the data at the aggregate level can provide a basic overview of the population attributes, the details about how an individual travels and participates in activities remain unknown. The limitations of traditional transportation models have hindered the development of transportation analysis. To overcome these limitations, many researchers began to turn to individual-based transportation modeling (Kitamura, 1988; Recker, 1995). The individual-based modeling is a technique which simulates individuals' decision in travel and activity participation to forecast the future status of a transportation system (Pritchard & Miller, 2012). This technique first emerged in the 1980s, and now it has been the most popular technology in transportation research (Kitamura, 1988; Lu & Pas, 1999). The individual-based modeling is popular because it outperforms the traditional modeling in many aspects. For example, the individual-based modeling can describe how individuals travel and participate in activities over a period of time in details, so it is easier than the traditional modeling to capture the differences in individuals' travel activity patterns (Pritchard & Miller, 2012; Müller & Axhausen, 2010).

In the individual-based transportation models, microdata are required. Microdata refer to the data at the individual or household level, and they provide the basis for individual-based transportation models, including the activity-based transportation models (Pritchard & Miller, 2012), such as the TRansportationANalysisSIMulation System (TRANSIMS). The TRANSIMS is an open-source, activity-based transportation simulation model developed by Travel Model

Improvement Program (TMIP) supported by United States Department of Transportation (McNally, 1996). But, the acquisition and application of microdata have been hindered by concerns like personal information privacy, data availability, and small sample size (Mohammadian, Javanmardi, & Zhang, 2010; Hermes & Poulsen, 2012). Therefore, sampled survey is still a major means to collect micro-level data for transportation modeling, although it is time-consuming and expensive (Mohammadian et al., 2010). The U.S. National Household Travel Survey (NHTS) is a nationwide travel survey data that have been widely used to analyze individuals' travel and activities. The NHTS is a comprehensive travel activity survey, and it can provide information about the trips, activities, and socio-demographic factors. However, the temporal interval between two NHTS surveys spans several years and the NHTS survey is not taken at regular interval. The most recent two NHTS were taken place in 2001 and 2009. A similar situation happens to the travel surveys in the regional or metropolitan level, such as the Chicago Regional Household Travel Inventory (CRHTI) conducted by the Chicago Metropolitan Agency for Planning (CMAP). The most recent two travel surveys by the CMAP were conducted in 1990 and 2007, and the gap between them is 17 years. The long gap is mainly due to the long collecting time and the high expense of the surveys (Mohammadian et al., 2010). Hermes and Poulsen (2012) also recognized that applications of microdata are limited by the relatively small sample size, and it is economically unreasonable to conduct a survey to get the information for all individuals.

On the other hand, many individual-based models require every individual being digitally represented. Although the sampled microdata only cover a small portion of the total population, together with other information (often the information at the aggregate level), they can be used as seeds to extrapolate the microdata for the whole population, which are anonymous and synthetic

but have the necessary statistical features. It has been supported by many researchers that it is reasonable to use sample data to create synthetic microdata for the whole population (Ballas et al., 1999; Hermes & Poulsen, 2012). The procedure that uses available sample data to generate synthetic population microdata with socio-demographic variables of interest is called population synthesis (Auld et al., 2009).

2.3.2 The IPF Algorithm

Population synthesis is an important topic in transportation research, and it is an indispensable step for activity-based models (Auld et al., 2009; Müller & Axhausen, 2010). Many population synthesizers are widely used, such as the Population Synthesizer module in TRANSIMS, the Household Attributes Generation System (HAGS), PopSynWin, PopGen, etc. The Population Synthesizer in TRANSIMS has been used in simulating human travel and activities in many cities, such as Portland and Chicago (McNally, 2008). As another example, PopGen has been used to create a synthetic population for Maricopa County in Arizona (Ye et al., 2009).

Among these population synthesizers, the most frequently used algorithm to generate the synthetic population is the Iterative Proportional Fitting (IPF) algorithm (Barrett et al., 2009). The basic IPF algorithm was first proposed by Deming and Stephan in 1940 (Ireland & Kullback, 1968; Fienberg, 1970), and then it was refined and improved by Stephan (1942) and Fienberg (1970). The IPF is an iterative algorithm that repeatedly scales cell values of a contingency table constrained by external marginal totals until it converges (Fienberg, 1970; Wong, 1992; Ye et al., 2009). Compared with other iterative methods, such as EM algorithm and Newton-Raphson method, the IPF algorithm is simpler, easier to understand, and more computationally efficient

(Lovelace et al., 2015). In addition, the IPF algorithm can maintain the interaction denoted by cross-product ratio (Mosteller, 1968) and can maximize the entropy (Johnston & Pattie, 1993). For these reasons, IPF algorithm has been employed in many disciplines, such as transportation and economics.

The IPF algorithm has different names in different fields (Lomax & Norman, 2016), for example it is called Fratar model in transportation engineering (Beckman et al., 1996) and RAS (Ranking And Scaling) in economics (Wong, 1992). In traditional four-step travel demand forecasting models, the IPF algorithm is used to estimate and update an origin-destination (OD) matrix (Bruton, 1985). Wong (1992) claimed that the potential of the IPF algorithm in the geography community has not been thoroughly realized, and he believed the IPF can be used to estimate and generate the micro-level data. In fact, the IPF algorithm has been a popular technique today to generate population microdata. For example, Simpson and Tranmer (2005) implemented the IPF algorithm by using the Statistical Package for the Social Sciences (SPSS) software to generate population by jointly using the sample survey microdata and census aggregate data. Lomax and Norman (2016) also applied the IPF method to estimate population microdata for a subsequent year, when the data source only contain the number of population with socio-economic variables for each area.

The earliest employed and the most established IPF-based method to generate population microdata can be traced back to the two-step IPF algorithm developed by Beckman et al. (1996). The two-step IPF algorithm is adopted in TRANSIMS population synthesizer to generate the synthetic population at individual and household levels by using the decennial census aggregate data and the Public Use Microdata Sample (PUMS) five percent microdata (Beckman et al., 1996). The PUMS data contain two files in which one is at the household level, and the other is

at the individual level. The PUMS data are sampled microdata in the Public Use Micro Area (PUMA) that mostly consists of several census tracts. Based on the definition in the U.S. Census Bureau, each PUMA roughly has a population of 100,000 (Wheaton et al., 2009). The two-step IPF algorithm can be described as follows: firstly, add up the aggregate census tract data in a single PUMA to form a combined census tract, then apply the IPF algorithm to update cell values of the contingency table in the PUMS data; secondly, the IPF algorithm is implemented again by using the result of the first step as the additional dimension marginal constraint (Beckman et al., 1996; Mohammadian et al., 2010). Based on the result generated by the two-step IPF algorithm, households and individuals are selected proportionally from the PUMS to synthesize the households and personal data (Beckman et al., 1996). Beckman's two-step IPF method has been used over the past 20 years; however, it has a few long-standing problems that have not been solved. The most outstanding issue is that Beckman's method does not fit the statistical distributions of the household and individual level socio-demographic variables simultaneously (Pritchard & Miller, 2012; Guo & Bhat, 2007). To be specific, Beckman's method can only fit the distribution at the household level, but it leaves the individual level variables unmatched (Guo & Bhat, 2007; Pritchard & Miller, 2012).

In addition to the two-step IPF method, many other approaches have been proposed to synthesize population with the capacity to control the probability distributions of the household and individual socio-demographic variables simultaneously. For instance, an alternative method for population synthesis is the combinatorial optimization (CO) developed by Williamson et al. (1998). Instead of using the multi-zone method like Beckman's, Williamson et al. (1998) adopted a zone-by-zone method to estimate the population (Pritchard & Miller, 2012). Williamson's algorithm randomly selects a subset from the sample survey to make it suitable for

a single zone. However, Williamson's method is very time-intensive (Pritchard & Miller, 2012). Monte Carlo method is used as another method for population synthesis (Frick & Axhausen, 2004). The Monte Carlo method is an approach that finds an approximate solution by repeatedly generating random numbers based on a probability distribution. It is a commonly used technique when a problem is too difficult to be solved by other regular methods. The Monte Carlo method has been used to solve many problems, such as the estimation of the value of π and the multiple integral problems. Because Monte Carlo method can be used to generate the entire population according to a prior probability distribution from samples, it is considered a proper method and has been applied by several researchers to synthesize population from sample microdata (Frick & Axhausen, 2004; Guo & Bhat, 2007; Ye et al., 2009; Farooq, Bierlaire, Hurtubia, & Flötteröd, 2013). Although several alternative methods are also used to generate synthetic population, the IPF algorithm is still the most extensively used algorithm because of its simplicity and computational efficiency. Therefore, in this dissertation, the IPF algorithm is chosen to synthesize population.

2.3.3 The Problems of the IPF Algorithm

The IPF algorithm has some problems that need to be resolved for population synthesis. The first problem is the occurrence of the zero-cells, and it has become a major source of error in population synthesis (Beckman et al., 1996; Guo & Bhat, 2007; Norman, 1999). The IPF algorithm needs to iteratively update cell values in a contingency table constrained by marginal totals, so all data are required to be cross-tabulated into a contingency table first. During the procedure of cross-tabulation, the zero-cells problem takes place. For example, suppose a sampled microdata has two variables: age and employment status. We further assume that the

age is divided into the following categories: 0-4 years old, 5-9 years old, 9-16 years old, 16-64 years old, and over 65 years old; and the employment status has two categories: employed and unemployed. Then, the sampled microdata can be cross-tabulated into a contingency table that contains the number of individuals with a specific variable category. In this contingency table, it is likely that the number of individuals who are 5-9 years old and are employed in the labor force is zero. The occurrence of zero-cells can lead to severe problems, such as the failure of convergence or even the runtime error to execute the IPF algorithm (Norman, 1999). In general, there are two methods to deal with the zero-cells. The first method is to introduce a very small positive number to replace the zeros. This has been used by Beckman et al. (1996) and Lomax and Norman (2016). The second method is to re-categorize the variables (Guo & Bhat, 2007). For example, if household income is the attribute to be used in the IPF algorithm, and some census tracts whose household income is lower than 10,000 dollars have zero households, re-categorizing the attribute (household income) with other categories (such as household income is between 10,001 to 20,000 dollars) can reduce the occurrence of the zero-cells. In this dissertation, these two methods will both be used to resolve the zero-cells problem.

Another challenge of the IPF-based synthetic population is that its computational intensity and complexity increases exponentially with the number of the attributes (Pritchard & Miller, 2012). It would be ideal but not computationally possible to generate a universal-purpose synthetic population that has all the characteristics of the real-world human beings. In this study, this challenge is dealt with by just incorporating the most relevant socio-demographic factors into population synthesis so that it reduces the computational intensity while keeping the crucial socio-demographic variables. More specifically, the synthetically generated population will only include those important socio-demographic characteristics relevant to travel activity patterns.

For instance, suppose there is a travel activity pattern that can be characterized by going to work in the morning, attending social recreations, and then returning to home in the evening; and further it is assumed that employment status, education level, and household incomes are the most relevant socio-demographic factors that influence this travel activity pattern, then we use the population synthesis to simulate a synthetic population with these relevant socio-demographic factors that fit a particular travel activity pattern, rather than creating the whole population with all socio-demographic variables.

2.3.4 Validation of the Synthetic Population

After the synthetic population is generated, it is necessary to evaluate the performance to see how accurate it is. This procedure is called validation of synthetic population. In essence, the validation is to measure the difference between the synthesized population and the true population. In general, there are two types of methods to validate the synthetic population: internal and external validation (Lovelace et al., 2015). The internal validation refers to the comparison between the aggregated synthetic population and the predefined attribute constraints; the external validation means the comparison between the synthetic population with other real micro-data that are available. Both of these two validations are important for population synthesis. However, due to the limited availability of external sources to validate the synthetic population, most of the validations are the internal validations. Therefore, in this research, we also examine the performance of the IPF algorithm by using the internal validation.

The internal validation is also referred to as the internal goodness-of-fit testing (Knudsen & Fotheringham, 1986; Pritchard & Miller, 2012). In previous work, the commonly utilized

goodness-of-fit indicators to verify the population synthesis include chi-square statistic, Root Mean Square Error (RMSE) (Simpson & Tranmer, 2005; Birkin, Turner, & Wu, 2006), Standardized Root Mean Square Error (SRMSE) (Pritchard & Miller, 2012), Total Absolute Error (TAE) (Wong, 1992), and coefficient of determination (Lovelace et al., 2015). For example, Simpson and Tranmer (2005) considered that the RMSE statistic is a good indicator to measure the overall accuracy, and they adopted it to verify the estimated population created by the IPF algorithm. The RMSE is defined as:

$$RMSE = \sqrt{\sum_i \sum_j (\hat{\theta}_{ij} - \theta_{ij})^2 / N} \qquad \text{(Equation 2.2)}$$

where the $\hat{\theta}_{ij}$ is the cell value in row i and column j of the contingency table from the simulated population, and θ_{ij} is the observed population, and N is the grand sum of the population. From Equation 2.2, we see that the RMSE indicates the standard deviation between the estimated and observed data. In addition, the variant of the RMSE, the SRMSE, is also commonly applied to evaluate the performance. For instance, Pritchard and Miller (2012) pointed out that the SMRSE is also a useful goodness-of-fit measure between simulated and observed population microdata, and it is actually equal to RMSE divided by its expected mean value.

As in another example, Wong (1992) presented a procedure of generating disaggregate level population data by the IPF algorithm and then evaluating its performance. Wong (1992) considered that among many goodness-of-fit measures that used to evaluate the performance of the IPF algorithm, the Total Absolute Error (TAE) is a reliable one. The TAE represents the sum of the absolute value of the difference between predicted value and observed value. The formula of TAE is listed as below:

$$TAE = \sum_i \sum_j |(\hat{\theta}_{ij} - \theta_{ij})| \qquad \text{(Equation 2.3)}$$

where the $\hat{\theta}_{ij}$ and θ_{ij} are the cell values in ith row and jth column of contingency tables from the simulated and observed population respectively.

In addition to the RMSE and TAE, Lovelace et al. (2015) considered the scatter plot a reasonable method to explore the overall goodness-of-fit at the early stage of validation. They believed that a scatter plot can be constructed for each geographic zone to see whether there is a linear relationship between aggregated synthesized population and attribute constraints. If the aggregated synthetic population is perfectly matched with the attribute constraints, all points will fall on the best fitting line. Lovelace et al. (2015) also used the coefficient of determination, R^2, to measure the goodness-of-fit. In statistics, the coefficient of determination is the square of the Pearson's correlation coefficient. The correlation coefficient is frequently used to describe the linear relationship between two numeric variables, and it is equal to the covariance of two variables divided by the product of the standard deviation of two variables. The range of coefficient of determination, R^2, is from 0 to 1. If R^2 is closer to 1, it indicates a good fitting between synthesized microdata and the observed data. In contrast, if the R^2 value is closer to 0, it means the overall goodness-of-fit is poor.

In this research, the scatter plot will be used first to plot the relationship between aggregate level estimated population against socio-demographic attribute constraints. Then, the Normalized RMSE is calculated to gauge the difference between the synthesized and the observed population microdata. Moreover, the probability distributions of the synthesized socio-demographic variables are plotted to see whether their distributions are similar as that of the observed variables. More details will be discussed in Chapter 3.2.7.

Chapter Three: Data and Methodology

3.1 Data Sources

Travel surveys have been utilized as main data sources for travel activity analysis. The travel survey data should at least consist of starting and ending time, activity duration, activity location (Anderson, 1971; Thornton, Williams, & Shaw, 1997). Non-spatial attributes should include individual characteristics such as activity purpose and transportation mode, while location information mainly consists of residential locations and activity locations (Anderson, 1971).

The National Household Travel Survey (NHTS) in the year of 2001 is well recognized as the first comprehensive nationwide household travel survey database about people's travel behaviors (Cohen, Moore, Sharp, & Giesbrecht, 2003; Pucher & Renne, 2004). It has been used in many travel activity pattern studies, such as those conducted by Bowman et al. (1998), Giuliano (2003), Collia et al. (2003), Pucher and Renne (2003), Pucher and Renne (2004), Tal and Handy (2010), and Pucher et al. (2011). The 2001 NHTS gathers and merges the information from the Nationwide Personal Transportation Survey (NPTS) and American Travel Survey (ATS). The NPTS is the predecessor of the 2001 NHTS, and it was conducted from the 1960's to 1995 (in 1969, 1977, 1983, 1990, and 1995). The ATS is a program financially supported by the Bureau of Transportation Statistics (BTS) with an aim to provide detailed data on long distance trips that are longer than or equal to 100 miles. The ATS is conducted every five years, and the most recent one is the 1995 ATS. The 2001 NHTS changes the definition for long distance trips to the ones of a minimum distance of 50 miles. The NHTS provides complete information for all-purpose trips, and it contains information about demographic characteristics such as income,

gender, race, and household lifecycle (Cohen, Moore, Sharp, & Giesbrecht, 2003). It also provides the trip-related data, for instance, daily trip number, travel mode, trip purpose, and trip time in minutes (Pucher & Renne, 2003).

The 2001 NHTS contains 69,817 records of household information, 160,758 records of personal characteristics, 139,382 rows of vehicles data, and 642,292 records of daily trips. Since the majority of travel modes are personal vehicles (87.24%), 3,000 individuals who use personal vehicles as the travel mode are randomly sampled for analyzing their activity patterns. Besides the transportation mode, some other important transportation attributes are extracted from the 2001 NHTS to construct a subset for the study. They include the household ID, person ID, trip number, starting and ending time, trip miles, activity type, and activity sequence.

The other travel survey data used in this study is the 2009 NHTS. It contains four files, which are household file (150,147 records), person file (308,901 records), vehicle file (309,163 records), and day trip file (1,167,321 records). Just like the 2001 NHTS, since the personal vehicle accounts for the majority of travel modes (87.48%), 3,000 individuals who use person vehicles are sampled from the whole data set. Those important transportation attributes used in the 2001 NHTS, like household ID, person ID, trip number, starting and ending time, trip miles, activity type, and activity sequence are also extracted from the 2009 NHTS for analysis.

3.2 Methodology

3.2.1 Representation of Human Travel Activities

In this study, two recent NHTS, the 2001 and 2009 NHTS databases are used to explore the human travel activity patterns. Table 3.1 below is an example of an individual's daily activity

record in the 2001 NHTS. It contains the information such as the trip number (TRIPNUM), trip starting time (STRTTIME), trip ending time (ENDTIME), travel distance in miles (TRPMILES), why travel from (WHYFROM), why travel to (WHYTO), and trip purpose summary (WHYTRP1S). It is worth noting that, in the 2001 and 2009 NHTS, the starting and ending time of trips are all stored in the format of military time. For example, in Table 3.1, the individual's starting time for the first trip is 830, and the ending time for the last trip is 1930. It indicates the individual starts his daily activity from 8:30 AM in the morning and returns to home at 7:30 PM in the evening.

Table 3.1 An example of a person's daily activity record in the 2001 NHTS

TRIPNUM	STRTTIME	ENDTIME	TRPMILES	WHYFROM	WHYTO	WHYTRP1S
1	830	845	0.556	01	82	09
2	930	945	0.556	82	01	12
3	1800	1802	1	01	82	09
4	1845	1910	5	82	40	07
5	1910	1930	5	40	01	12

The travel distance in miles (TRPMILES) in Table 3.1 is the travel distance for the particular number of trip, not cumulative distance from home. In order to represent each individual's activity as a pattern vector chained by travel, the travel distance in miles (TRPMILES) in Table 3.1 is transformed into the distance to home, which is the starting point of the trip. The distance from the activity location to an individual's home is also used in previous work, such as Pas (1988). In this study, the distance to home is calculated according to the following steps: (1) if an individual starts the trip from home, the distance to home after completing this trip is equal to the trip miles traveled from home; (2) if an individual is returning to the home, the distance to home after finishing the trip is assigned as zero; (3) if an activity is followed by a trip returning

to home, the distance to home is trip miles traveled from activity sites to home; and (4) otherwise, the cumulative distance is used. Figure 3.1 maps the whole procedure.

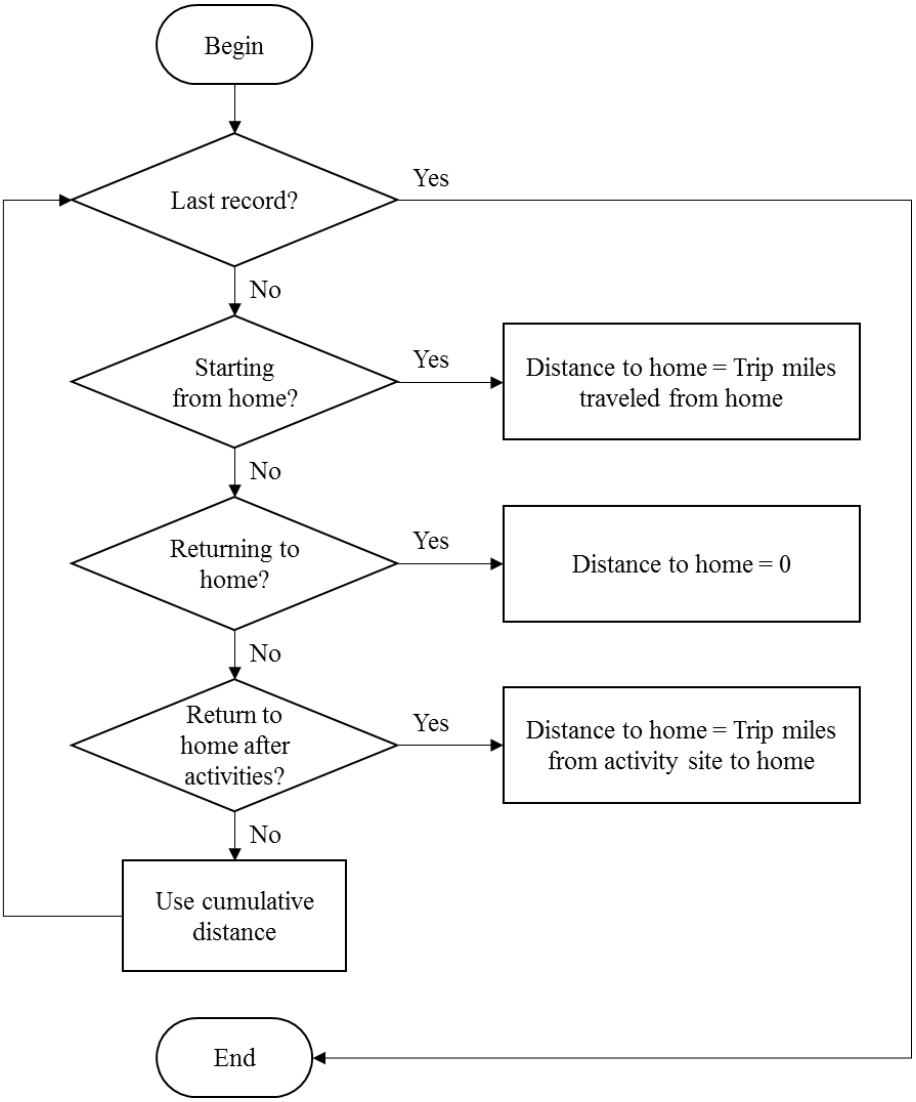

Figure 3.1 The procedure of calculating the distance to home

Upon the completion of the conversion, the distance represents the distance to home. Afterwards, the individual's daily activity records can be represented as activity pattern vectors as shown in Figure 3.2 below.

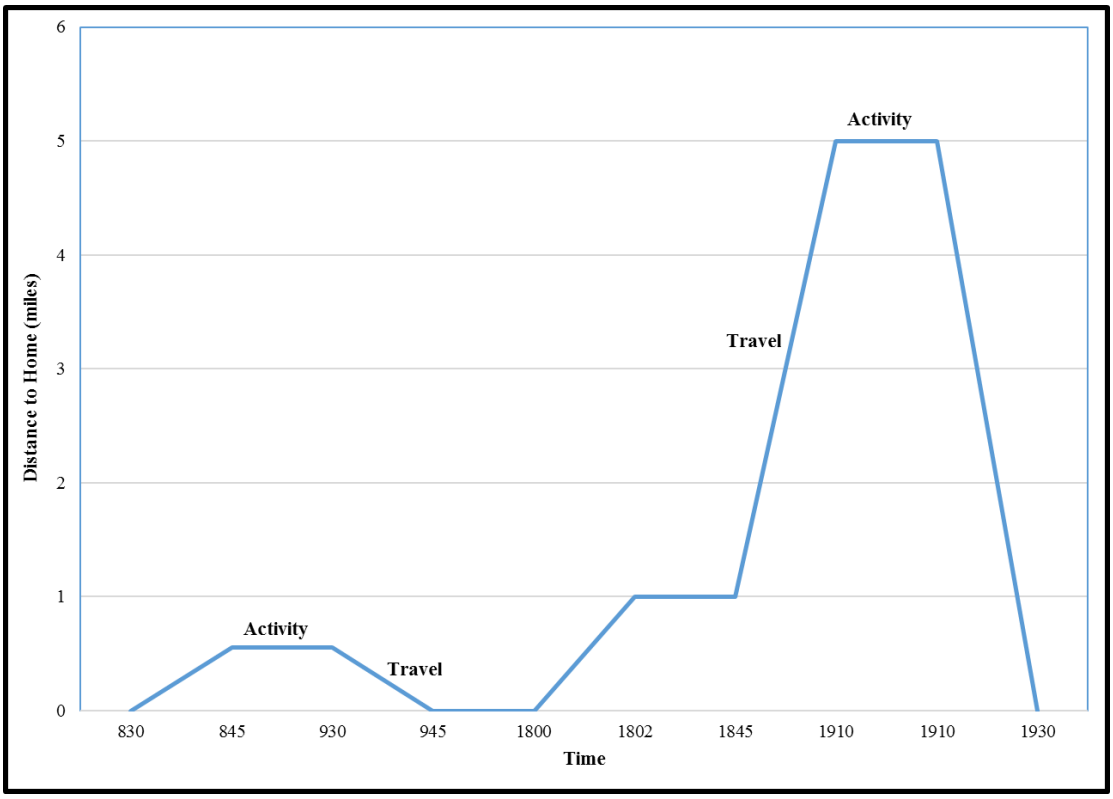

Figure 3.2 An example of activity pattern vector

In Figure 3.2, any line segment with the zero slope stands for activity participation because the distance to home remains constant during that time period. The line segments with non-zero slopes are considered as travel behaviors. By using this notation, an individual's activity pattern can be represented as a chain of activities in time sequence, and all activities are linked by travel

behaviors. In addition to the distance to home, other attributes, such as activity time, activity duration, trip purpose, and activity sequence, are all included in the activity pattern vector.

3.2.2 Dissimilarity in Travel Activities

An individual's daily travel activities have many attributes. Some important travel and activity attributes can be used to compare whether individuals have similar travel activity patterns or not, such as starting and ending time, activity type, travel distance, travel mode, and activity sequence. In this study, two attributes, i.e. the distance to home and activity sequence are chosen to calculate the dissimilarity/similarity of individuals' travel activities. This section discusses the method to calculate the dissimilarity in distance to home. Because there are hardly any two persons having exactly the same daily travel activities, pattern vectors representing individuals' daily travel activities may have different lengths, or different trip starting and ending time. Therefore, a linear interpolation of distance according to time is needed before such comparison can be performed.

It is critical to determine an appropriate time interval before performing the linear interpolation. A large time interval could miss information; on the other hand, if the daily travel activities are sliced minute by minute, the dataset will become too huge to process. Figure 3.3 provides the distribution of time intervals of travel and activities in the 2001 and 2009 respectively. In Figure 3.3, the red solid line represents the time interval distribution in the 2001 NHTS, and the blue solid line is for 2009 NHTS. It is observed that their distributions are very similar, and most of the time intervals are between 5 to 10 minutes. Hence, in this study, the time interval is set to 5

minutes which is fine enough to catch important travel and activity details while the dataset is still manageable.

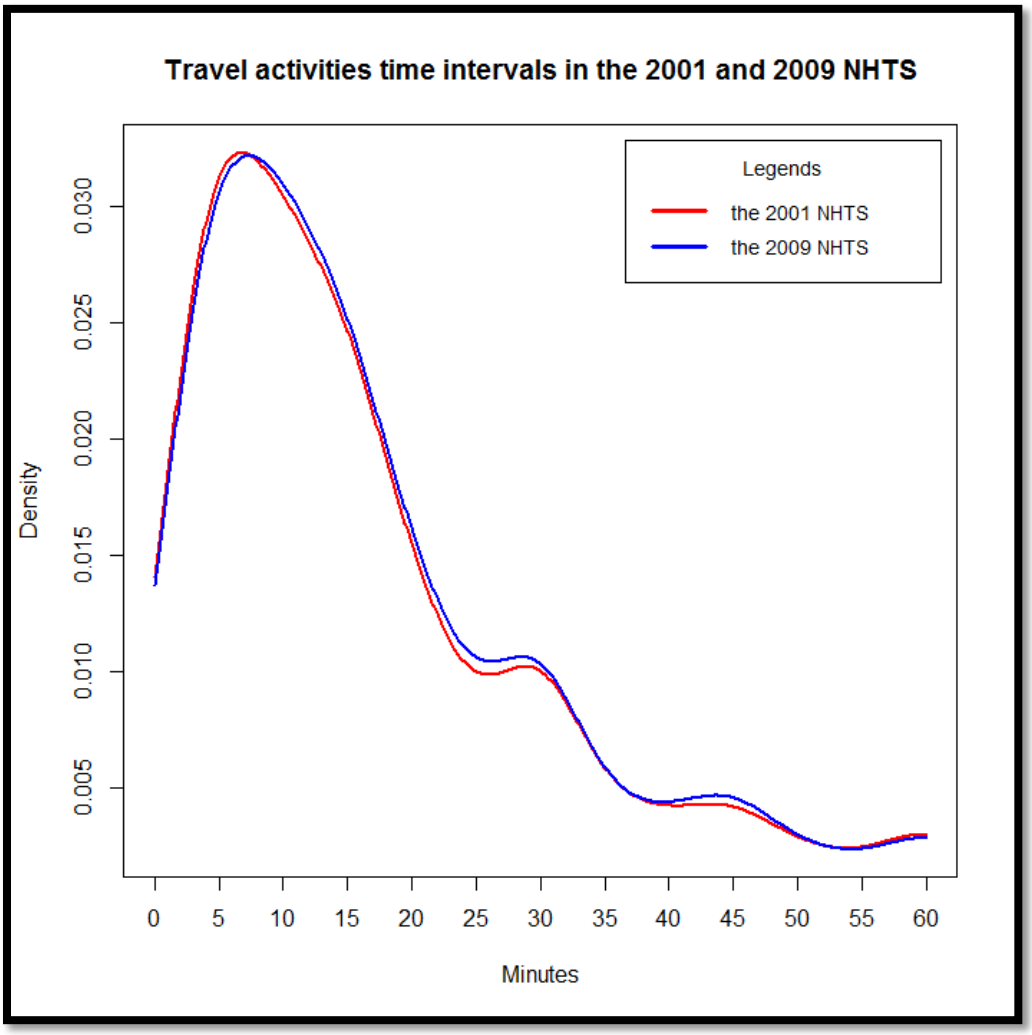

Figure 3.3 Travel activities time intervals in the 2001 and 2009 NHTS

In the linear interpolation, a determination about what time the interpolation starts is also important. Figure 3.4 provides the descriptive analysis for the trip starting and ending time in a day in the 2001 and 2009 NHTS. In Figure 3.4, we can see that most of the individuals start their daily trips between 7:00 AM and 8:00 AM in the morning, and finish their last trips

between 5:00 PM and 6:00 PM in the afternoon. More importantly, we find that it starts from 3:00 AM that increasingly more people begin their daily trips. As such, the linear interpolation starts at 3:00 AM in this study. In order to ensure the consistency of the activity pattern vector, any military time which is less than 300 are transformed by adding 2400. For instance, 0:50 AM is converted into 2450. In addition, before the linear interpolation, the distance values at 3:00 AM in the morning are calculated first. There are three possible values for the distance to home at 3:00 AM: (a) if an individual returns to home before 3:00 AM and has no outside activities, the distance to home at 3:00 AM is set to zero; (b) if an individual is still outside and is driving home, the linear interpolation is applied to calculate the distance to home at 3:00 AM; and (c) if an individual is still outside and is participating in an activity, the distance from the activity site to the home is used as the distance to home at 3:00 AM.

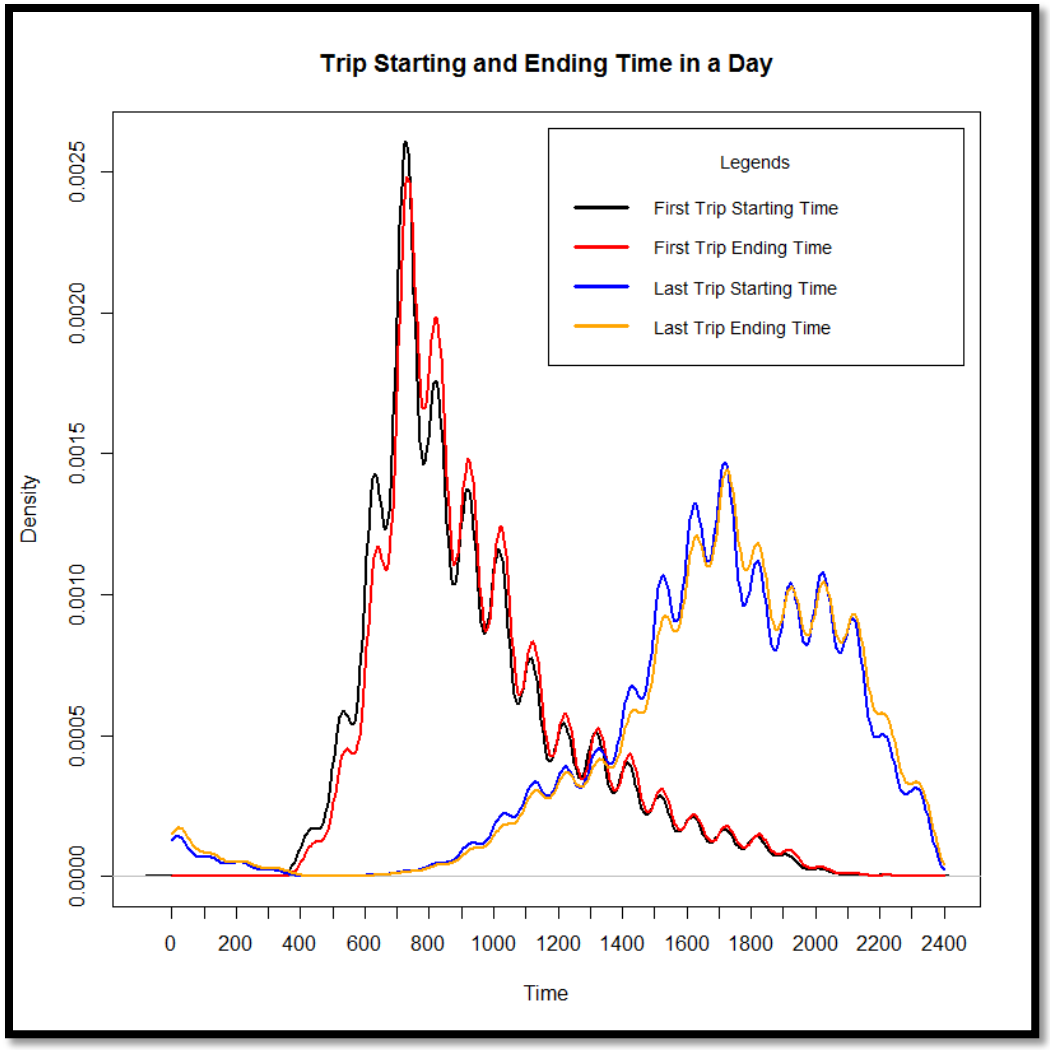

Figure 3.4 The distribution of trip starting time and trip ending time

Military time format in the dataset needs to be treated carefully because of its unique property. For example, an arithmetic time sequence from 8:50 AM to 9:10 AM with 5-minute intervals is (850, 855, 900, 905 and 910). Apparently, the difference between 855 and 900 is 45 in military format, instead of 5 (minutes) as we expected. Hence, data transformation process is necessary when time is used in the interpolation. The simplest solution, which is also the method adopted in this research, is converting the military time into minutes elapsed since 12:00 AM. By this

method, the arithmetic time sequence (850, 855, 900, 905, 910) will be converted into (530, 535, 540, 545, 550), and thus the actual time difference is 5 minutes, rather than 45 minutes.

After the linear interpolation, a distance to home matrix is available in which each row represents an individual's daily travel activities in the form of vectors with the same length of 288 elements (24 hours * 60 (minutes/hour) / 5 (minutes/interval) = 288 intervals). The row-wise Euclidean distance is applied to calculate the dissimilarity between any two vectors. For vectors $A = (A_1, A_2, \dots, A_{n-1}, A_n)$ and $B = (B_1, B_2, \dots, B_{n-1}, B_n)$, the Euclidean distance d between vectors A and B is $d = \sqrt{\sum_{i=1}^{n}(A_i - B_i)^2}$. As a result, a symmetric dissimilarity matrix is constructed in which the values in the major diagonal line are all zero and the other values represent the Euclidean distance between individuals' distance-to-home vectors.

3.2.3 Dissimilarity in Activity Sequence

In this research, the dissimilarity between individuals' travel activities are based on their activity sequences. To conduct the activity sequence analysis, the trip purposes for all individuals are summarized and encoded by English capital letters. Table 3.2 and Table 3.3 show the summarized trip purposes and their corresponding encoded letters in the 2001 and 2009 NHTS respectively.

Table 3.2 The summarized trip purposes in the 2001 NHTS

Value	Summarized Trip Purpose	Encoded Letter
01	To work	A
02	Work-related	A
03	School	B
04	Religious	B
05	Medical/dental	C
06	Shopping	D
07	Other family & personal	F
08	Social recreation	E
09	Eat meals	I
10	Serve passenger	G
11	Return to work	A
12	Home	H
13	Others	J

Table 3.3 The summarized trip purposes in the 2009 NHTS

Value	Summarized Trip Purpose	Encoded Letter
01	Home	H
10	Work	A
20	School / Religious activity	B
30	Medical services	C
40	Shopping/Errands	D
50	Social recreation	E
60	Family personal	F
70	Transport someone	G
80	Meals	I
97	Others	J

As we can see in Table 3.2 and Table 3.3, there are 13 types of summarized trip purposes in the 2001 NHTS; but in the 2009 NHTS, there are only 10 types of summarized trip purposes. In order to make the two datasets compatible, some categories of trip purposes in the 2001 NHTS

have to be generalized and then combined according to the 2009 NHTS. For example, in the 2001 NHTS, to work (Value=01), work-related (Value=02), and return to work (Value=11) are all encoded with the letter of 'A'. School (Value=03) and Religious (Value=04) are combined together and encoded with the letter of 'B'. Apart from the above-mentioned trip purposes, the travel behavior is abbreviated and encoded as "T" for convenience. Therefore, the person's daily activity sequence in Table 3.1 can be represented as a sequence of letters: T → I → T → H → T → I → T → F → T → H. Like the distance to home, the interpolation is also implemented for the activity sequence for every 5 minutes within 24 hours. But, for the activity sequences, a constant interpolation is applied, which standardizes the activity sequences for comparison purpose.

After the matrix of activity sequences has been constructed, the sequence analysis can be performed. A widely accepted method to conduct the sequence analysis is the global pairwise sequence alignment developed by Needleman and Wunsch (1970). In the following section regarding activity sequence analysis, the Needleman-Wunsch's global pairwise sequence alignment is applied to align any two pairs of activity sequences.

The sequence alignment method is based on a scoring scheme to indicate the similarity/dissimilarity between two sequences. In this study, the Levenshtein edit distance which was developed by Vladimir Levenshtein in 1965 is applied to establish the scoring scheme (Ding, 2009; Joh, Arentze, & Timmermans, 2001). The Levenshtein edit distance is based on the cost of three types of character operations: insert, delete, and substitute. The edit cost is one if one gap is inserted or deleted, and the cost of the substitution operation is one if the two sequences do not match. The Levenshtein edit distance can be formulated in the Equation 3.1.

$$cost = \begin{cases} insert\ gap: & cost = 1 \\ delete\ gap: & cost = 1 \\ substitution: & match = 0;\ \ unmatch = 1 \end{cases} \qquad \text{(Equation 3.1)}$$

By applying the Levenshtein edit distance definition, a symmetric substitution matrix containing the capital letters A to Z like Table 3.4 can be constructed for the subsequent pairwise activity sequence alignment.

Table 3.4 Substitution matrix by Levenshtein edit distance

	A	B	C	…	Z
A	0	1	1	1	1
B	1	0	1	1	1
C	1	1	0	1	1
…	1	1	1	0	1
Z	1	1	1	1	0

Based on the substitution matrix, the Needleman-Wunsch's global pairwise sequence alignment can be carried out as follows (Needleman & Wunsch, 1970). Suppose there are two sequences A and B with the length of m and n respectively. The first step is to initialize a matrix with (m + 1) rows and (n + 1) columns. Starting with a value of zero in the top-left cell, and the value increases by one in the first row and first column according to the gap cost defined by the Levenshtein edit distance. For example, if A = "TATH" and B = "TJJTH", then the initialization matrix can be constructed like in Table 3.5.

Table 3.5 Initialization matrix for activity sequence alignment

	-	T	J	J	T	H
-	0	1	2	3	4	5
T	1					
A	2					
T	3					
H	4					

After the initialization, the next step is to fill out the remaining values in the matrix as illustrated in Table 3.6. Starting from the cell in the second row and second column, the value of a cell is derived from the minimum value among its neighbors from three directions (diagonal, left, and top) as illustrated in Table 3.6. The diagonal number is added by its corresponding value with the same subscripts in the substitution matrix, but the numbers on the left and top are added by the gap cost defined by the Levenshtein edit distance.

Table 3.6 Filling out the matrix by Levenshtein edit distance

	-	T	J	J	T	H
-	0	1	2	3	4	5
T	1	0	1	2	3	4
A	2	1	1			
T	3	2				
H	4	3				

After filling out the entire matrix, the trackback procedure can begin from the bottom-right corner of the matrix as presented in Table 3.7. It is worth noting that no matter what kind of pathway is employed, the cost between these two sequences remains unchanged. In this case, the

difference between activity sequences A and B is two. With the Needleman-Wunsch's global sequence alignment method described above, any two sequences can be compared and the dissimilarity matrix can be formed.

Table 3.7 Trackback pathway

	-	T	J	J	T	H
-	0	1	2	3	4	5
T	1	0	1	2	3	4
A	2	1	1	2	3	4
T	3	2	2	2	2	3
H	4	3	3	3	3	2

3.2.4 Cluster Analysis

The basic principle of cluster analysis is to divide a set of data into several clusters, so that observations within one cluster are similar and observations between clusters are different (Jung, Park, Du, & Drake, 2003; Fraley & Raftery, 1998). Among various clustering algorithms, *k*-means clustering and hierarchical clustering are the two basic methods. The term "*k*-means" was first used by MacQueen (1967), and its formal algorithm was first published by Lloyd (1982). The *k*-means clustering can partition the dataset into *k* different clusters so that the within cluster sum of squares (WCSS) is minimized (MacQueen, 1967; Lloyd, 1982). The WCSS is the total sum of squared deviation of each observation to its corresponding clustering center, and its formula is defined as follows:

$$WCSS = \sum_{i=1}^{k} \sum_{\boldsymbol{X} \in C_i} ||\boldsymbol{X} - C_i||^2 \qquad \text{(Equation 3.2)}$$

where k denotes the number of clusters, C_i means the ith clustering center.

The *k*-means algorithm starts from randomly generating *k* (*k* is usually specified by users) clustering centers as the initial seeds, and then it iteratively goes through the assignment step (assigning each observation to its closest center) and the update step (recalculating the new clustering centers after assignment step) until it reaches the convergence (Wagstaff, Cardie, Rogers, & Schrödl, 2001). Here, the convergence is referred to the situation in which the assignments of all observations will not change by carrying out one more iteration. Today, the *k*-means algorithm is widely used, and is considered one of the most important methods in clustering analysis because of its simplicity and efficiency (Zhu, Wang, & Li, 2010; Ng, 2000).

Hierarchical clustering is another approach in the family of clustering methods. The hierarchical clustering is a kind of deterministic algorithm that builds a tree-like hierarchy structure to partition data into clusters. For hierarchical clustering, although it is not necessary to specify the number of clusters as the *k*-means does, it requires a dissimilarity matrix which represents the overall difference between observations. The construction of the dissimilarity matrix is based on a specific distance metric. The selection of the distance metric is important, because different distance metrics will lead to different clustering results. In most cases, the distance metric used in the hierarchical clustering is the Euclidean distance. After the dissimilarity matrix is constructed, observations are progressively merged based on certain linkage criteria until all points are in a single cluster. The popular choices of linkage criteria include single linkage (minimum distance), complete linkage (maximum distance), Ward's method (minimum variance) etc. In this research, the Ward's method is used because of it is most commonly used.

In this dissertation, the *k*-means and hierarchical cluster analysis methods will both be applied to partition data into different human activity patterns. The *k*-means clustering is employed first to determine the number of clusters to be partitioned. The WCSS is plotted by the number of

clusters first, and then we use the elbow method proposed by Robert Thorndike in 1953 to determine the appropriate number of clusters. Specifically, the WCSS drops sharply with the increase of the cluster numbers at the beginning, and then the dropping rate slows down. The number of clusters is chosen at the "elbow" point such that increasing one more number of clusters will not greatly change the WCSS. After the number of clusters is determined, the hierarchical clustering analysis is conducted. The hierarchical clustering requires a dissimilarity matrix that represents the overall difference between observations, so two dissimilarity matrices, computed based on distance and activity sequence respectively, are integrated to get an overall dissimilarity matrix. Before integration, a normalization procedure is necessary. What is worth to mention is that there are two criteria that need to be met simultaneously for the normalization: (1) these two matrices should be symmetric after normalization process, and (2) no negative values are allowed in both matrices. To meet the above constraints, feature scaling is adopted. Feature scaling is a type of statistical normalization method that brings all values into the range of [0, 1] by this transformation: $X' = (X - min)/(max - min)$. Specifically, values in the upper and lower triangular matrices are all normalized by using this transformation. The normalization is applied to the whole matrix, so that it keeps the two dissimilarity matrices symmetric. Two normalized matrices are then merged together by a simple mathematical addition operation.

The whole procedure regarding travel activity analysis is presented as a flow chart in Figure 3.5. In summary, an individual's daily travel and activities is represented as a sequence of activities linked by travel behaviors. Distance to home (converted from the travel distance) and activity sequence are interpolated at 5-minute interval first, and then Euclidean distance and pairwise sequence alignment are applied to develop dissimilarity matrices separately. Then, two

dissimilarity matrices are normalized and integrated into a dissimilarity matrix that represents the overall distance between observations. After that, the hierarchical clustering analysis is conducted based on the integrated dissimilarity matrix to partition data into several travel activity patterns.

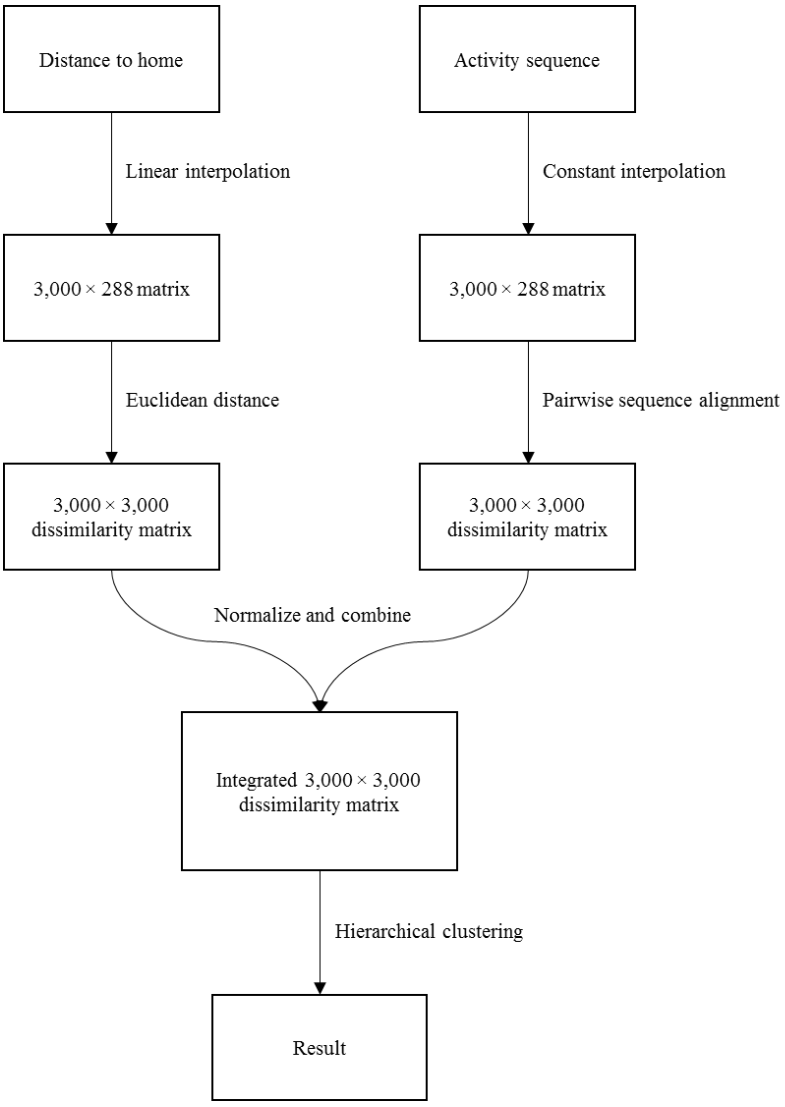

Figure 3.5 The flow chart on travel activity analysis

3.2.5 Socio-Demographic Correlates of the Major Travel Activity Patterns

Individuals' travel activity patterns are correlated with their socio-demographic characteristics (Joh, Arentze, Hofman, & Timmermans, 2002). The random forest classification model is employed to examine their relationship to social-demographic characteristics. To introduce the random forest model, the decision tree learning is described first.

Decision tree learning is one of the most popular prediction models. It creates a tree-like diagram from top to down that partitions the data into subsets containing observations with similar properties. During the procedure of constructing a decision tree, it keeps selecting the best explanatory variable to split the data at each step until the tree stops splitting. Different decision tree learning algorithms use different metrics to determine the best explanatory variable to split the tree, and three of them are introduced here. They are ID3, C4.5, and Classification and Regression Trees (CART). The ID3 was first proposed by Ross Quinlan (1986), and it was one of the earliest algorithms to implement the decision tree learning. In ID3, the metric that is used to split data is based on the information gain, which equals to the change of the entropy. The tree node is split by using the explanatory variable which maximizes the information gain. The C4.5 algorithm was also developed by Ross Quinlan (1993), and it was considered as the successor of the ID3 method. But, the metric to select attribute to split the data is the information gain ratio (also known as the normalized information gain). The CART was first proposed by Breiman et al. (1984), and it is another famous algorithm that implements the decision tree learning. Unlike the ID3 and C4.5 methods in which they use information gain as the metric during the procedure of constructing a decision tree, the CART algorithm calculates Gini impurity to decide how to split the dataset (Lewis, 2000). To be specific, the explanatory variable which minimizes the Gini impurity is chosen to split the tree node. The decision tree

learning keeps selecting the best variable to split the tree until the tree stops splitting. In most cases, the decision tree learning stops splitting under one of the following conditions: (1) no more explanatory variables left to split; (2) all data in the given child node are pure; that is to say, the values of the dependent variable in the given child node are the same (Lewis, 2000).

According to the type of the dependent variable, the decision tree can be divided into classification tree and regression tree, where the dependent variable is categorical data and continuous data respectively. In this research, the dependent variable, travel activity patterns, is categorical data, thus the classification tree is considered an appropriate method to explore the relationship between travel activity patterns and the socio-demographic variables. The decision tree learning has been widely used because of its advantages. For example, the decision tree is a procedure that is easy to understand and explain. It does not require the user to learn a lot of background knowledge about the learning process to understand the features of the data. Moreover, the decision tree learning is a non-parametric method. So, it is unnecessary to make the prior assumptions of the distribution of variables as other statistical methods do (Prasad, Iverson, & Liaw, 2006). However, some limitations of the decision tree learning should not be neglected. For example, the overfitting problem and low prediction accuracy were the two concerns (Lewis, 2000). The consequence of the overfitting problem is bad, because it might lead to a poor predictive performance.

To overcome the limitations of decision tree learning method, Breiman (2001) proposed the random forest model. The random forest model is a kind of ensemble learning classifier, and it is considered as the enhanced version of the decision tree learning method. As the name suggests, instead of just generating a single decision tree, the random forest algorithm trains a "forest" that includes a large number of classification or regression trees from bootstrap samples,

and let those trees "vote" for the best solution (Breiman, 2001). Compared with the decision tree learning method, the random forest model has many advantages. Firstly, many researchers believed that the random forest model is an optimal classifier, not only because it improves the predictive accuracy, but also it overcomes the overfitting problem in the decision tree learning method (Liaw & Wiener, 2002). Secondly, another advantage of using the random forest model is that it can provide the ranking of the importance of the explanatory variables (Svetnik et al., 2003). In this research, it needs to find out the prominent variables that contribute to the distinct travel activity patterns, and therefore, the random forest is deemed as an appropriate method. Thirdly, the random forest model is a robust classifier, because its performance is better than many other classification algorithms, such as decision tree learning, Linear Discriminant Analysis (LDA), Multinomial Logistic Regression (MNL), and Support Vector Machine (SVM) (Liaw & Wiener, 2002; Svetnik et al., 2003). For these reasons, the random forest model is used in this dissertation to analyze the relationship between socio-demographic factors and major travel activity patterns, and to find out the prominent socio-demographic variables that contribute to individuals' different travel activity patterns. The explanatory variables that are entered into the random forest model come from the household, person, and vehicle files in the 2001 and 2009 NHTS datasets respectively. To be specific, in the person file, the age, gender, race, employment status, and education level are extracted for the random forest model. Some of the household socio-demographic factors including household income, household size, and household life cycle are subset from the NHTS database. The rest of the socio-demographic variables come from the vehicle file, which emphasizes more the travel characteristics of a household and a person, such as travel day in a week, the number of vehicles per household, and the number of drivers.

3.2.6 Population Synthesis with the Prominent Socio-Demographic Characteristics

After examining the prominent socio-demographic characteristics, the procedure of population synthesis is carried out to synthesize population microdata for travel activity analysis. The methodology provided here is an IPF-based approach to generate a complete microdata of all population. The IPF algorithm is a type of iterative method. An iterative method is a process that sets an initial value, and then iteratively refines the solution until it meets the convergence criteria. This is how the IPF works: assume that there is a two-way contingency table, and in the table, ${p_{ij}}^{(k)}$ is the element at the ith row, jth column, after kth iteration, and Q_{i+} and Q_{+j} are the known marginal totals for rows and columns. The next iteration is computed according to the following equations:

$${p_{ij}}^{(k+1,\ a)} = \frac{{p_{ij}}^{(k)}}{\sum_j {p_{ij}}^{(k)}} * Q_{i+} \qquad \text{(Equation 3.3)}$$

$${p_{ij}}^{(k+1,\ b)} = \frac{{p_{ij}}^{(k)}}{\sum_i {p_{ij}}^{(k)}} * Q_{+j} \qquad \text{(Equation 3.4)}$$

where the ${p_{ij}}^{(k+1,\ a)}$ is the row operation and the ${p_{ij}}^{(k+1,\ b)}$ involves the column operation.

As illustrated in Equation 3.3 and Equation 3.4, the IPF algorithm iteratively updates the contingency table constrained by the predefined Q_{i+} and Q_{+j} until it reaches the convergence. There are two conditions under which the IPF algorithm stops: (1) the IPF algorithm converges. This happens when after $k+1$ iterations, the differences of all cell values between iteration k and $k+1$ are less than the predefined tolerance value. In other words, $|{p_{ij}}^{(k)} - {p_{ij}}^{(k+1)}| < \delta$, where δ denotes the predefined tolerance. The tolerance is usually a very small positive number, and it is set to 10^{-6} in this study. (2) the number of iterations exceeds the maximum number of iterations, which is set to be 1,000 in this study. In other words, if the IPF routine is carried out

for more than 1,000 times and the convergence criteria have not been met, the IPF algorithm will stop.

Figure 3.6 shows the procedure of how the IPF algorithm works in detail. The IPF algorithm requires seed data, which provide a sample size of population data with socio-demographic attributes, and target data, which only contain the total count of population with socio-demographic attributes, but the cell values for the target data at the beginning are unknown. In Figure 3.6, the contingency table on the upper-left quadrant is the cross-classified seed data, and it contains the number of population with the socio-demographic attributes of race and gender. The value in each cell represents the number of population with the combined socio-demographic attributes. For example, in Figure 3.6, the seed data shows that the number of people who are white and female is 63. The contingency table on the upper-right quadrant is the target data, which only provides the marginal total constraints, but the specific cell values are unknown. The initial solution for the cell values can be provided by using the sample data, but their row and column sums do not agree with the marginal totals now. In the Iteration 1(a), the IPF routine is applied to update the cell values to make the row sums equal to the row marginal totals as illustrated in Equation 3.3. However, after the Iteration 1(a), we find that the column sums are not equal to the column marginal totals. In the Iteration 1(b), the IPF algorithm scales the cell values to make the column sums equal to the column marginal totals as illustrated in Equation 3.4. After that, the row sums might not be equal to the marginal row totals this time. Upon the completion of the Iteration 1(a) and the Iteration 1(b), the 1st round of iteration is finished. Then, the second round of iteration of IPF algorithm starts. Similarly, we find only the row sums fit the marginal row totals in the Iteration 2(a), and only the column sums fit the marginal column totals in the Iteration 2(b). But, it is obviously to note that, compared with the

1st round of iteration, the differences between the estimated values and observed marginal total constraints after the 2nd round of iteration are reduced. For example, after the Iteration 1(a), the total number of white people is 2144.89. But after the Iteration 2(a), the total number of white people has been updated to 2198.39, which is closer to the observed marginal total constraint, 2205. Then, the IPF algorithm continues to update the cell values in the contingency table until it meets the convergence criteria. That is to say, when the difference in cell values between two consecutive iterations does not exceed the specified tolerance value, the IPF algorithm converges and the routine stops.

In Figure 3.6, the contingency table on the bottom-left side is the result after implementing the IPF algorithm. Eventually, the IPF algorithm converges after a number of iterations, and the total number of the synthesized population who are white and female is turned to be 1256.28 in the final target data.

Seed Data

	White	Non-white	Total
Male	45	108	153
Female	63	37	100
Total	108	145	**253**

Target Data

	White	Non-white	Total
Male	?	?	3105
Female	?	?	1955
Total	2205	2855	**5060**

Iteration 1(a)

	White	Non-white	Total
Male	913.24	2191.76	3105
Female	1231.65	723.35	1955
Total	2144.89	2915.11	**5060**

Iteration 1(b)

	White	Non-white	Total
Male	938.83	2146.57	3085.4
Female	1266.17	708.43	1974.6
Total	2205	2855	**5060**

Iteration 2(a)

	White	Non-white	Total
Male	944.80	2160.20	3105
Female	1253.60	701.40	1955
Total	2198.39	2861.61	**5060**

Iteration 2(b)

	White	Non-white	Total
Male	947.63	2155.22	3102.85
Female	1257.37	699.78	1957.15
Total	2205	2855	**5060**

Final Target Data

	White	Non-white	Total
Male	948.72	2156.28	3105
Female	1256.28	698.72	1955
Total	2205	2855	**5060**

Figure 3.6 The interaction procedure of IPF algorithm

3.2.6.1 Seed Data

To conduct the population synthesis by the IPF algorithm, as illustrated in Figure 3.6, two types of data are required; one is seed data, and the other is target data (Guo & Bhat, 2007). The seed data usually use the detailed sample survey, such as PUMS data in the United States. The PUMS data are sampled microdata that provide socio-demographic information about households and the associated individuals (Stopher, Greaves, & Bullock, 2003), so they are involved with two levels: one is at the household level, and the other is at the individual level. The socio-demographic information includes age, gender, household incomes, education level, employment status, etc., but the respondents' personal information (such as names and locations) is removed to protect the privacy (Stopher et al., 2003; Mohammadian, Javanmardi, & Zhang, 2010). It is worth stating that, after the year of 2012, the PUMS data change its geography boundary and code to match with the census data. Therefore, the seed data used in this research is the 2013 5-year PUMS data. According to the 2013 5-year PUMS data, there are 5,940 records in the household-level data, and 12,639 records in the individual-level data for Milwaukee County in Wisconsin.

3.2.6.2 Target Data

The target data are from the census data, and they are cross-classified into a multi-way contingency table with marginal totals. The target data provide the marginal totals of an area (e.g., a census tract), such as the total number of the female, or the total count of households with incomes ranging from $40,000 to $50,000. Before the year of 2000, the target data usually use data from the Summary File 3 (SF3) provided by the US Census Bureau. The SF3 is no longer available after 2000 census. Alternatively, US Census Bureau offers the American Community

Survey (ACS) Summary File to replace the SF3. In this research, the target data used are from the 2013 5-year ACS Summary File at the census tract level. Based on the 2013 ACS 5-year summary files, there are 379,637 households and 950,527 individuals within the 298 census tracts in the Milwaukee County in Wisconsin.

3.2.6.3 The 2010 State-based PUMA Equivalency File

The seed data and target data are collected at two different levels, i.e., Public Use Micro Area (PUMA) and census tract. The PUMA is the geographic unit to report the PUMS data. Usually, a single PUMA is composed of a number of census tracts, and it approximately has a population of 100,000 (Wheaton et al., 2009). Thus, it needs to establish a crosswalk linkage that connects PUMAs with census tracts. The 2010 state-based PUMA equivalency file includes the geographic relationship between PUMAs and census tracts, and it is used in this research. For the Milwaukee County in Wisconsin, there are eight PUMAs that correspond to the 298 census tracts. When examining the data, two empty census tracts (ID='980000' and '990000') are found in the 40101 and 40301 PUMA zones, and they are removed from the data. Table 3.8 shows the PUMAs and their corresponding census tracts for the Milwaukee County in Wisconsin. Figure 3.7 shows the eight PUMAs in the Milwaukee County.

Table 3.8 The 2010 equivalency file for the Milwaukee County in Wisconsin

PUMA	No. of Census Tracts	No. of Census Tracts After Removing Empty Zones
40101	33	32
40301	26	25
40701	46	46
41001	30	30
41002	41	41
41003	51	51
41004	35	35
41005	36	36

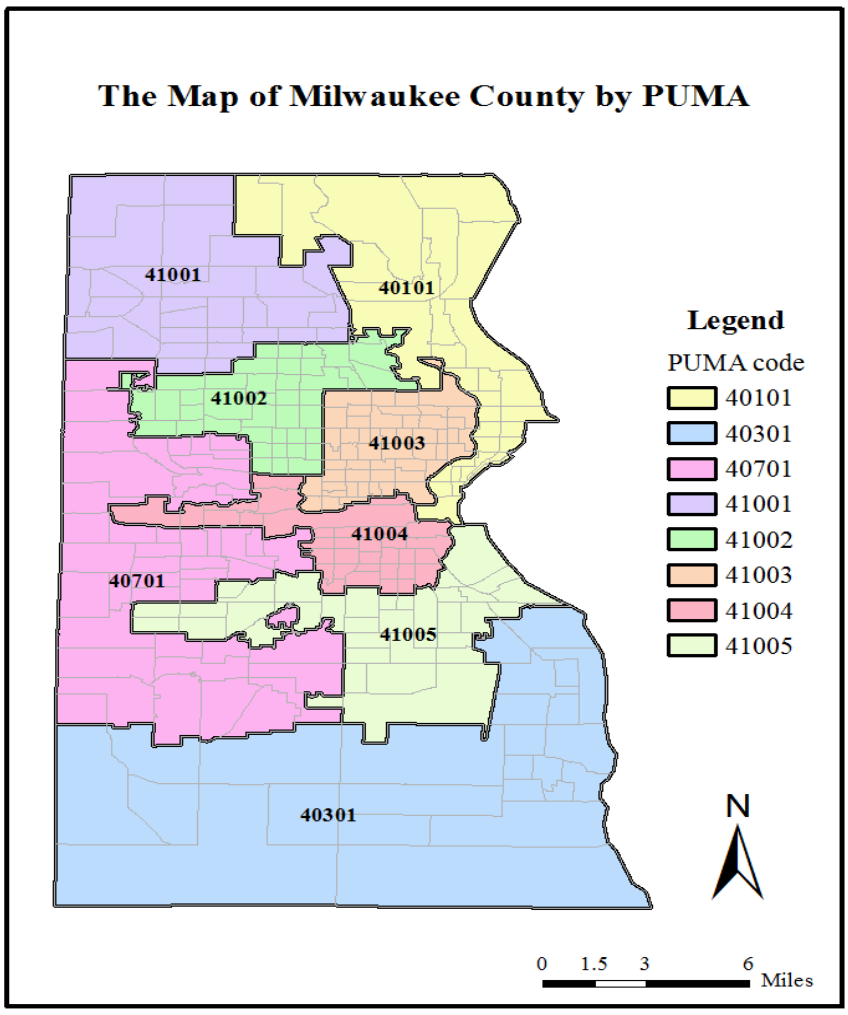

Figure 3.7 The map of Milwaukee County by PUMA

3.2.6.4 The Principles to Select Socio-Demographic Variables

After both the seed data and target data are prepared, the next step is to select socio-demographic variables from two respective datasets for population synthesis. Selecting the variables is based on the following principles. First of all, the variables to be synthesized should exist at both datasets (PUMS data and ACS Summary File). For example, suppose the age is the prominent socio-demographic variable identified by the random forest model, then the age variable should both exist at the PUMS data and ACS Summary File. Otherwise, the age variable will not be selected for population synthesis. Secondly, since socio-demographic variables are selected from the PUMS data and ACS Summary File respectively, the data consistency should be ensured, i.e., the total number of individuals and households should match. For example, in the ACS Summary File, it only provides the employment status for persons who are over 16 years old. But, in the PUMS data, persons with all ages are sampled. Therefore, to ensure the data consistency, the number of individuals who are under 16 years old should also be extracted from the ACS Summary File, and be added into the target data contingency table for the variable of employment status. Otherwise, the number of categories will be different, and it will cause the IPF algorithm to fail.

3.2.6.5 Error Checking

Before the population synthesis, error checking is a necessary step to ensure that the IPF will not fail to proceed or fail to converge during the runtime. The error checking involves the following steps:

(1) To ensure that the number of categories is the same. Because the seed and target data come from different sources, it is highly possible that the number of categories between the seed and target data is different. If the categories are different, it is essential to re-categorize data to ensure that the number of categories of the variables between the seed and target data is exactly the same before the IPF algorithm is conducted.

(2) To verify that the marginal totals in a multi-way contingency table match for all related variables. For example, the total number of individuals calculated from the individual-level socio-demographic variables should be identical to all variables. Similarly, the total number of households should also be the same for all household-level socio-demographic characteristics.

(3) To pre-process zero-cells. In an IPF-based algorithm, the occurrence of zero-cells is a common problem that needs to be resolved. Zero-cells can lead to the failure of convergence of the IPF algorithm. In this dissertation, the zero-cells are pre-processed by replacing them with a very small positive number (10^{-6}). At the same time, the zero marginal totals should also be treated correctly. Zero marginal totals refer to a situation in which the sums of certain categories are zeros. The IPF algorithm implements the division operation where the marginal totals serve as divisors, so the zero marginal totals will make the IPF algorithm fail to proceed and raise an error during the runtime. This problem is handled by re-organizing the categories of variables. More specifically, if cell values across a category of a variable are all zeros, then the category should be combined with other categories.

3.2.6.6 Fit the Individual and Household Level Simultaneously

In this dissertation, the IPF algorithm is used to compute the weights for all individuals in a single PUMA zone first. The weights generated by the IPF algorithm at the individual level are aggregated by the same households, and then carry out the IPF algorithm at the household level to calculate the weights for all households within the PUMA zone. The IPF-generated weight is the measure of representativeness for individuals and households, and they represent the number of times that individuals and households need to be synthesized within the PUMA zone. After the generation of synthetic population in a single PUMA is completed, the same procedure is repeated zone-by-zone to synthesize all PUMA zones within the Milwaukee County. The detailed steps are listed as follows.

First of all, the individual level samples from the PUMS data are used to generate a multi-way contingency table for one PUMA. Because there is no prior knowledge about individuals' representativeness, it is assumed that all individuals are equally representative for all census tracts within the PUMA and their initial weights are all set to ones. Then, the IPF algorithm is applied to iteratively update the contingency table constrained by predefined marginal totals in the ACS Summary File until it converges.

Secondly, the IPF-estimated weights are proportionally (by individuals' weights) assigned to individuals that belong to a specific variable category across all census tracts within the PUMA. As such, the assigned weight for each individual represents the number of times that the individual needs to be synthesized.

Thirdly, generate a multi-way contingency table from the PUMS data at the household level and then carry out the IPF algorithm. Unlike the individual level at which the initial weights are all

set to ones, to fit the individual and household levels simultaneously, the initial weights at the household level are computed by aggregating the weights of individuals who belong to the same households. Afterwards, the IPF-estimated weights are proportionally (by households' weights) assigned to households that belong to a specific variable category across all census tracts within the PUMA, such that the assigned weight for each household represents the number of times that the household needs to be synthesized. When the weights for individuals and households are both available, it is required to round the weights to be integers. The rounded weights represent the number of times to synthesize individuals and households across all census tracts within the PUMA.

At the last step, a similar procedure is applied to the rest of PUMA areas in the Milwaukee County to produce all synthetic individuals and households.

3.2.7 Methods to Verify the Synthetic Population

Upon completion of the population synthesis, it is necessary to verify the synthetic population to see how accurate it is. In this dissertation, firstly, the scatter plot is drawn for every PUMA within the Milwaukee County to conduct a preliminary evaluation. Within a PUMA, the synthesized individuals and households are aggregated at the census tract level and then plotted against all socio-demographic variables constraints across all census tracts. To examine the difference between the observed population and the synthesized population, the Normalized RMSE (NRMSE) percentage is applied to verify the synthetic population. The NRMSE is expressed as an error percentage, and its lower value indicates less deviation between the observed and the simulated data. The relation between NRMSE and RMSE is defined as:

$$NRMSE = \frac{RMSE}{max(\theta_{ij}) - min(\theta_{ij})} * 100\% \qquad \text{(Equation 3.5)}$$

where the $max(\theta_{ij})$ and $min(\theta_{ij})$ denote the maximum and minimum observed population respectively.

Since the synthetic population represents the full population, the distributions of all attributes between the observed and the synthesized data should be identical. Therefore, the synthetic population can be verified by comparing the probability distributions of socio-demographic variables (Hermes & Poulsen, 2012). For example, after Beckman et al. (1996) used the two-step IPF algorithm to generate the synthetic population, they obtained the joint probability of household size and the number of vehicles in each household, and then compared it with the joint distribution of true population to evaluate the performance of the two-step IPF algorithm. In this dissertation, this method is also chosen to verify the synthetic population. In particular, continuous socio-demographic variables are represented by probability density and cumulative density plots, and categorical socio-demographic variables are displayed as bar charts.

Chapter Four: Results and Discussion

4.1 Results of the Travel Activity Analysis on the 2001 and 2009 NHTS

4.1.1 Major Travel Activity Patterns in NHTS

This section reports the results of the travel activity analysis, and describes the major travel activity patterns in the 2001 and 2009 NHTS. By using the method provided in Chapter Three, an individual's daily travel and activities is represented as a sequence of activities chained by travel behaviors, with the attributes of distance to home and activity sequence. The dissimilarity matrices of distance to home and activity sequence are derived from Euclidean distance calculation and global pairwise sequence alignment respectively. The matrices are then normalized and integrated into a final dissimilarity matrix for clustering analysis. In the clustering analysis, the variation of within cluster sum of squares (WCSS) by the number of clusters is plotted for distance to home and activity sequence respectively as demonstrated in Figure 4.1, to determine the appropriate choice of clustering numbers. From Figure 4.1 we observe that the shapes of two graphs are similar. The WCSS drops down quickly with the increase of the number of clusters at the beginning, and then it slows down. The number of clusters can be reasonably set to five. This is because after the "elbow" turning point around the cluster number five, the increase of clustering numbers will no longer greatly affect the value of WCSS. That is to say, the individual travel activities in the 2001 and 2009 NHTS can be classified into five different groups.

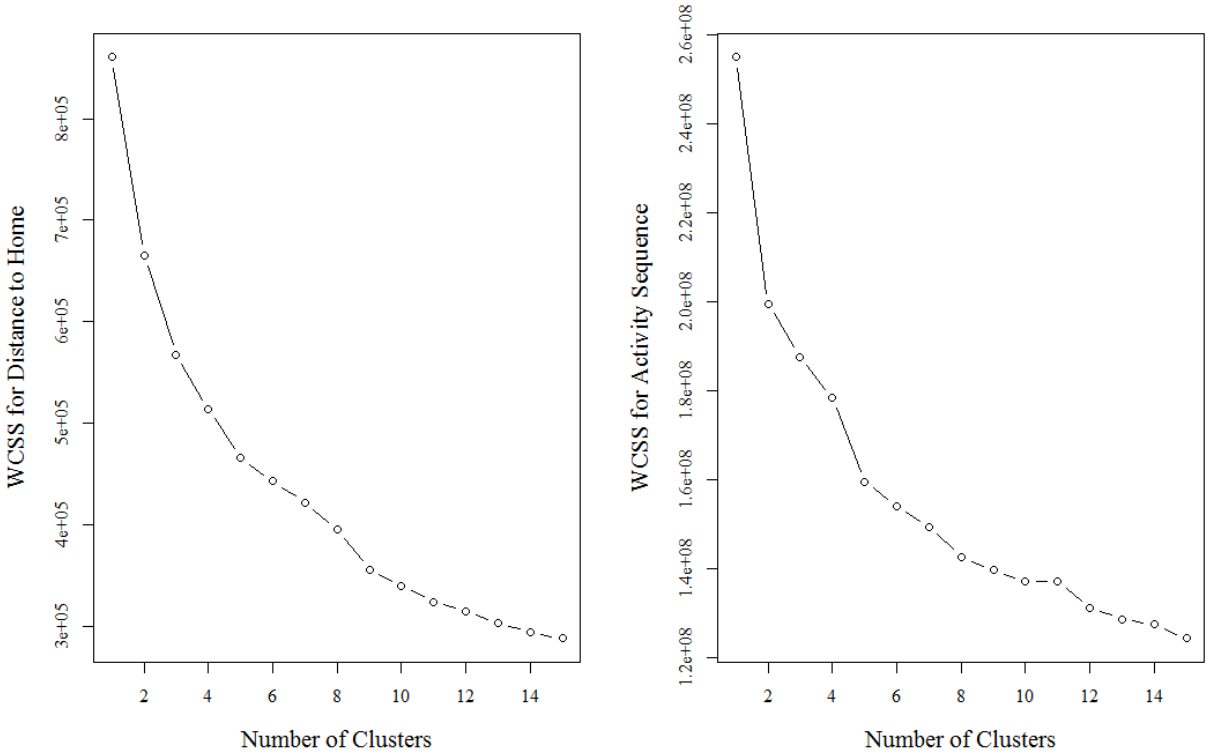

Figure 4.1 The decrease of WCSS by the number of clusters for distance to home and activity sequence

After the number of clusters is determined, the hierarchical clustering analysis is conducted based on the normalized, integrated dissimilarity matrix. The travel activity patterns are partitioned into five different groups by the hierarchical clustering method, and the results are presented in Figure 4.2 and Figure 4.3. For an easier visual interpretation of the clustering analysis results, the majority values of the distance to home and the activity sequence are drawn for every group. The size of each activity pattern group is summarized in Table 4.1 and Table 4.2.

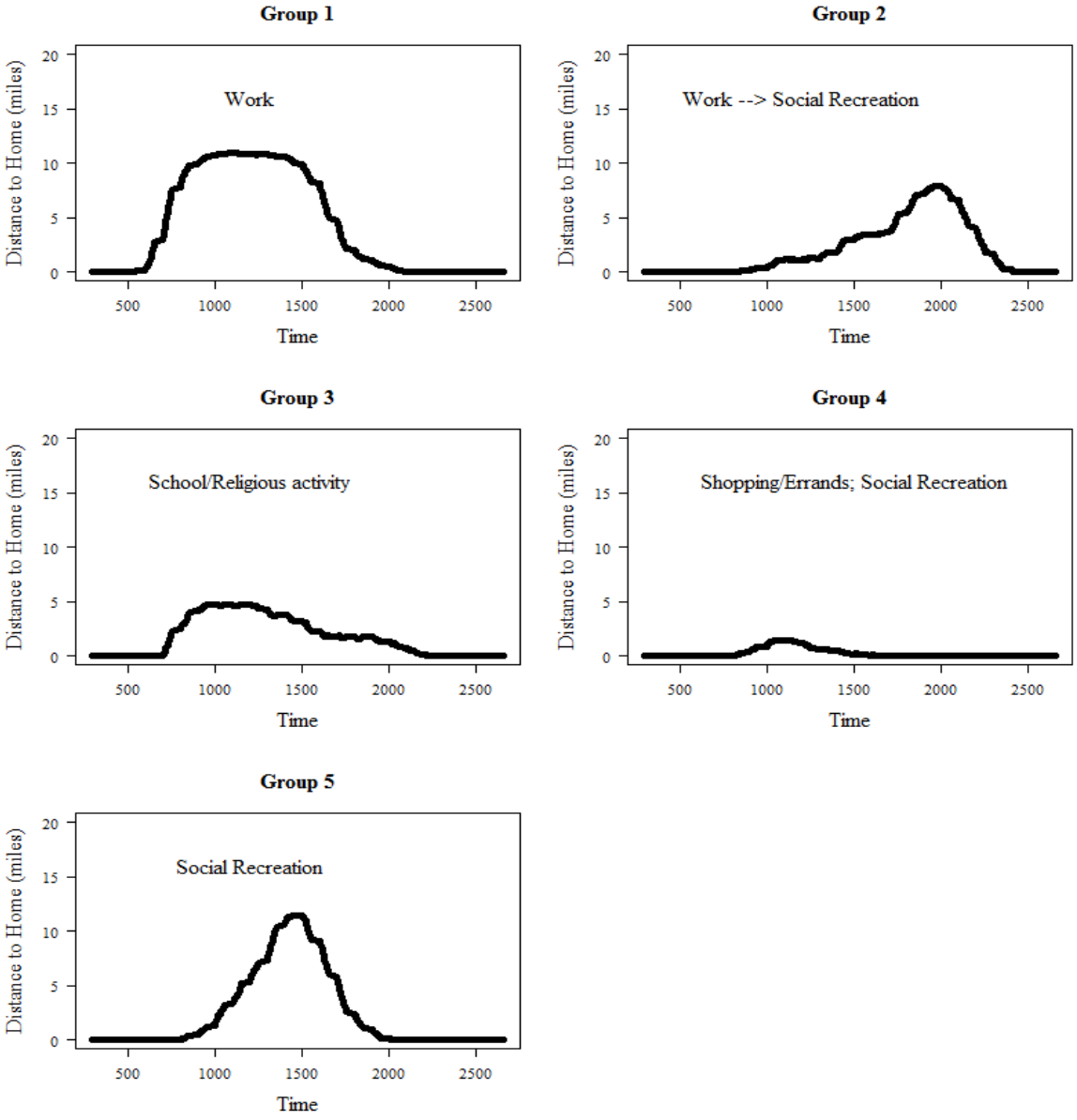

Figure 4.2 Activity pattern groups in the 2001 NHTS

Table 4.1 Sample size in activity pattern groups in 2001 NHTS

Group	1	2	3	4	5	Sum
Size	1,060	358	116	1,109	357	3,000

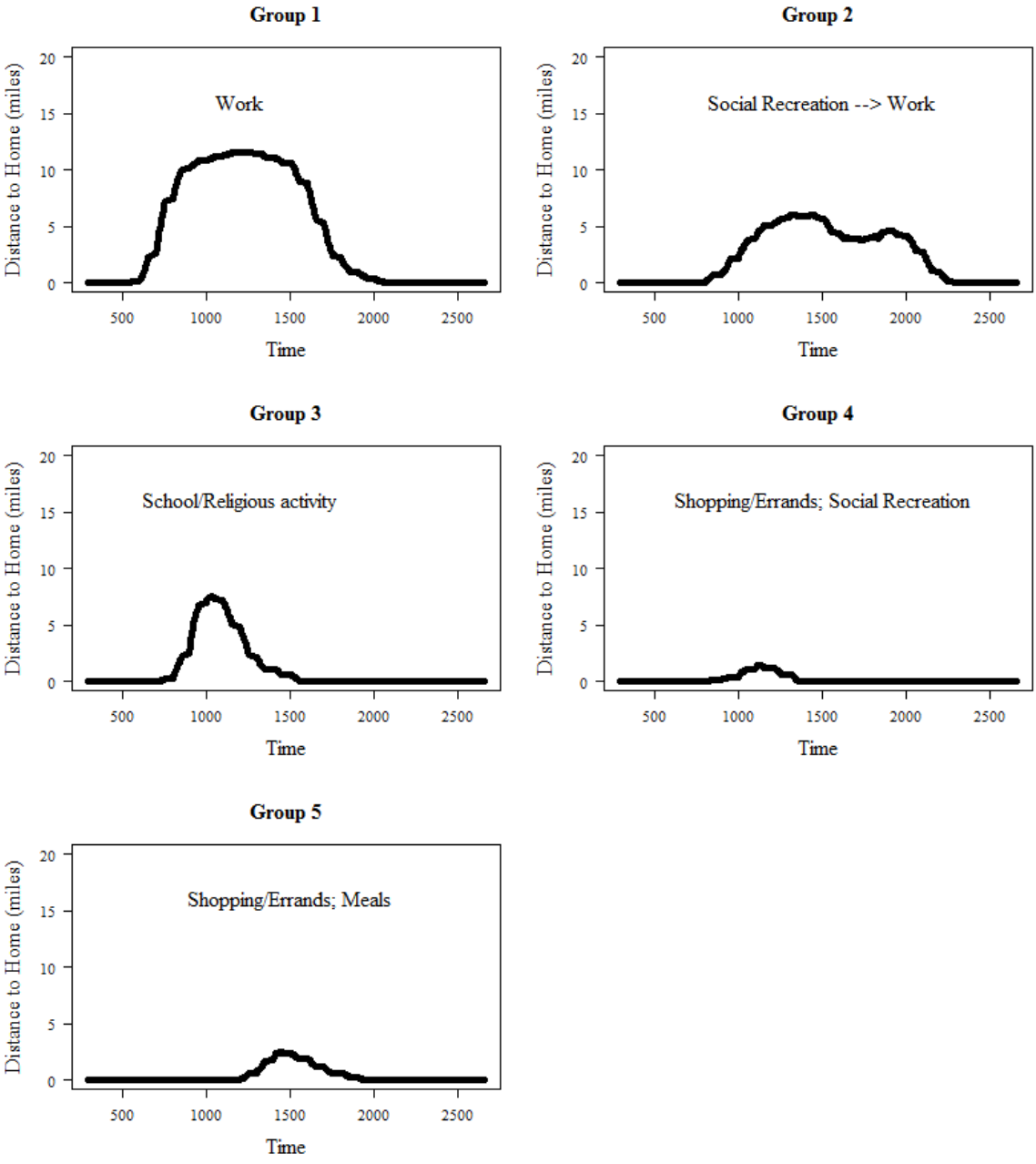

Figure 4.3 Activity pattern groups in the 2009 NHTS

Table 4.2 Sample size in activity pattern groups in 2009 NHTS

Group	1	2	3	4	5	Sum
Size	933	584	274	605	604	3,000

4.1.2 Description of Activity Pattern Groups in the 2001 and 2009 NHTS

In the 2001 NHTS, for Group 1, it is noticed that there is only one type of activity, which is working. The majority of individuals in this group go to work in the morning, stay at the workplace until they return to homes in the afternoon. The travel distance for people who go to work as shown in Figure 4.2 Group 1 is about 10 miles. We can find that travel activity pattern for Group 1 in the 2009 NHTS is almost identical as that of the 2001 NHTS. In the 2009 NHTS, the majority of activities falling into Group 1 are also the single-stop working activities. The travel distance for people who go to work as shown in Figure 4.3 Group 1 is a little more than 10 miles. The sizes of Group 1 of the 2001 and 2009 NHTS are 1060 and 933 respectively. Approximately one-third of the sampled population is in this group.

Group 2 in the 2001 NHTS is characterized by multi-stop activities. There are 358 individuals in this group. The majority of individuals in this group go to work in the morning, then attend social recreation after work, and then return to home at night. The travel distance is less than 10 miles from home. But in the 2009 NHTS, the situation is different from that in the 2001 NHTS. Firstly, the number of individuals in Group 2 is larger in 2009, which is 584. Secondly, although these individuals also perform multi-stop activities, they appear late workers because they go to social recreation first, and then go to work in the afternoon, and then go home after work at night. Lastly, the travel distance is only about 5-7 miles.

The travel activity patterns of Group 3 in the 2001 and 2009 NHTS appear to be the results of school/religious activities. In the 2001 NHTS, people go to school or participate in religious activities in the morning, and return to home at about 8 PM in the evening. There are only 116 individuals in this group, and the travel distance for people who engage in school/religious activity as shown in Figure 4.2 Group 3 is less than 5 miles in 2001. The travel activity pattern

of Group 3 in the 2009 NHTS has the similar features of Group 3 in the 2001 NHTS. Its major travel activity is also school/religious activity, but people go out in the morning and return to home about 3 PM in the afternoon. In addition, there are 274 individuals in this group in 2009, and the travel distance is about 7-8 miles from home.

For Group 4, the travel activity patterns in the 2001 and 2009 NHTS are very similar. There are no obvious predominant activities, as the activities are miscellaneous. Most individuals' activities are shopping, running personal errands, and social recreation. The travel distance is only 3 miles at most. The size of the group is different between the 2001 and 2009 NHTS. In the 2001 NHTS, there are as many as 1109 individuals in this group; but in the 2009 NHTS, there are only 605.

For Group 5, the travel activity patterns are totally different between the 2001 and 2009 NHTS. In the 2001 NHTS, 357 individuals are classified into this group, and the representative travel activity pattern is social recreation. People go out at about 10 AM in the morning, traveling more than 10 miles for social recreation, and return to home at about 8 PM in the evening. In the 2009 NHTS, 604 individuals are classified into this group, and multiple different activities are involved. The activities include shopping, running personal errands, and eating meals. The travel distance is less than 5 miles, relatively shorter than that in the 2001 NHTS.

In summary, in the 2001 and 2009 NHTS, five types of travel activity patterns are found. Among the five patterns, differences are observed in Group 2 and Group 5 respectively. In contrast, three out of the five patterns are similar, and they are single-stop working activity (Group 1), school/religious activity (Group 3), and miscellaneous activities characterized by shopping, running personal errands, and social recreation (Group 4).

4.2 Prominent Socio-Demographic Characteristics

4.2.1 The Accuracy of the Model

This section reports the analysis results of the random forest model, and the prominent socio-demographic variables that correlate to the major travel activity patterns. As a result of the hierarchical clustering analysis, individuals' travel activities are classified into five categorical groups. The random forest model is used to explore the correlation between travel activity patterns and socio-demographic variables. The performance of the random forest model in the 2001 and 2009 NHTS are presented as confusion matrices in Table 4.3 and Table 4.4 below.

Table 4.3 Confusion matrix in the 2001 NHTS

Predicted / Observed	Group 1	Group 2	Group 3	Group 4	Group 5	Accuracy
Group 1	958	18	4	78	2	90.38%
Group 2	154	30	11	156	7	8.38%
Group 3	8	5	61	36	6	52.59%
Group 4	251	27	27	780	24	70.33%
Group 5	69	7	11	251	19	5.32%
Overall Accuracy = 61.60%						

Table 4.4 Confusion matrix in the 2009 NHTS

Predicted / Observed	Group 1	Group 2	Group 3	Group 4	Group 5	Accuracy
Group 1	848	42	21	7	15	90.89%
Group 2	114	326	76	16	52	55.82%
Group 3	34	53	66	82	39	29.93%
Group 4	70	102	340	28	65	56.2%
Group 5	90	100	122	23	269	44.54%
Overall Accuracy = 62.17%						

The confusion matrix is a table which is used to evaluate the performance of a classification model (Kohavi & Provost, 1998). In a confusion matrix, rows and columns represent the observed and the predicted classes respectively, and the values in the diagonal represent the number of observations that are correctly classified. A confusion matrix can provide the information on the classification model's overall accuracy, as well as the accuracy for each class. The overall accuracy is the proportion of the number of observations in total that are correctly classified, and the accuracy for each class is the percentage of the number of observations that are correctly classified in each group. In both cases, a higher percentage means a better accuracy. Table 4.3 provides a confusion matrix for the 2001 NHTS. In Table 4.3, the overall accuracy of the random forest classification model is 61.6%, which means that 61.6% of observations are assigned to the correct groups. Among the five groups, three groups' accuracies are over 50%, and they are Group 1, Group 3, and Group 4. Moreover, Group 1, which is characterized by a typical single-stop working activity, has the highest accuracy of 90.38%. On the other side, the lowest accuracy is in Group 5, and its accuracy is 5.32%, which indicates that only 5.32% of observations are correctly classified to Group 5.

Table 4.4 is the confusion matrix for the 2009 NHTS. The overall accuracy of the random forest classification model is 62.17%, just a little higher than that of the 2001 NHTS. In particular, three out of five groups' accuracies are higher than 50%, and they are Group 1, Group 2, and Group 4. In the 2009 NHTS, the highest accuracy is still in Group 1, the single-stop working activity. The Group 3 has the lowest accuracy, and it is featured by school/religious activity, because only 29.93% of the individuals are classified into the correct group, which is the lowest in this matrix.

4.2.2 The Importance of Socio-Demographic Variables

The random forest model also provides the information on the importance for all socio-demographic variables. Table 4.5 and Table 4.6 display the importance of socio-demographic variables in the 2001 and 2009 NHTS respectively. The importance of variables is measured by two indicators: the mean decrease in accuracy and the mean decrease in Gini. A large value of the mean decrease in accuracy or Gini suggests the high importance of the variable in the model. For example, in Table 4.5, the variable "number of adults" is less important than the variable "number of workers", because 16.15 is smaller than 29.97 in terms of the mean decrease in accuracy.

Table 4.5 The importance of socio-demographic variables in the 2001 NHTS

	Mean Decrease in Accuracy	Mean Decrease in Gini
Household Incomes	16.21	***219.74***
Number of Adults	16.15	44.48
Number of Workers	***29.97***	94.19
Household Life Cycle	***28.44***	***121.26***
Household Size	17.99	79.5
House Ownership	8.95	32.11
Urban/Rural Indicator	1.54	107.32
Age	***47.4***	***252.31***
Gender	6.79	39.51
Race	8.19	50.81
Employment Status	***63.31***	***155.34***
Education Level	1.98	97.45
Travel Day	***57.97***	***215.61***
Number of Drivers	16.11	46.73
Number of Vehicles	19.8	83.86

Table 4.6 The importance of socio-demographic variables in the 2009 NHTS

	Mean Decrease in Accuracy	Mean Decrease in Gini
Household Incomes	11.54	***216.98***
Number of Adults	12.14	46.02
Number of Workers	***23.78***	99.28
Household Life Cycle	***24.69***	***115.18***
Household Size	12.42	70.45
House Ownership	2.28	22.16
Urban/Rural Indicator	0.13	91.74
Age	***25.82***	***218.96***
Gender	6.35	43.06
Race	1.4	50.9
Employment Status	***71.83***	***154.4***
Education Level	2.7	107.21
Travel Day	***49.73***	***201.57***
Number of Drivers	13.2	51.23
Number of Vehicles	18	93.64

In Table 4.5 and Table 4.6, the five highest values of the mean decrease in accuracy and Gini are highlighted as bold and italic fonts. In Table 4.5, it is worth noting that four out of these five socio-demographic variables are common between the mean decrease in accuracy and Gini, and they are household life cycle, age, employment status, and travel day. The fifth variable of high mean decrease in accuracy is the number of workers. The fifth variable of high mean decrease of Gini is the household income. Therefore, six prominent socio-demographic variables in the 2001 NHTS will be used in the population synthesis. In Table 4.6, the similar situation can be observed. In the 2009 NHTS, the four significant socio-demographic variables in common with the 2001 NHTS are household life cycle, age, employment status, and travel day, and the remaining two variables are household incomes and number of workers. To summarize, six

socio-demographic attributes are prominent to contribute to the distinct travel activity patterns, and they are age, travel day, employment status, number of workers, household incomes, and household life cycle. These variables will be used in population synthesis.

4.2.3 Variable Selection

According to the results from the random forest model, the most important socio-demographic variables to be used in population synthesis are age, travel day, employment status, number of workers, household incomes, and household life cycle. The variable "travel day" is not available at both PUMS data and ACS Summary File, so it is not considered in this research. Age and employment status at the individual level, the number of workers and household incomes at the household level are selected from the PUMS and ACS Summary File respectively. For the household life cycle, although it does not exist in the both data sources, a variable with the closest meaning can be used to replace it. The household life cycle is a composite indicator that integrates the information about marital status, the number of adults, and the number of youngest children for a household. At the PUMS data and ACS Summary File, a variable named "household type" is the most similar variable, because it provides the data about households by householder's marital status, family type, and the number of children under 18 years old. Therefore, in this research, the household type is used to substitute the household life cycle. In summary, age and employment status at the individual level, the number of workers, household incomes, and household type at the household level are selected for the population synthesis.

4.3 Validation of the Synthetic Population

4.3.1 Scatter Plot and Internal Goodness-of-fit

This section reports the results of validation of the synthetic population for all eight PUMAs in the Milwaukee County. Scatter plot is drawn first for each PUMA, and then internal goodness-of-fit testing is conducted against the marginal total constraints. Also, the NRMSE is calculated for each PUMA to see how accurate they are. The details are presented in Figure 4.4 and Figure 4.5 below.

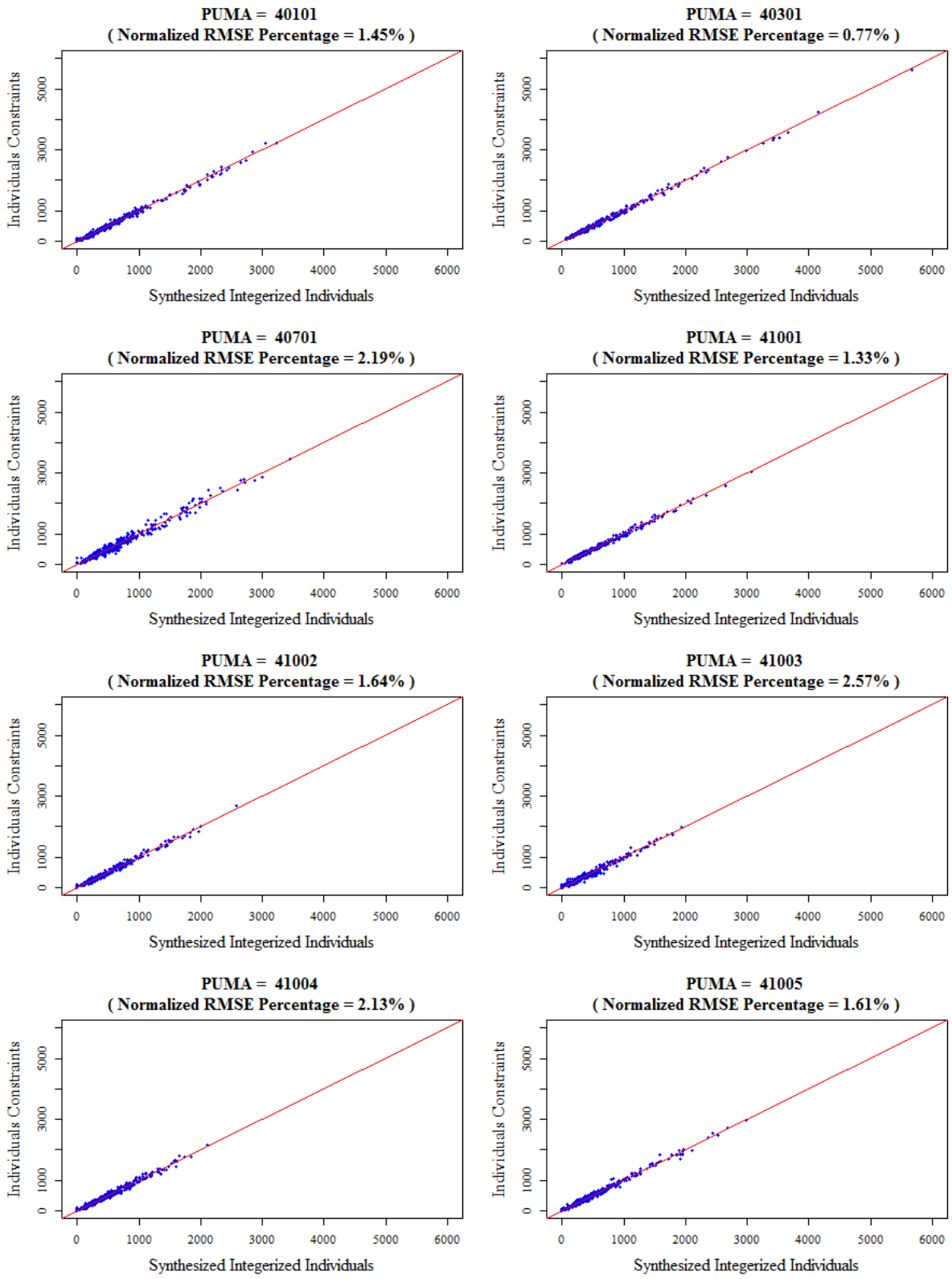

Figure 4.4 The overall fit at the individual level for all PUMAs in the Milwaukee County

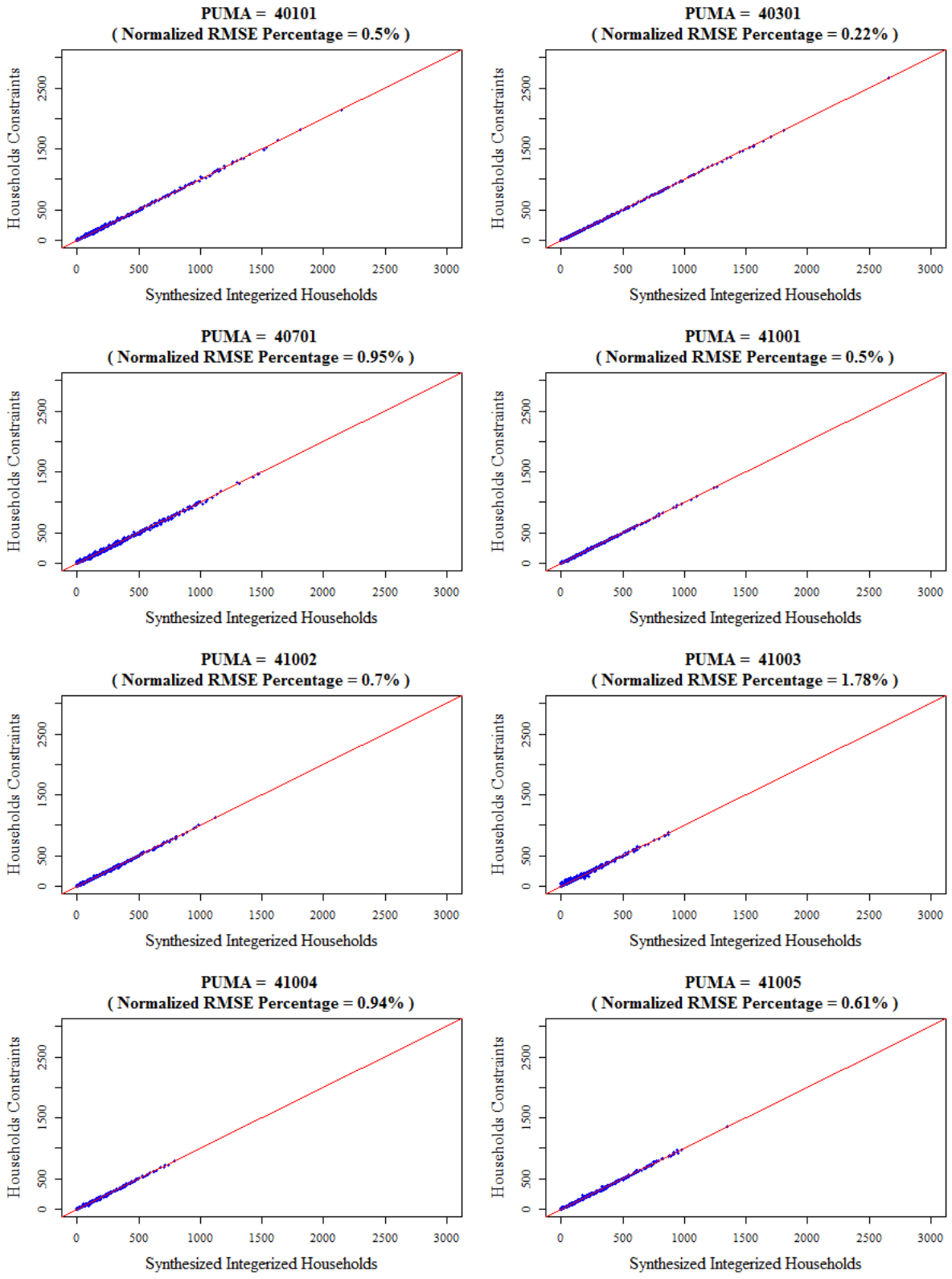

Figure 4.5 The overall fit at the household level for all PUMAs in the Milwaukee County

Figure 4.4 and Figure 4.5 present the overall goodness-of-fit at the individual and household levels for all the eight PUMAs in the Milwaukee County. In each scatter plot, the x value represents the number of synthesized individuals and households, and the y value denotes the observed marginal total constraints. The red line $y = x$ represents the best fitting line. If the synthesized values are exactly the same as the observed constraints, all points would fall on the best fitting line. In most cases, the line of best fit passes through most of the points, and a less deviation between points and the line indicates a better fit.

The NRMSE on each scatter plot gauges the synthesized results. Since it is an error measurement, a lower value means a better fit. As shown in Table 4.7, NRMSE values are all less than 3% at the individual level. The maximum error is 2.57%, occurring at the 41003 PUMA zone. The overall goodness-of-fit improves at the household level, since only one PUMA zone has a value greater than 1% in NRMSE, and the rest of seven PUMA zones have values all less than 1%. The maximum error is still at the 41003 PUMA zone, but the NRMSE has been reduced from 2.57% at the individual level to 1.78% at the household level.

Table 4.7 NRMSE in PUMAs for the Milwaukee County

PUMA	Individual-level NRMSE	Household-level NRMSE
40101	1.45%	0.5%
40301	0.77%	0.22%
40701	2.19%	0.95%
41001	1.33%	0.5%
41002	1.64%	0.7%
41003	2.57%	1.78%
41004	2.13%	0.94%
41005	1.61%	0.61%

4.3.2 Comparison with the Distributions of Socio-Demographic Variables in the PUMS Data

In addition, the synthesized socio-demographic variables are plotted to see whether these variables have the similar distributions as that of the observed variables in the PUMS data. The distributions of continuous variables are represented by probability density and cumulative density charts, and the categorical variables are plotted as grouped bar charts. Figure 4.6 to Figure 4.10 below show the distributions for age, employments status, household incomes, the number of workers, and household type respectively.

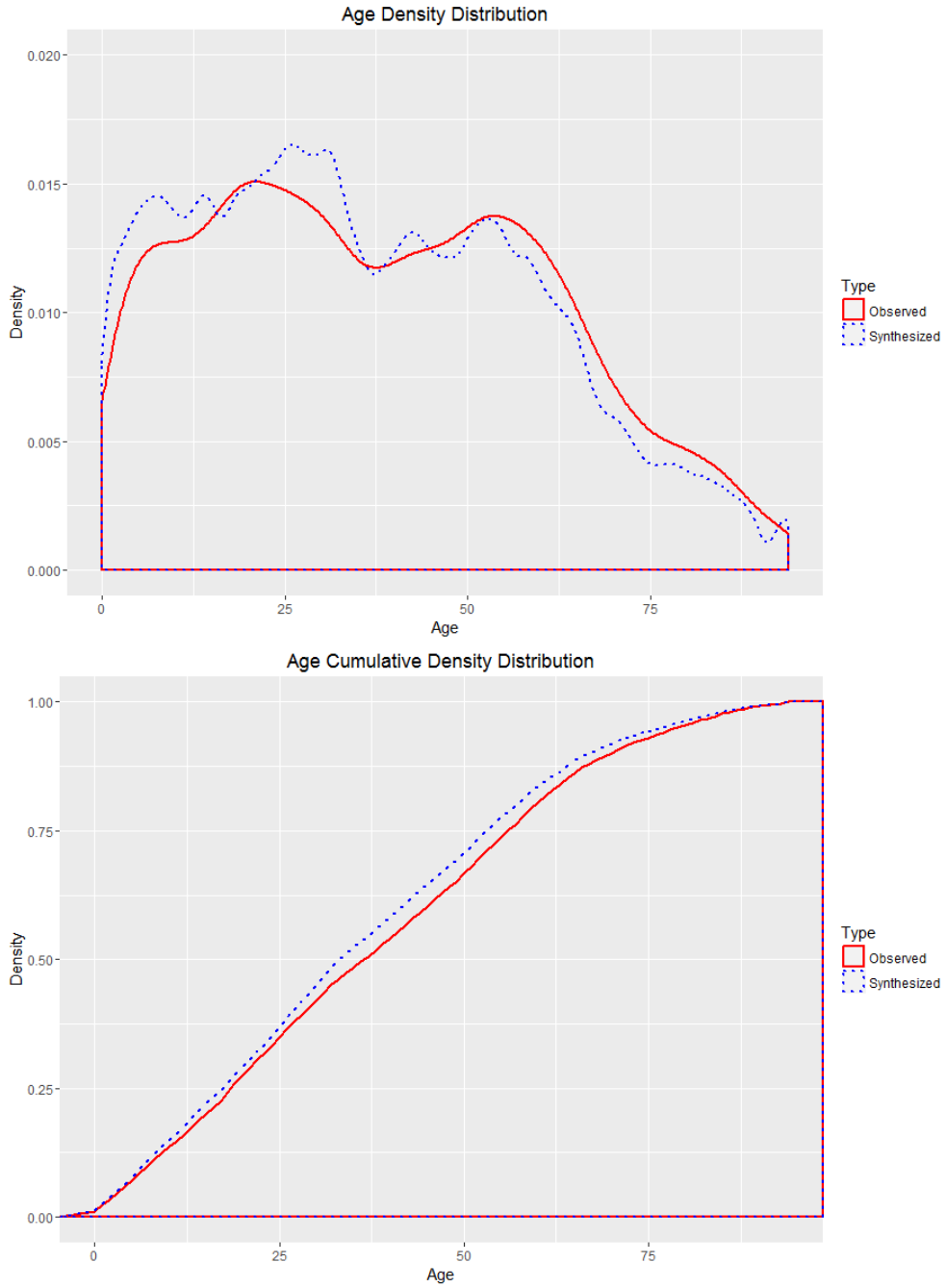

Figure 4.6 Density and cumulative density distribution for age

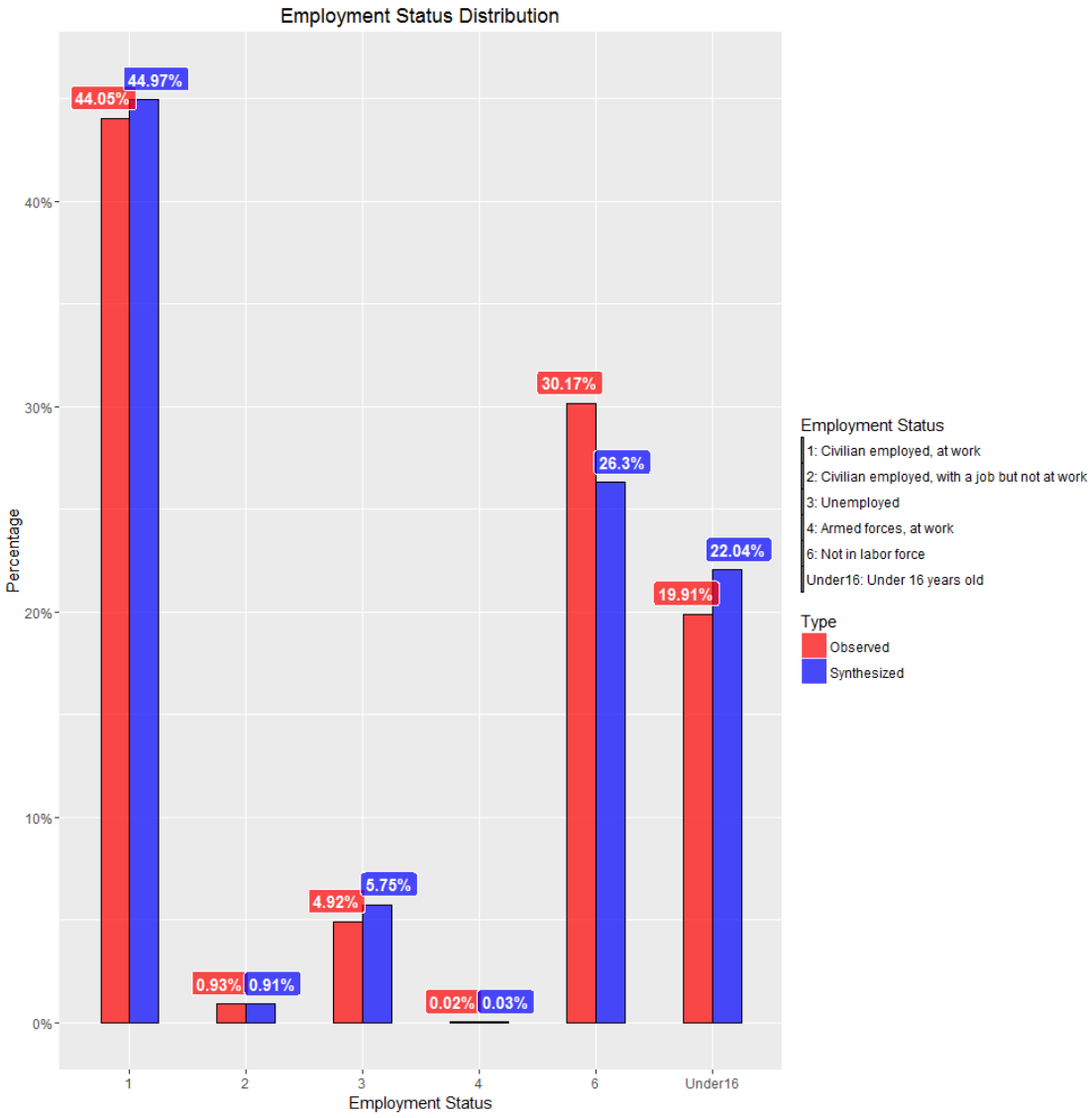

Figure 4.7 Employment status in the observed and the synthesized individual data

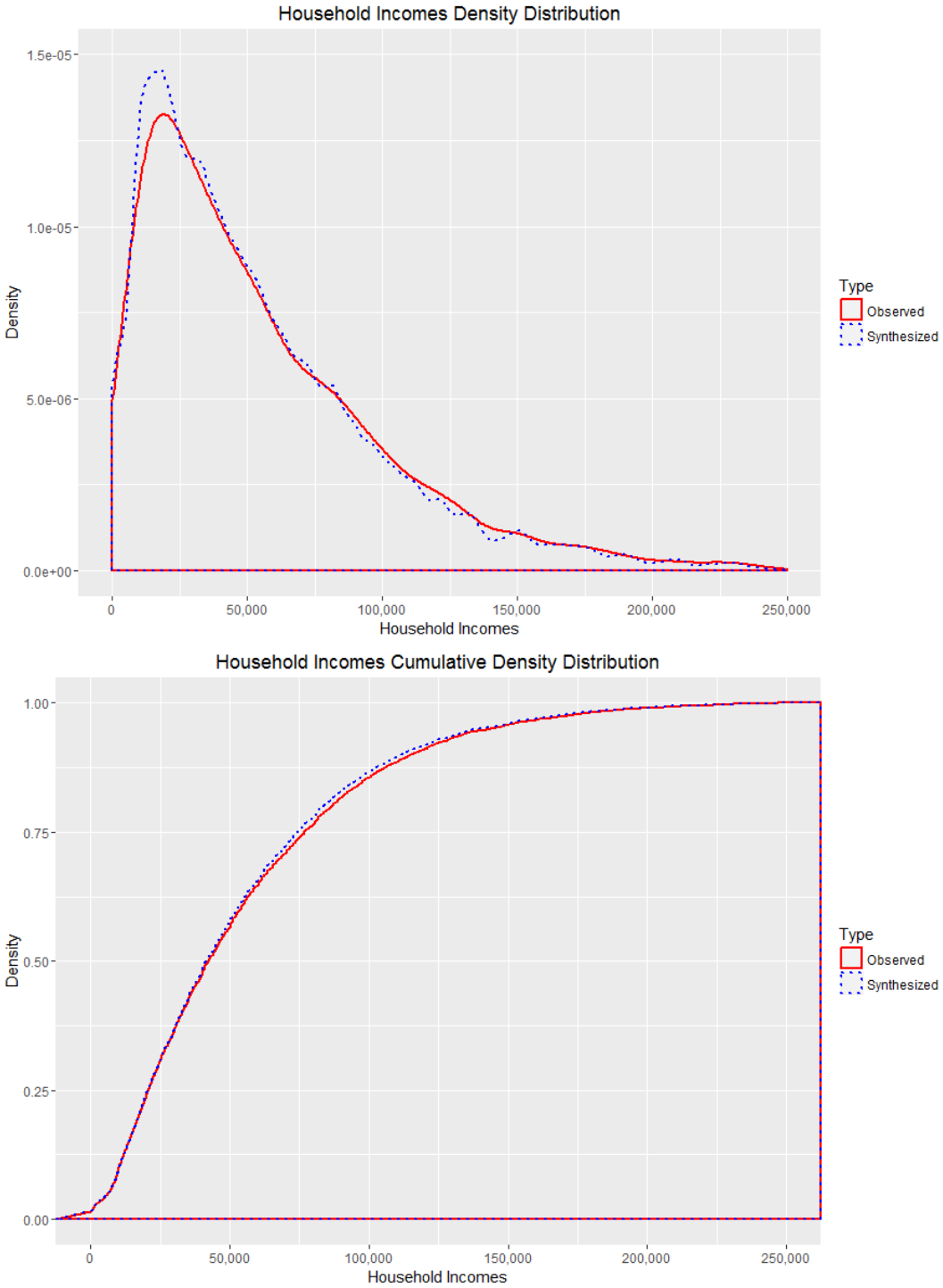

Figure 4.8 Density and cumulative density for household incomes

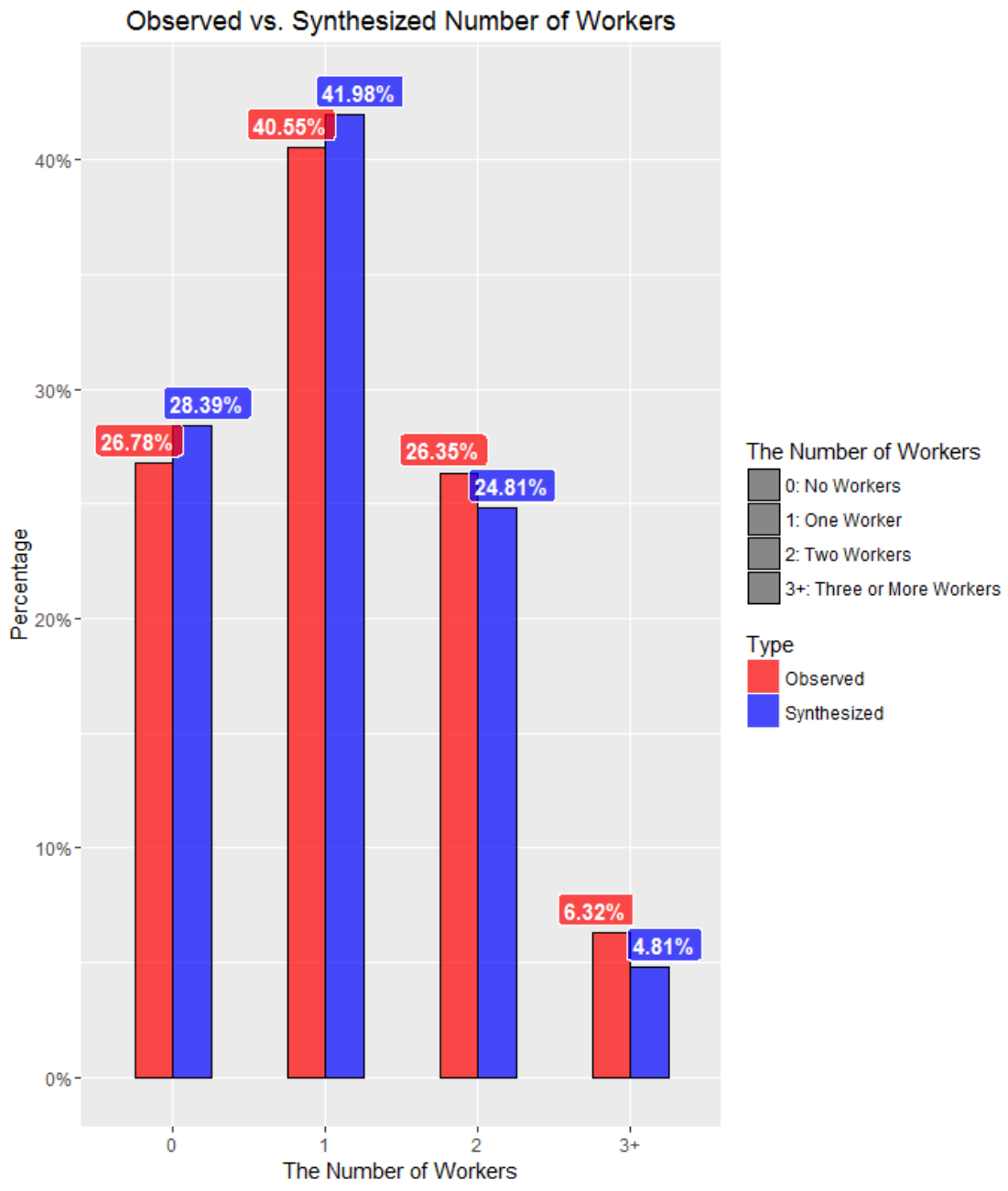

Figure 4.9 Number of workers in the observed and the synthesized household data

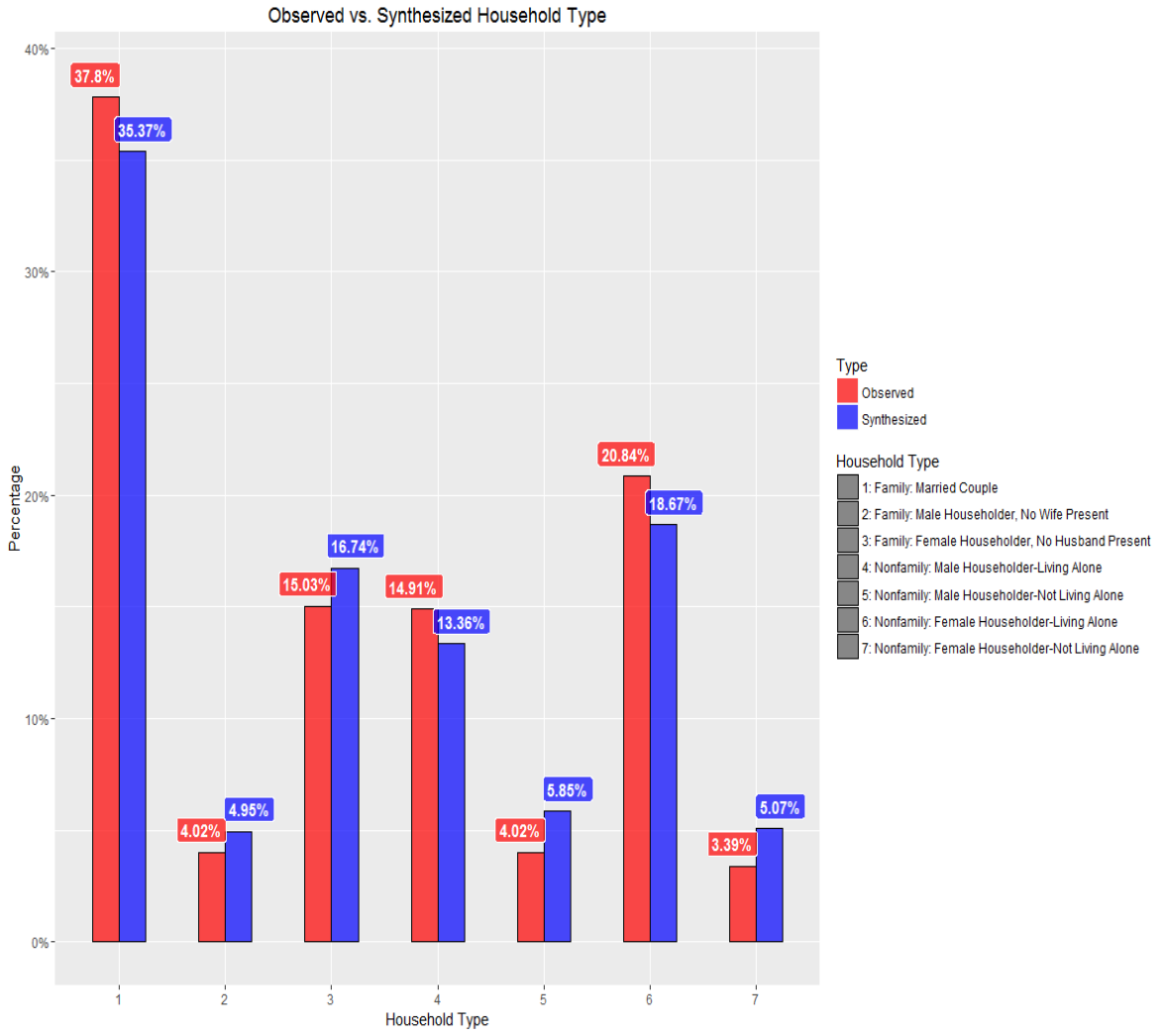

Figure 4.10 Household type in the observed and the synthesized household data

At the individual level, Figure 4.6 describes the density and the cumulative density distributions of the age variable of the synthetic individuals and the observed PUMS data. The solid red line indicates the observed age values, and the dotted blue line represents the synthesized age values. We can notice that the distribution of the synthesized data is similar to that of observed data. Figure 4.7 shows the comparison of the proportions of all categories for the employment status between the observed and the synthesized data. Although the discrepancy is around 3% for the

fifth and the sixth category, the differences for the previous four categories are all smaller than 1%.

At the household level, Figure 4.8 compares the density and the cumulative density distributions of the household incomes. Like Figure 4.6, the solid red line and the dotted blue line denote the household income values from the observed PUMS data and the synthesized data respectively. It is worth noting that their distributions are almost the same. For the number of workers, Figure 4.9 presents the proportions for each category by bar charts. The discrepancies across all categories are less than 2%. Figure 4.10 plots the proportions of all categories for the household type in the observed PUMS data and the synthesized data. The differences are all smaller than 2% except the first category (the married couple family household) and the sixth category (the nonfamily household which female householders live alone). In summary, Figures 4.6 to Figure 4.10 present a reasonable overall goodness-of-fit for five synthesized socio-demographic variables.

Chapter Five: Summary, Conclusions, and Future Research

5.1 Summary and Conclusions

This dissertation has achieved its three research objectives outlined in Chapter One, including analyzing the 2001 and 2009 NHTS to explore the major travel activity patterns, investigating the prominent social-demographic variables that correlate to the major travel activity patterns, and synthesizing the population microdata with those prominent social-demographic factors for travel activity analysis. This dissertation demonstrated a method to generate the synthetic population with socio-demographic attributes prominent for travel activity analysis. Firstly, it represented individuals' daily travel behaviors and activities as pattern vectors, derived dissimilarity matrices for travel distance and activity sequence, and normalized and integrated dissimilarity matrices for clustering analysis that partitions travel activities into discrete groups. The clustering analysis revealed that there are five types of major travel activity patterns in the 2001 and 2009 NHTS. Among these five patterns, three of them (Group 1, Group 3, and Group 4) are similar. To be specific, they are single-stop working activity, school/religious activity, and miscellaneous activities characterized by shopping, running personal errands, and social recreation. Differences are observed in Group 2 and Group 5 between the 2001 and 2009 NHTS.

Based on the results of the clustering analysis, the random forest classification model was employed to explore the correlation between socio-demographic factors and major travel activity patterns, and to investigate the most important variables that correlate individuals' travel and activities. The random forest model can improve the predictive accuracy and provide the ranking of the importance of the explanatory variables. The results of the random forest model

suggested six correlated socio-demographic characteristics, and they were age, travel day, employment status, the number of workers, household incomes, and household life cycle. Then, this dissertation incorporated these prominent socio-demographic variables into the population synthesis.

The IPF-based method was used to synthesize population with the prominent socio-demographic attributes for the Milwaukee County. This dissertation designed a method to aggregate the weights at the individual level after implementing the IPF algorithm, and then applied them to the household level to ensure the probability distribution at different levels can be fitted simultaneously. After the population synthesis, the validation process was conducted to evaluate the performance of the IPF-based algorithm. The overall goodness-of-fit was evaluated by scatter plot and NRMSE. The distributions of socio-demographic variables in the synthesized data were compared with that of the observed PUMS data to see their accuracy. The results of the validation process indicated that the overall goodness-of-fit is reasonable, and the maximum value of the NRMSE can be managed under 3%.

In conclusion, this dissertation focused on the travel activity analysis and its associated social, economic, and demographic variables, and generated a full size of synthetic population microdata with relevant socio-demographic variables. It was found that there are five major travel activity patterns in the 2001 and 2009 NHTS. From 2001 to 2009, only two of the five major patterns have changed, and the rest of three patterns remain the same. In addition, by exploring the relationship between socio-demographic variables and major travel activity patterns, one can find that the most important socio-demographic variables are age, travel day, employment status, the number of workers, household incomes, and household life cycle. Finally, this dissertation used these important socio-demographic variables to synthesize

population microdata that fit particular travel activity patterns. The traditional and well-established method to generate synthetic population is the Beckman's two-step IFP algorithm, but it can only match the socio-demographic variables' distributions at the household level. This dissertation proposed a new method to address this issue, by aggregating the IPF-generated weights of individuals who belong to the same households and then applying the weights to the household level. The results of the validation process suggested that, compared with the traditional two-step IFP algorithm, the population synthesis algorithm proposed in this dissertation can fit the distributions at the individual and household levels simultaneously.

5.2 Future Research

Like much previous research on travel activity analysis, this dissertation studied the individuals' daily travel activity patterns based on the travel survey data. In this case, the travel survey data was based on the 2001 and 2009 NHTS. The NHTS is well recognized as a comprehensive national household travel survey data, and it has found many applications in travel activity analysis. However, the NHTS does not provide detailed location information, especially, the exact geographic locations and the routes to participate in activities are not available. In future research, travel survey data with detailed location information can be included in the travel activity analysis. With the rapid development of location-tracking technology such as GPS, individuals' travel and activity data with location information could be easily acquired.

In addition, in the 2001 and 2009 NHTS, the majority of travel modes are personal vehicles (both over 87% as illustrated in Chapter 3.1), so this dissertation only analyzed those data that use

personal vehicles. To conduct a more comprehensive analysis, other travel modes should be included in the future research.

Finally, this dissertation used the zone-by-zone method to synthesize a full size of population microdata for the Milwaukee County in Wisconsin. Since population synthesis is a computationally intensive algorithm, applying this zone-by-zone method to generate the synthetic population for the entire country will be a time-consuming procedure. For the future research, a more efficient method which can synthesize population in parallel is necessary, if we want to generate the synthetic population for the entire country.

References

Abbott, A., & Forrest, J. (1986). Optimal matching methods for historical sequences. *Journal of Interdisciplinary History, 16*(3), 471-494.

Abbott, A., & Hrycak, A. (1990). Measuring resemblance in sequence data: An optimal matching analysis of musicians' careers. *American Journal of Sociology, 96*(1), 144-185.

Anderson, J. (1971). Space-time budgets and activity studies in urban geography and planning. *Environment and planning, 3*(4), 353-368.

Auld, J. A., Mohammadian, A., & Wies, K. (2009). Population synthesis with subregion-level control variable aggregation. *Journal of Transportation Engineering, 135*(9), 632-639.

Ballas, D., Clarke, G., & Turton, I. (1999). Exploring microsimulation methodologies for the estimation of household attributes. *4th International Conference on GeoComputation, Mary Washington College, Virginia, USA.*

Barrett, C. L., Beckman, R. J., Khan, M., Kumar, V., Marathe, M. V., Stretz, P. E., . . . Lewis, B. (2009). Generation and analysis of large synthetic social contact networks. *Winter Simulation Conference* (pp. 1003-1014). IEEE.

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

Bhat, C. R. (1995). A heteroscedastic extreme value model of intercity travel mode choice. *Transportation Research Part B: Methodological, 29*(6), 471-483.

Birkin, M., Turner, A., & Wu, B. (2006). A synthetic demographic model of the UK population: Methods, progress and problems. *Regional Science Association International British and Irish Section, 36th Annual Conference.*

Bowman, J. L., Bradley, M., Shiftan, Y., Lawton, T., & Ben-Akiva, M. E. (1998). Demonstration of an activity based model system for Portland. *8th World Conference on Transport Research.* Antwerp, Belgium.

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

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and regression trees.* CRC Press.

Bruton, M. J. (1985). *Introduction to transportation planning.*

Cohen, M., Moore, C., Sharp, J., & Giesbrecht, L. (2003). *Highlights of the 2001 National Household Travel Survey.*

Collia, D. V., Sharp, J., & Giesbrecht, L. (2003). The 2001 national household travel survey: A look into the travel patterns of older Americans. *Journal of safety research, 34*(4), 461-470.

Deming, E. W., & Stephan, F. F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *The Annals of Mathematical Statistics, 11*(4), 427-444.

Ding, G. (2009). *Deriving Activity Patterns from Individual Travel Diary Data: A Spatiotemporal Data Mining Approach.* Doctoral dissertation, The Ohio State University.

Farooq, B., Bierlaire, M., Hurtubia, R., & Flötteröd, G. (2013). Simulation based population synthesis. *Transportation Research Part B: Methodological, 58*, 243-263.

Fienberg, S. E. (1970). An iterative procedure for estimation in contingency tables. *The Annals of Mathematical Statistics*, 907-917.

Fraley, C., & Raftery, A. E. (1998). How many clusters? Which clustering method? Answers via model-based cluster analysis. *The computer journal, 41*(8), 578-588.

Franklin, J. (1995). Predictive vegetation mapping: geographic modelling of biospatial patterns in relation to environmental gradients. *Progress in Physical Geography, 19*(4), 474-499.

Frick, M., & Axhausen, K. W. (2004). Generating synthetic populations using IPF and monte carlo techniques. *Swiss Transport Research Conference.*

Giuliano, G. (2003). Travel, location and race/ethnicity. *Transportation Research Part A: Policy and Practice, 37*(4), 351-372.

Guo, J. Y., & Bhat, C. R. (2007). Population synthesis for microsimulating travel behavior. *Transportation Research Record: Journal of the Transportation Research Board, 2014*(1), 92-101.

Hägerstrand, T. (1970). What about people in regional science. *Papers in regional science, 24*(1), 7-24.

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

Hanson, S., & Hanson, P. (1980). Gender and urban activity patterns in Uppsala, Sweden. *Geographical Review*, 291-299.

Hanson, S., & Hanson, P. (1981). The travel-activity patterns of urban residents: dimensions and relationships to sociodemographic characteristics. *Economic geography*, 332-347.

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

Ireland, C. T., & Kullback, S. (1968). Contingency tables with given marginals. *Biometrika, 55*(1), 179-188.

Janelle, D. G., & Goodchild, M. F. (1983). Transportation indicators of space-time autonomy. *Urban Geography, 4*(4), 317-337.

Janelle, D. G., Goodchild, M. F., & Klinkenberg, B. (1988). Space-time diaries and travel characteristics for different levels of respondent aggregation. *Environment and Planning A, 20*(7), 891-906.

Joh, C.-H., Arentze, T. A., & Timmermans, H. J. (2001). A position-sensitive sequence-alignment method illustrated for space-time activity-diary data. *Environment and Planning A, 33*(2), 313-338.

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

Johnston, R. J., & Pattie, C. J. (1993). Entropy-maximizing and the iterative proportional fitting procedure. *The Professional Geographer, 45*(3), 317-322.

Jung, Y., Park, H., Du, D., & Drake, B. L. (2003). A decision criterion for the optimal number of clusters in hierarchical clustering. *Journal of Global Optimization, 25*(1), 91-111.

Kahn, M. E. (2005). The death toll from natural disasters: the role of income, geography, and institutions. *Review of Economics and Statistics, 87*(2), 271-284.

Kitamura, R. (1988). An evaluation of activity-based travel analysis. *Transportation, 15*(1-2), 9-34.

Knudsen, D. C., & Fotheringham, A. S. (1986). Matrix comparison, goodness-of-fit, and spatial interaction modeling. *International Regional Science Review, 10*(2), 127-147.

Kohavi, R., & Provost, F. (1998). Confusion matrix. *Machine learning, 30*(2-3), 271-274.

Kwan, M.-P. (1998). Space-time and integral measures of individual accessibility: a comparative analysis using a point-based framework. *Geographical Analysis, 30*(3), 191-216.

Kwan, M.-P. (1999). Gender and individual access to urban opportunities: a study using space–time measures. *The Professional Geographer, 51*(2), 211-227.

Kwan, M.-P. (2000). Interactive geovisualization of activity-travel patterns using three-dimensional geographical information systems: a methodological exploration with a large data set. *Transportation Research Part C: Emerging Technologies, 8*(1), 185-203.

Kwan, M.-P. (2004). GIS Methods in Time-Geographic Research: Geocomputation and Geovisualization of Human Activity Patterns. *Geografiska Annaler: Series B, Human Geography*, *86*(4), 267-280.

Kwan, M.-P., & Lee, J. (2004). Geovisualization of human activity patterns using 3D GIS: a time-geographic approach. *Spatially integrated social science, 27*.

Kwan, M.-P., & Ren, F. (2008). Analysis of human space-time behavior: Geovisualization and geocomputational approaches. *Understanding Dynamics of Geographic Domains, CRC Press, New York*, 93-113.

Lenntorp, B. (1976). *Paths in space-time environments: A timegeographic study of movement possibilities of individuals* (Vol. 44). Lund: Royal University of Lund, Department of Geography.

Levinson, D., & Kumar, A. (1995). Activity, travel, and the allocation of time. *Journal of the American Planning Association, 61*(4), 458-470.

Lewis, R. J. (2000). An introduction to classification and regression tree (CART) analysis. *Annual meeting of the society for academic emergency medicine in San Francisco*, (pp. 1-14).

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R news, 2*(3), 18-22.

Lloyd, S. (1982). Least squares quantization in PCM. *IEEE transactions on information theory, 28*(2), 129-137.

Lomax, N., & Norman, P. (2016). Estimating Population Attribute Values in a Table: “Get Me Started in” Iterative Proportional Fitting. *The Professional Geographer, 68*(3), 451-461.

Lovelace, R., Birkin, M., Ballas, D., & van Leeuwen, E. (2015). Evaluating the performance of Iterative Proportional Fitting for spatial microsimulation: new tests for an established technique. *Journal of Artificial Societies and Social Simulation, 18*(2), 21.

Lu, X., & Pas, E. I. (1999). Socio-demographics, activity participation and travel behavior. *Transportation Research Part A: Policy and Practice, 33*(1), 1-18.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, 281-297.

McNally, M. G. (1996). *An activity-based microsimulation model for travel demand forecasting.*

McNally, M. G. (2008). *The four step model.* Retrieved from UC Irvine: Center for Activity Systems Analysis: http://escholarship.org/uc/item/0r75311t

Miller, H. J. (1991). Modelling accessibility using space-time prism concepts within geographical information systems. *International Journal of Geographical Information System, 5*(3), 287-301.

Miller, H. J. (1999). Measuring space-time accessibility benefits within transportation networks: basic theory and computational procedures. *Geographical analysis, 31*(1), 1-26.

Mohammadian, A. K., Javanmardi, M., & Zhang, Y. (2010). Synthetic household travel survey data simulation. *Transportation Research Part C: Emerging Technologies, 18*(6), 869-878.

Mohammadian, A., & Bekhor, S. (2008). Travel behavior of special population groups. *Transportation, 35*(5), 579-583.

Mosteller, F. (1968). Association and estimation in contingency tables. *Journal of the American Statistical Association*, 1-28.

Müller, K., & Axhausen, K. W. (2010). Population synthesis for microsimulation: State of the art. *ETH Zürich, Institut für Verkehrsplanung, Transporttechnik, Strassen-und Eisenbahnbau (IVT)*.

Needleman, S. B., & Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology, 48*(3), 443-453.

Ng, M. K. (2000). A note on constrained k-means algorithms. *Pattern Recognition, 33*(3), 515-519.

Norman, P. (1999). Putting iterative proportional fitting on the researcher's desk.

Openshaw, S. (1984). The Modifiable Areal Unit Problem. In *Concepts and techniques in modern geography.* Norwick: Geo Books.

Pas, E. I. (1984). The effect of selected sociodemographic characteristics on daily travel-activity behavior. *Environment and Planning A, 16*(5), 571-581.

Pas, E. I. (1988). Weekly travel-activity behavior. *Transportation, 15*(1-2), 89-109.

Pereira, J., & Itami, R. M. (1991). GIS-based habitat modeling using logistic multiple regression-A study of the Mt. Graham red squirrel. *Photogrammetric Engineering and Remote Sensing, 57*(11), 1475-1486.

Prasad, A. M., Iverson, L. R., & Liaw, A. (2006). Newer classification and regression tree techniques: bagging and random forests for ecological prediction. *Ecosystems, 9*(2), 181-199.

Pritchard, D. R., & Miller, E. J. (2012). Advances in population synthesis: fitting many attributes per agent and fitting to household and person margins simultaneously. *Transportation, 39*(3), 685-704.

Pucher, J., & Renne, J. L. (2003). Socioeconomics of urban travel: evidence from the 2001 NHTS. *Transportation Quarterly, 57*(3), 49-77.

Pucher, J., & Renne, J. L. (2004). Urban-Rural differences in mobility and mode choice: Evidence from the 2001 NHTS. *Bloustein School of Planning and Public Policy, Rutgers University*.

Pucher, J., Buehler, R., Merom, D., & Bauman, A. (2011). Walking and cycling in the United States, 2001–2009: evidence from the National Household Travel Surveys. *American Journal of Public Health, 101*(S1), S310-S317.

Quinlan, J. R. (1986). Induction of decision trees. *Machine learning, 1*(1), 81-106.

Quinlan, J. R. (1993). C4. 5: Programming for machine learning. *Morgan Kauffmann*, 38.

Recker, W. W. (1995). The household activity pattern problem: General formulation and solution. *Transportation Research Part B: Methodological, 29*(1), 61-77.

Recker, W. W., McNally, M. G., & Root, G. S. (1985). Travel/activity analysis: pattern recognition, classification and interpretation. *Transportation Research Part A: General, 19*(4), 279-296.

Recker, W., McNally, M. G., & Root, G. S. (1981). Application of pattern recognition theory to activity pattern analysis.

Root, G. S., & Recker, W. W. (1981). *Toward a dynamic model of individual activity pattern formulation.* Institute of Transportation Studies and School of Engineering, University of California, Irvine.

Roseman, C. C. (2010). Migration as a spatial and temporal process. *Annals of the Association of American Geographers, 61*(3), 589-598.

Shoval, N., & Isaacson, M. (2007). Sequence alignment as a method for human activity analysis in space and time. *Annals of the Association of American geographers, 97*(2), 282-297.

Simpson, L., & Tranmer, M. (2005). Combining sample and census data in small area estimates: iterative proportional fitting with standard software. *The Professional Geographer, 57*(2), 222-234.

Stephan, F. F. (1942). An iterative method of adjusting sample frequency tables when expected marginal totals are known. *The Annals of Mathematical Statistics, 13*(2), 166-178.

Stopher, P. R., Greaves, S., & Bullock, P. (2003). Simulating household travel survey data: application to two urban areas. *82nd Annual Meeting of the Transportation Research Board.* Washington, DC.

Svetnik, V., Liaw, A., Tong, C., Culberson, J. C., Sheridan, R. P., & Feuston, B. P. (2003). Random forest: a classification and regression tool for compound classification and QSAR modeling. *Journal of chemical information and computer sciences, 43*(6), 1947-1958.

Tal, G., & Handy, S. (2010). Travel behavior of immigrants: An analysis of the 2001 National Household Transportation Survey. *Transport Policy, 17*(2), 85-93.

Tal, G., & Handy, S. L. (2005). The travel behavior of immigrants and race/ethnicity groups: an analysis of the 2001 national household transportation survey. *Institute of Transportation Studies*.

Thornton, P. R., Williams, A. M., & Shaw, G. (1997). Revisiting time-space diaries: an exploratory case study of tourist behaviour in Cornwall, England. *Environment and Planning A, 29*(10), 1847-1867.

Thrift, N. (2002). The future of geography. *Geoforum, 33*(3), 291-298.

Wagstaff, K., Cardie, C., Rogers, S., & Schrödl, S. (2001). Constrained k-means clustering with background knowledge. *ICML, Vol. 1*, pp. 577-584.

Wheaton, W. D., Cajka, J. C., Chasteen, B. M., Wagener, D. K., Cooley, P. C., Ganapathi, L., . . . Allpress, J. L. (2009). Synthesized population databases: A US geospatial database for agent-based models. *Methods report (RTI Press), 10*, 905.

Williamson, P., Birkin, M., & Rees, P. H. (1998). The estimation of population microdata by using data from small area statistics and samples of anonymised records. *Environment and Planning A, 30*(5), 785-816.

Wilson, W. C. (1998). Activity pattern analysis by means of sequence-alignment methods. *Environment and Planning A, 30*(6), 1017-1038.

Wong, D. W. (1992). The Reliability of Using the Iterative Proportional Fitting Procedure. *The Professional Geographer, 44*(3), 340-348.

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

Zhu, S., Wang, D., & Li, T. (2010). Data clustering with size constraints. *Knowledge-Based Systems, 23*(8), 883-889.

Curriculum Vitae

Hong Zhuo

Education

Ph.D., University of Wisconsin-Milwaukee, August 2017

Major: Geography

M.A., University of Toledo, May 2012

Major: Geography

B.S., Minjiang University, June 2008

Major: Resources, Environment, and Urban Planning

Dissertation Title: Synthesizing Population for Travel Activity Analysis

Awards/Honors: Chancellor's Graduate Student Awards, September 2012

Affiliations: American Association of Geographers (AAG)