Title
Tour Complexity, Variability, and Pattern using Longitudinal GPS Data

Author
Shi, Xiaoxia

Publication Date
2017

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA,
IRVINE

Tour Complexity, Variability and Pattern using Longitudinal GPS Data

DISSERTATION

Submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Transportation Science

by

Xiaoxia Shi

Dissertation Committee:
Professor Douglas Houston, Chair
Professor Michael G. McNally
Professor Jean-Daniel Saphores

2017

DEDICATION

In loving memory of my grandparents

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to express the deepest appreciation to my committee chair, Professor Douglas Houston. Without his guidance and help this dissertation would not have been possible.

I would like to thank Wantzu (Ashley) Lo, Thi Bich Thuy (Vanessa) Luong, Dongwoo Yang, and Gaby Abdel-Salam who helped with my research and provided encouragement along the way. In addition, I am grateful to the many people and parties who contributed to the Expo Study, on which this dissertation is built on. I would like to thank Professor Marlon Boarnet, Professor Douglas Houston, and Professor Steven Spears for leading this study, and the study participants and thank the research assistants who supported data collection and processing. The Expo Line Study was supported by the California Air Resources Board, the Haynes Foundation, the Lincoln Institute of Land Policy, the San Jose State Mineta Transportation Institute, the Southern California Association of Governments, the University of California Transportation Center, the University of California Multi-Campus Research Program on Sustainable Transportation, and the University of Southern California Lusk Center for Real Estate.

I alone am responsible for the results, interpretations, and any errors in this work.

CURRICULUM VITAE

Xiaoxia Shi

2010	B.E. in Urban Planning, Southeast University, Nanjing, China
2012	Master in Urban and Regional Planning, University of California, Irvine
2013 - 2016	Graduate Student Researcher, University of California, Irvine
2015	Teaching Assistant, University of California, Irvine
2017	M.S. in Statistics, Ph.D. in Transportation Science, University of California, Irvine

FIELD OF STUDY

Travel Behavior and Geospatial Data Mining

ABSTRACT OF THE DISSERTATION

Trip Chaining Complexity, Variability, and Pattern using Longitudinal GPS Data

By

Xiaoxia Shi

Doctor of Philosophy in Transportation Science

University of California, Irvine, 2017

Professor Douglas Houston, Chair

Trip chaining is a common phenomenon generally known as linking multiple activities and trips in one travel process. A good understanding about trip chaining complexity is important for travel demand model development and for transportation policy design. However, most of the existing studies on trip chaining limit the complexity classification scheme on number of trips chained and neglect other dimensions that also elevate the degree of complexity. The purpose of this study is to develop a new approach, Tour Complexity Index (TCI), that integrates the multi-dimensional nature of trip chaining into the complexity assessment. The study contains three analysis components. The first component introduces the TCI approach as a trip chaining complexity measure that not only considers number of trips chained but also includes the spatial relationship across destinations, the route arrangement, and the urban environment of the destinations. By comparing descriptive statistics and generalized linear model results from TCI approach with those from traditional approach, we find that the TCI approach offers more information regarding trip chaining and mode choice. The application of TCI is further demonstrated in the following

components. The second component investigates the intrapersonal daily and weekly travel variability with travel characterized by TCI and mode choice. The result reinforces an argument in current literature that the common single-day travel survey may produce biased estimation due to the day-to-day variance in travel behavior. Result also finds that proximity to a new transit service from place of residence is connected with a decline in variability. The third component explores a framework for travel pattern recognition where pattern is characterized by TCI as well. The discrepancy analysis which is a generalized analysis of variance (ANOVA) method is applied to associate individual characteristics with travel pattern. In addition, both components use Sequential Alignment Method (SAM) for travel pattern representation. The TCI approach and proposed analysis frameworks are validated using the longitudinal GPS trajectory data collected between 2011 and 2013 at west Los Angeles area for Expo Study.

Chapter 1 Introduction

Background

The major goal of Senate Bill 375 (SB375) is to reduce greenhouse gas (GHG) emission through adoption of a series of sustainable urban development strategies. One of its main strategies is encouraging people to use other travel modes as an alternative to gasoline-consuming vehicle. We know little about, as more residents transfer from auto-oriented travel to other modes, how will their activity arrangements be influenced? Investigation into such travel behavior adjustment is important for understanding the potential impact of policies for compact development such as SB375.

The research seeks to understand how urban residents link trips and group activities, and how these patterns vary by (1) travel mode, (2) built environment, (3) day of the week, and (4) individuals with different social-demographical background. It helps assess ways in which these patterns are potentially impacted by the construction of a new light rail transit service. It sheds light on how predictable people are in terms of the interpersonal and intrapersonal variability in the way multiple activities are chained together. It will also provide implications on the development of activity-based travel demand theories and empirical demand models.

Policy Context

Adopted in 2006, Assembly Bill 32 (AB 32), Global Warming Solutions Act of 2006, sets a statewide GHG emission reduction goal in California. SB375, the Sustainable Communities and Climate Protection Act of 2008, is a state legislation designed to support AB 32 on vehicle-related GHG reduction. It requires local and regional governments to adopt and incorporate a development plan (Sustainable Communities Strategy, known as SCS) that provides greater integration of land use and transportation planning in order to reduce travel demand and to encourage non-auto travel.

The Regional Transportation Plan (RTP) is a long term transportation plan developed by local Metropolitan Planning Organizations (MPOs), such as Southern California Association of Government (SCAG), every four or five years. It aims to provide a vision for future regional transportation investment over a period of 20 years. Under the requirement per SB375, each MPO, like SCAG, must incorporate a Sustainable Communities Strategy (SCS) into their RTP to reduce GHG emission. The SCS puts the local level development focus on the denser, mixed-use, transit-oriented communities that could lead to less vehicle miles traveled (VMT) and GHG emission. A set of targets were developed by Air Resource Board (ARB) in 2011 to ensure a region's SCS meets the preset regional goal on vehicle-related GHG reduction if it is adopted. The measurement includes: VMT, daily trips per household and share of various travel modes. The latest regional target for the SCAG region is 9% per capita by 2020 and 16% by 2035 (SCAG, 2016).

SCAG's currently-adopted 2012-2035 RTP/SCS covers six counties in Southern California. It is implemented to facilitate the reduction of VMT and GHG through a strategy that bundles land use, transportation, transportation demand management and regional transit project. Its target is "to improve our environment quality while providing mobility for our residents" (SCAG, 2016). The plan emphasizes the transit-oriented-development (TODs) within the existing and planned transit corridors. This 2012-2035 SCS introduces a definition for the concept of "High-Quality Transit Areas" (HQTA), which are areas within 1.5 miles of a well-serviced transit stop. Substantial future distribution of new job and housing in the region is targeted for these designated HQTAs. The expected target is 51% of new employment growth and 53% of new housing. The significance of the integration of land use and public transportation facility lies in its potential to potentially increase alternative transportation options and improve regional accessibility and mobility (SCAG, 2016).

However, the question that how people would regroup their activity location and schedules in response to a new transportation option remains limited. A "tour" is a basic analysis unit in travel behavior context which refers to the grouping and sequence of trips (trip chaining); understanding how tour characteristics change over time and across modes could provide important insights for policy makers. Besides observing how the travel patterns and tour behavior changes over time, it is also important to have a better understanding how people with different needs differ in travel behavior. This could eventually help improve the overall urban mobility by assisting researchers to develop more accurate travel demand models and to design policies that reflect different population groups' needs.

The following content of this introductory chapter is organized in this way: the research framework is first illustrated graphically as an overview of the entire study. The research motivation, goal and purpose are briefly summarized as a complementary part to the framework diagram. Then, the research objectives, questions, background, expected outcomes, assumptions and limitations are presented at a component-specific level. The chapter ends with a discussion about the significance and contribution.

Research Framework

The research draws from a multi-day, multi-year GPS-based dataset from the longitudinal Expo Line Study of travel behavior near a new light rail service in Los Angeles (discussed in Chapter

2). It is a three-component research framework (Figure 1-1) that first investigates the tour complexity to obtain a better understanding on the mechanism of trip chaining behavior (Component I), then examines the within-individual difference in trip chaining behavior through out the week and across years. Last but not least, the research evaluates the between-individual deviance in scheduling and tendency of complex trip chaining behavior.

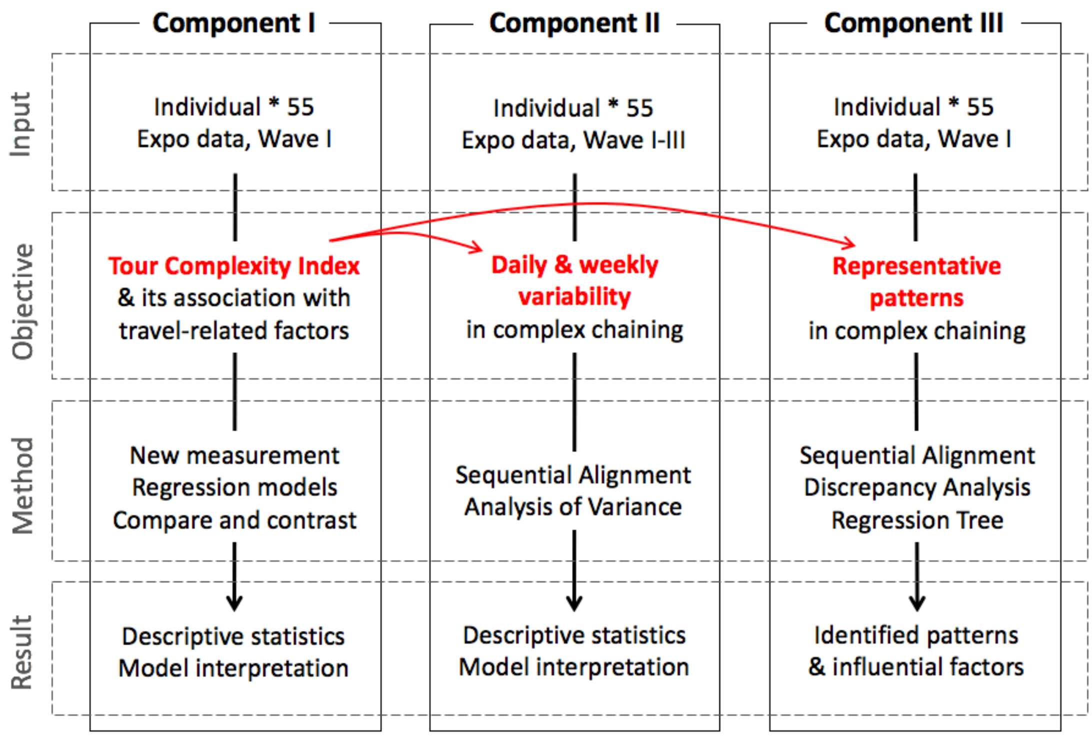


Figure 1-1 Research Framework

The complexity of trip chaining behavior is an emerging topic in travel behavior research field. A better understanding about trip chaining complexity is important for travel demand modeling and for transportation policy evaluation. Previous studies of trip-chaining have mostly paid attention to the magnitude of trip chaining (i.e. how many intermediate trips are inserted in a tour). This component aims to extend the conventional approach in trip chaining complexity study into a multi-dimension approach. Using the evidence from a GPS trajectory database, this study is able to examine how trip chaining complexity level can be further discriminated according to the spatial relationship between activity locations, the route choice, and the urban environment of activity locations. This component also offers an exploration about how GPS trajectory data can contribute to trip chaining studies, considering the existing trip chaining literature is dominated by traditional travel surveys and heavily relies on the self-reported information such as trip purpose and other people presents. The tour complexity measurement discussed in this component more directly incorporates routing and location characteristics into the overall evaluation compared to the traditional approach.

With the measurement of tour complexity is established, the next questions are whether the tour complexity is stable across time for the same individual and whether the tour complexity pattern is closely related with characteristics of individuals. The former question leads to the analysis Component II while the latter directs to Component III. Attention has been received about intrapersonal (within the same individual) day-to-day travel behavior variability since 1980s (Pas, 1987). The clarification of this variability issue is very important in that it directly affects the travel survey design, data collection method, and modeling techniques. However, the amount of research in this area is still limited, partially due to the availability of suitable datasets. More specifically, the investigation into variability requires repeated measurement for the same individuals over a certain period of time. Expo tour dataset contains three waves of week-long travel data for all participants so it is ideal for intrapersonal variability study. On the other hand, trip chaining has been a key element to travel demand modeling development, from trip-based, tour-based, to activity-based model. But no research has extended the travel behavior variability involving tours to any dimension beyond tour frequency. In fact, none of the reviewed literature has specifically focused on the intrapersonal variability in trip chaining. Therefore, Component II would be the first study that examines the within-individual variability of trip chaining behavior with trip chaining characterized by complexity.

While Component II targets at intrapersonal variability of trip chaining, Component III focuses on extracting representative tour patterns, of which the logic is identical to analyze interpersonal (between individuals) deviance. The importance of representative travel-activity pattern lies in that it is the foundation to the development of transportation policy and travel demand modeling. The recognition of representative patterns enables the assessment of the connection between travel patterns and social-demographical characteristics. The relationship then provides operational guidance to further theoretical and empirical research in travel behavior. Most of the literature about travel-activity pattern focused on the activity pattern rather than travel pattern. Indeed, activity induces travel. But an investigation on travel pattern could offer information that is directly related with the transportation network. In Component III, the tendency on choosing different complexity level of trip chains is analyzed as an aspect of the overall travel-activity pattern. The process of a home-based tour is also a process to meet out-home activity needs. The activity needs determine the choice of destinations, schedules and routing of the tour. Trip chaining pattern provides a new perspective from which researchers can understand the overall travel-activity pattern. In addition, Component II and III are the first attempts about interpersonal and intrapersonal variability of trip chaining behavior that based on the urban context of southern California.

Component I: Tour Complexity and Mode Choice

Objective

To apply an expanded "tour complexity" measurement for assessing the connection between trip chaining complexity and travel-related factors through analysis of GPS-based tour data collected before and after the opening of a new transit facility.

Research Goal and Question

1. Propose a tour complexity measurement and compare it with the traditional binary classification scheme
2. Examine the relationship between tour complexity and public transit usage
3. Investigate how tour complexity is affected by travel-related factors
4. Explore how activity location influences tour complexity and mode choice

Background

A trip chain or tour is defined as a series of grouped trips involving multiple stops and activities. Most chained trips originate and end at home or work place. The intermediate stops or legs may include a change in travel mode or trip purpose. Table 1-1 presents a summary of three popular definitions on "tour/chained trip/trip chain" from the previous studies. In 2003, FHWA adopted an operational definition of "trip chain" that the duration of intermediate stops within a trip chain should not exceed 30 minutes. A stop longer than 30 minutes signals the end of a tour. Some studies use a different definition of a "trip chain" and defined that a tour should start and end at home (Jing Ma, Mitchell, & Heppenstall, 2014). In these two cases, either temporal or location constraint is applied. Though there is no universal definition of what constitutes a trip chain, it is widely agreed that trip chaining is very common. The Australia national household travel survey indicates that about half of work trips and over a third of non-work trips involve more than one stop outside the origins and the destinations (Currie & Delbosc, 2011). A Beijing study has found 52.1% of the home-based work tours have more than one leg (Jing Ma et al., 2014). In U.S., the 2001 National Household Travel Survey (NHTS) reports that 27% (18.8 million) of weekday commuters have chained trips; the report also claims that the share of trip chaining may be increasing over the years (McGuckin, Zmud, & Nakamoto, 2005).

Table 1-1 Definition of "Tour"

Definition	Research
a series of grouped trips with intermediate stops not exceeding 30-minute activity duration	McGuckin et al., (2005); Santos, McGuckin, Nakamoto, D., & Liss (2011)
a series of trips that starts and ends at home; any trips outside of home are considered as legs	Lee & McNally (2006); Bowman (2000)
a tour that starts and ends at either home or the location where the primary activity occurs, usually work place	Jou & Mahmassani (1997); Currie & Delbosc (2011)

Promoting trip chaining has been proposed as a trip reduction strategy (McGuckin et al., 2005). Consider the situation that a person has three activities planned outside of home for a day: work, grocery shopping and buying a cup of coffee. Without trip chaining, he/she needs to make three home-based tour to complete these activities: {Home - Work - Home}, {Home - Coffee - Home} and {Home - Shopping - Home}. With trip chaining, this person can minimize the total number of tours from three to one: {Home - Coffee - Work - Shopping - Home}. The total time cost and traveled distance are reduced as a result. However, does reducing the number of trips by chaining help reach the higher-priority goal of reducing VMT? By comparing commuters who chain and who do not, McGuckin et al. (2005) concludes that people who chain have higher annual miles traveled than those who travel directly between home and work. The author hypothesizes that people who chain trips tend to take advantage of the saved time from commuting tour and take more personal trips. It is still a hot topic for debate in travel behavior area.

Tour complexity level is a parameter commonly used to characterize trip chaining behavior. After being labeled with the complexity level (e.g. “simple”, “complex”), the relationship between chaining behavior and other interested variables can be assessed. In most trip chain studies, if two or more trips are inserted into a tour, it is classified as a complex tour. For instance, a tour is a simple tour if it is composed of {Home - Work - Home} and a tour is classified as a complex tour if it is composed of {Home - Work - Shopping - Home}. Several studies find that trip chains are becoming increasingly complex due the dispersion of activity clusters and the increasing value of time (Hensher & Reyes, 2006; Maat & Timmermans, 2006; Xianyu, 2013). Researchers argue that commuters tend to insert personal activities on the way to/from home in order to reach a higher living efficiency and to save the time comparing with making separated trips (Xianyu, 2013). A few studies, however, reach a contradictory conclusion that no evidence shows modern life leads to a bigger demand for complex tours. Currie et al. (2011) using data collected in Melbourne shows that between 1994 to 1999 the complexity of chained trips were quite stable and even declined slightly, by 0.025 legs per tour per year.

Three factors are most analyzed with tour complexity: 1) travel mode, 2) individual and household characteristics, and 3) trip purpose. Some studies have argued that public transportation is a less suitable mode for complex tour activity than vehicles (Hensher & Reyes, 2006). Private vehicle trips are more likely to be associated with more complex tours than public transit trips (Hensher & Reyes, 2006; Xianyu, 2013). McGuckin et al. (2005) has found that people who chain trips are significantly more likely to use private vehicle than those who do not chain. Another study has demonstrated that non-work transit tours are more complex than non-work vehicle tours. If multiple public transit modes are examined separately, train and tram chains are generally more complex than vehicle chains while bus chains are less complex (Currie & Delbosc, 2011). A broad range of social-demographical characteristics are found to be influential to tour complexity as well, such as gender, income, marital status and household structure. For example, females are found to be more likely to make non-work stops in a tour than males (Hensher & Reyes, 2006). This can be explained by the assumption that females share greater household obligations. According to 2001 NHTS data, for home-to-work tour, the most common inserted trip purpose is picking-up/dropping-off passengers (35%) while the most common purpose for work-to-home tour is shopping (36%) (McGuckin et al., 2005). The data also shows a trend that commuters tend to insert the shopping trip into the work-home tour. The inclusion of shopping and pick-up/drop-off trip thus makes a simple tour {Home - Work - Home} into a complex tour { Home - Work - Shopping - Home } or { Home - Pick-up/Drop-off - Work - Home }.

Moreover, evidence suggests that the built environment impacts tour complexity. For example, transit-based tours have higher complexity in CBD areas than suburban areas (Currie & Delbosc, 2011). The same study also points out that rail-based tours are more complex than bus-based tours because rail stations usually contain greater concentrations of activity locations. A 2003 Seattle study has found that a highly accessible retail center would lead to on average less complex tours for households living in the nearby neighborhood (Krizek, 2003). But the study also finds that households' destination choice, particularly for maintenance trips, has no significant linkage with the presence of nearby retail center. The households in this study tend to complete their maintenance trips far away from the neighborhood (Krizek, 2003).

The difference in tour complexity by travel mode is often considered to be associated with the degree of flexibility with each mode. Many researchers argue that constraints in schedule, route, station/stop and uncertainty associated with in public transit limit the level of complexity in a single tour (Hensher & Reyes, 2006; Ye, Pendyala & Gottardi, 2007). Several studies take this argument one step further by examining the causal relationship between tour complexity and travel mode. From a disaggregate level, Xianyu (2013) finds that though the relationship varies among individuals, a majority of commuters determine the tour plan prior to the choice on the mode. Another study analyzes this interdependency using a simultaneous bi-variate probit model on the 2000 Swiss travel survey data (Ye, Pendyala & Gottardi, 2007). It finds a causal structure in which trip complexity preceded the mode choice fits the sample better than the one using a causal structure that assumes mode choice as the primary determinant. The conclusion could be applied to both work and non-work tour.

Contribution

A review of the existing literature has revealed a problem in trip chaining study. That is, over-simplified method of tour characterization. Most trip chaining analyses use a single binary measurement to classify a simple or complex tour (i.e., the number of intermediate trips/legs) (Currie & Delbosc, 2011; Hensher & Reyes, 2006; Xianyu, 2013). If more than one intermediate trip presents in a tour, the tour is defined as "complex". Otherwise, the tour is defined as "simple". For example, {Home - Work - Home} is a simple tour while {Home - Coffee - Work - Home} is a complex tour. However, a {Home - Coffee - Work - Home} tour is defined as complex as a {Home - Coffee - Drop-off - Work - Shopping - Laundry - Pick-up - Home} tour. The simplified definition is most likely used in previous studies due to the limited trip information provided by the travel surveys. Most previous researches did not use the detailed spatial information obtained from the travel survey, such as route and geographical relationship between chained destinations. The GPS travel data from the Expo study enables the researcher to explore the possibility of adding new measurements variable to trip chaining and tour analysis, such as route features, stop and destination location, leg length (trip distance), activity duration, and land use diversity, etc.

In addition, Component I examines the relationship between tour complexity and travel mode, given previous studies have provided contradictive viewpoints and results. Some studies state that people are less likely to take public transit because vehicles could meet their demand on tour complexity better (Hensher & Reyes, 2006). In general, previous findings regarding work-based tours consistently indicate vehicle tours are more likely to be complex than transit tours. On the other side, several studies point out that for non-work tours, the average complexity of transit tours is higher than vehicle tour. They have found the number of average legs per chain in transit-based tours are higher than that in vehicle-based tours (Currie & Delbosc, 2011). Such discrepancies imply that the underlying assumption of a causal relationship between vehicle mode and tour complexity requires further validation.

Policy implication

The findings of the study can provide policymakers a better understanding on the relationship between mode choice and tour complexity. For instance, if the demand of a complex tour is found to be high during weekends and most complex tours are associated with vehicle tours, policymakers could promote the use of public transit by encouraging transit agency to provide more service on weekends on certain routes. In addition, the expanded measurement method of "tour complexity" could help transportation planners and policymakers to better identify the mechanism of trip chaining behavior.

Component II: A Longitudinal Analysis of Daily Travel Variability

Objective

To assess the within person day-to-day variability from the perspective of trip chaining complexity and mode choice using the longitudinal feature of Expo dataset.

Research Goal and Question

1. Demonstrate the application of the tour complexity measurement in longitudinal study
2. Validate the existence of intrapersonal day-to-day travel behavior variation with travel being jointly characterized by tour complexity and mode choice
3. Test a hypothesis that variability differs across population segments
4. Investigate the impact of a new transit service on travel variability

Background

As discussed above, trip chaining complexity is closely related with trip purpose and activities. As a result, the complexity changes over time as well. A study taken place in Adelaide, Australia, indicated that the share of complex tours decreased from weekday to weekend. On weekdays, 31% of the tours were complex; on Saturdays, the proportion dropped to 23%; and it decreased to 16% on Sundays (Primerano et al., 2007). In addition, the difference of tour complexity over the week is found to be less significant as the number of vehicles in the households increased. Thus, examining tours across one day, a week or even a sequence of years could provide a more robust way to track the schedule and travel behavior of individuals and households (Krizek, 2003). However, a formal longitudinal analysis of trip chaining behavior is rare in existing literature. It may be due to two obstacles: one, extra efforts in extracting trip chaining information from travel diary regardless report-based or GPS-based, and two, the limited availability of longitudinal dataset. There is one study that has focused on the repetition of tours occurs over a course of a week (Stopher & Zhang, 2011). The study shows a low level of repetition for both simple and complex tours, where the traditional simple/complex classification scheme is applied.

Although not directly target on trip chaining, many studies have investigated the day-to-day variability in general travel behavior since 1980s and confirmed that the amount of deviation too high to be compensated by the random sampling mechanism (Hanson & Huff, 1982; Pas & Sundar, 1995; Raux et al., 2016; Stopher & Zhang, 2011). Following it, debates been raised about the cycle of travel pattern. For instance, whether the behavior repeats daily, weekly, or monthly. It has significant meaning to clarify this issue because it affects travel survey design, model estimation and result interpretation. For example, research has found that the modeling residual is negatively related with the length of the repetition cycle selected for the model (Raux et al., 2016).

After the confirmation of the existence of intrapersonal variability overtime, the question regarding the determinants of such a variability then follows up. Individual's demographical and social-economical characteristics is often hypothesized being strongly related with level of intrapersonal variability. However, little support can be found from the reviewed literature. Hanson & Huff (1982) have found that non-employed housewives and these women's full-time employed husbands do not differ significantly in travel variability. The study by Kitamura & Van Der Hoorn (1987) finds no evidence to support the hypothesis neither with travel variability characterized by trip frequency and time allocation. However, a more recent study by Raux et al. (2016) has found that gender, age, occupation, and family size are significantly influential to intrapersonal travel variability. But similar with Hanson & Huff (1982), employment does not show significant impact on intrapersonal daily variability.

Contribution

In summary, very few trip chaining studies has used longitudinal data to examine weekly patterns of tour complexity. Meanwhile, contradictory results are found about the relationship between personal variability with social-demographical factors. A few studies have examined complexity patterns by time of day. But the scale of analysis is within a day. They have found that the complex tours are more likely to occur during morning/evening peak hours because commuters tend to insert personal stops on their way to work or back home (McGuckin et al., 2005; Xianyu, 2013). There is one available study of tour complexity pattern by day of week (Primerano et al., 2007). However, this study is based on the travel survey in another urban context, Australia, and use a relatively old data collected during the year of 1999. Component II of this dissertation examines the weekly patterns of tour complexity using recently collected data under the urban context of Southern California. Last but not least, these studies follows the traditional binary classification scheme to label tours. For home-based tour with only one stop, it is classified as "simple tour"; for other tour with more than one stops, it is classified as "complex tour". The analysis of Component II is based on an extended tour complexity definition which provides a more comprehensive examination than the conventional approach.

Krizek (2003) analyzed the relationship between neighborhood accessibility and tour complexity. The study found that people living in a higher accessible neighborhood tended to make more simple tours because of convenience of travel. So I hypothesize that the average complexity level in experimental group is lower than control group. By comparing the longitudinal data, the complexity level in experimental group is expected to decrease after the opening of Expo line. Thus, this component also compares the level of tour complexity variability across the three years from 2011 to 2013.

Policy implication

The benefit of an explorative investigation of intrapersonal day-to-day variability in trip chaining behavior is that it offers feasible guidance in tour-based and activity-based urban transportation modeling. Current practice contains a shortcoming of neglecting the day-to-day variability in travel demand. Travel analysis based on multi-day survey data provides richer temporal information for travel and other behavioral responses to policy change. It also offers a better approach to evaluate the impact on travel behavior by the change of supply side, such as enhanced non-motorized travel facilities.

Component III: Recognition of Representative Travel-Activity Patterns Based on Tour Complexity

Objective

To identify weekly travel patterns and their association with population segments by characterizing travel pattern based on tour complexity via Sequential Alignment Method (SAM) and discrepancy analysis.

Research Goal and Question

1. Extract representative weekly travel patterns based on trip chaining complexity

2. Investigate personal and family characteristics that are influential to the weekly pattern
3. Experiment with the method of discrepancy analysis and regression tree

Background

The focus of Component III is expanded to pattern extraction. In line with Component II, the underlying assumption of Component III is that travel-related theories and practice that is developed based a single-day measurement could be biased due to a substantial amount of periodical variability within each individual. Therefore, it is important to study the travel-activity pattern using datasets that contain multi-day travel records, which is the major advantage of Expo tour dataset.

The existing research in travel-activity pattern suggests a two-step analysis framework is most appropriate. The first step is representative pattern extraction. There are two general approaches: 1) to study the variables and measurements that are extracted from the observed travel-activity and then to perform principle component analysis based on these variables (Hanson & Huff, 1986; Jun Ma & Goulias, 1997); 2) to study the observed behavior as a feasible analysis unit and directly analyze the similarity/dissimilarity between observations (Recker et al., 1985). Component III follows the latter. Questions addressed in the first step include how many typical patterns in the sample and what the pattern characteristics are. The second step is often performed with the purpose of relating interested exploratory variables with the extracted representative patterns. The methods include descriptive analysis, ANOVA and contingency table (Recker et al., 1985; Saneinejad & Roorda, 2009; Wilson, 2001). Questions answered in the second step include whether individual characteristics are strongly associated with the travel-activity pattern, and if so, what the patterns look like for certain population groups. The second step is not a necessity as the primary goal of pattern recognition is completed in the first step. However, it is valuable in that the linkage between pattern and social-economical features provides important information for transportation planners and policy makers.

Because travel-activity pattern is a subject so complicated that it is impossible to measure using a well-established method. The methodology to extract pattern becomes an area that most research effort is focused on. Early research has explored and presented some pattern representation schemes that rely on visualization, such as Hägerstrand trajectory (Hagerstrand, 1970) and space/time trajectory (Recker et al., 1985). However, such diagram-based methods could incorporate limited dimensions of travel information, i.e. time and geographical location. It would become inconvenient and computationally complicated when more travel aspects are added in. As a result, a few researchers start to seek alternatives and Sequential Alignment Method (SAM) is one of the promising methods. Originally developed in bioinformatics field for DNA and protein study, SAM is first introduced into travel behavior field by Wilson (1998). The researchers have found that the observed travel events and activities can be coded as a sequence (or "trajectory"). The advantage of SAM is it can be simple with a single attribute coded and can be complex with multiple attributes, depending on the research needs. But one of the limitations of SAM is that it requires categorization and results in loss of information. Another widely-mentioned limitation of SAM is the parameter setting in distance measurements, such as weighting, operational boundary, computational cost, and etc. (Joh et al., 2001). As a result, most of the reviewed studies choose the uni-dimensional approach with a default setting in distance computation, which is discussed in Chapter 5 (Kim, 2014; Wilson, 2008). Nevertheless, SAM has been demonstrated to be superior than other alternatives in pattern recognition (Joh et al., 2001). Hence, Component III continues to use SAM for dissimilarity measurement.

Contribution

Component III is designed as an extension of Component II which shifts the focus from intrapersonal variability in complex trip chaining behavior to interpersonal dissimilarity. Meanwhile, it is not difficult to see that all the relevant studies have focused on activity patterns although they are designed to be beneficial for both activity and travel demand research. Activity and travel behavior are different concepts because activity could be home-based while travel is explicitly related with needs outside of home. Component III directly focuses on the travel behavior which offers a new perspective from which travel behavior and travel demand can be studied. Moreover, the reviewed literature often studies the relationship between pattern and social-demographical factors in a posterior way, which is lack of generality from the perspective of modeling. Component III incorporates social-demographical factors into the clustering process via a tree-structured regression method developed based on discrepancy analysis. Therefore, it results a more feasible implication for model developers.

In addition, previous research efforts have been put on representative pattern recognition based on activity types/trip purpose on a daily basis (Joh et al, 2002; Kim, 2014; Saneinejad & Roorda, 2009; Wilson, 2001, 2008). Trip purpose, together with other traveling context including other people presented, is one of the limitations of GPS-based travel behavioral study. Because GPS data is passively recorded and is not context-aware. The data usually does not come with trip purpose. But such data is becoming more and more popular in behavioral research and it costs much less to collect longitudinal travel information using GPS. Hence, it is important to explore analytical framework that could fully utilize GPS data and does not rely on contextual information such as trip purpose.

Policy implication

By extracting typical patterns from observed weekly travel trajectories, Component III has important implication for further theoretical research and the development of tour-based or activity-based travel demand modeling. Meanwhile, association between the travel-activity patterns with population segments is a critical subject for model developers and urban transportation planners. Because the two-step analysis simplifies the complex human travel behavior based on quantifiable population segments and thus, provides an operational guidance for model development and policy assessment.

Chapter 2 Expo Tour Dataset

Introduction

The data used in this dissertation is a subset of the longitudinal Global Positioning Systems (GPS) data collected as part of the quasi-experimental before-and-after evaluation study of the travel behavior impacts of Los Angeles Metro's Expo light rail line on nearby residents (Figure 2-1). Personal 7-day GPS travel records were collected for the same study subjects in 2011 before and in 2012 and 2013 after the Expo line began service in early 2012 (Figure 2-2). This dissertation analyzes GPS-based travel patterns for a subsample of 55 Expo study subjects for whom complete 7-day movement tracts are available for each of the three study periods.

The use of these longitudinal GPS-based data for analysis of trip making and tour complexity has two major advantages. First, spatial and temporal information is directly available from the GPS record, such as activity location, departure/arrival time, and route used. For this reason, the resolution of the analysis can be higher than that of traditional travel surveys. Moreover, the travel statistics retrieved from such passively recorded trace of each participant's movement and can therefore provide relatively more objective and reliable data compared to data collected using conventional travel diaries for which travel information is manually reported by travelers themselves. It is widely agreed that the travel diaries often contain errors and they are hard to correct due to lack of "ground truth". In addition, the type of misreporting varies across individuals making it is also difficult to apply systematic correction to the factor to the data (Houston, Luong, & Boarnet, 2014). To the contrary, any suspicious or unclear patterns in a dataset created from GPS data can be always reviewed and corrected based on the original trace.

The second major advantage of the dataset is its longitudinal aspect. The physical movement of each participant is tracked for seven consecutive days during each wave of Expo survey. Hence both weekday and weekend travel behaviors are sampled. A single individual can have a mixture of different activity patterns across day of week. Some patterns may occur on a daily basis, such as commuting between home and work place, while some patterns may occur on a weekly basis, such as maintenance trips undertaken during weekends. Therefore, the multi-day travel data is more likely to capture a larger pattern set for each individual in the sample. Furthermore, the dataset contains repeated measurement for each participant from 2011 to 2013 during a time in which a major event occurred (the opening of Exposition Line in April, 2012), which enables analysis of the impact of neighborhood environment on travel patterns.

There are also drawbacks in this dataset. The first disadvantage is the small sample size. It not only limits the generalizability of the findings but also constrains the complexity level of potential modeling techniques. Considering a hypothetical mixed effects model at the home-based tour level applied to all three waves' data, the model would need to control for the correlation between tours made by the same person and also control for the dependency between tours occurred during same wave. In addition, if the response variables and independent variables are a mixture of numerical, categorical and binary data types, which is usually the case in travel behavior research, extra care would need to be taken in the choice of modeling method. Under this setting, a complex model often fails to converge. Moreover, the validity of distribution assumptions is hard to assess for a small sample. However, the limited sample size reduces the burden of data processing to a great extent and in turn allows more careful examination through all aspects of the dataset.

The second drawback of this GIS-based dataset is that trip purpose information, a common focus for most of the travel behavior and activity pattern studies, is unavailable from the Expo Study and cannot be included in this dataset. Despite this limitation, this dissertation makes important contributionsin demonstrating the effectiveness of GPS data in the field of travel behavior research to examine for analysis of trip making and tour complexity. Lack of participant reported trip information, such as trip purpose and number of people involved in the trip, is a common challenge in studies using passively-recorded data. In addition to the reliability (discussed above), GPS data are generally easier to obtain and cheaper to process nowadays than traditional travel survey data (which usually collect trip purpose information).

The following sections in this chapter provide an overview of the Expo Line before-and-after Study, followed by a detailed discussion regarding dataset design and construction. Some descriptive statistics for the study sample are presented in the last part of this chapter.

Expo Study

The data used in this dissertation is a subset of the longitudinal GPS data from Expo Line before-and-after study. Collected by USC and UCI researchers, the Expo data set was designed as a longitudinal, before-and-after study of public transit and travel behavior in the Los Angeles urban area. Personal travel data were collected before and after the opening of the Expo Line. The 7-day travel records, including a paper-based trip and mileage logs (not used in the research) and GPS mobile tracking data (used in the research), were collected for one week in each of the consecutive three years at a similar time period (Figure 2-2).

Figure 2-1 Metro Rail System Map

Source: http://www.metro.net/projects_studies/exposition/images/expo_ph1_fact_sheet.pdf

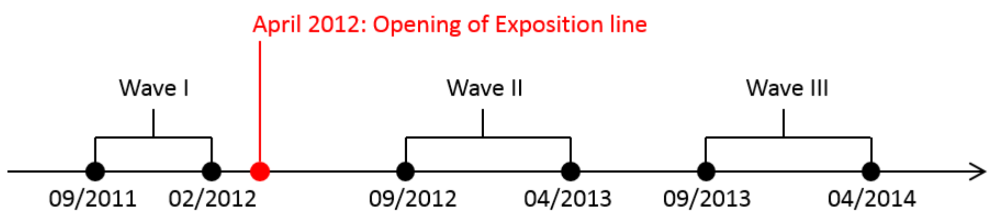


Figure 2-2 Timeline of Expo Study Data Collection

Phase 1 of the Expo light rail line is located to the west of Los Angeles downtown (Figure 2-1) and service began in April, 2012 (Figure 2-2). This Phase 1 segment has ten new stations plus two existing stations. The neighborhoods near the six stations of the eastern portion of this segment were excluded to avoid the bias from the existing service of Metro Blue light rail line and Metro Silver rapid bus line. The three stations closest to the University of Southern California campus were also excluded. The reason is that the residents of this area are mainly comprised by university students whose travel behavior pattern could not be generalized to the non-student residential groups along the corridor (Boarnet et al., 2013). As a result, approximately half of the Expo study participants lived within ½ mile from the six Expo stations on the western portion of this segment. The other half of participants lived from ½ mile to 3 miles from these stations (Figure 2-3). The neighborhoods in the service area of Expo line are predominantly lower income and minority, with moderate income neighborhoods in the Culver City (Boarnet et al., 2013).

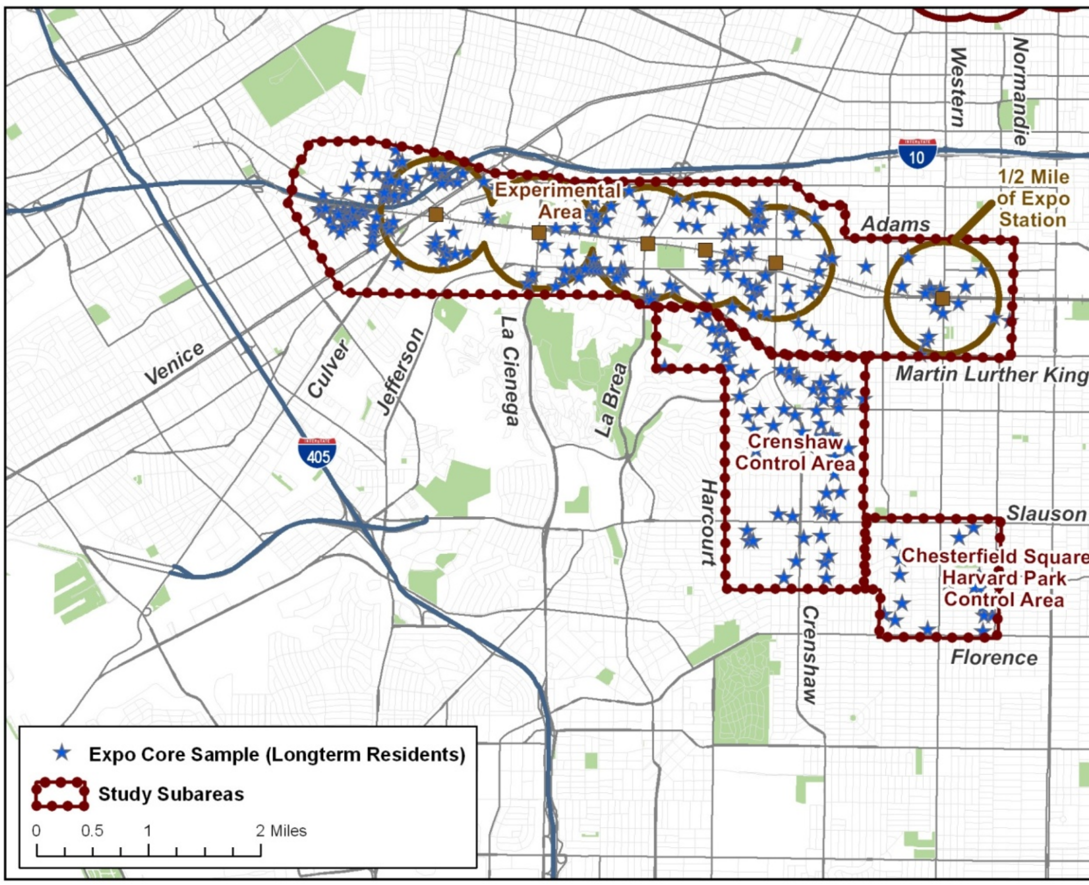


Figure 2-3 Expo Study Core Sample Approximated Residential Location

Source: Boarnet et al. (2013)

For each wave, the participating households were separated into three subgroups corresponding with different survey instruments based on their willingness to participate based on initial screening questions: web-based, paper-based and mobile tracking. This dissertation analyzed data from the mobile tracking group, and for households in this group one adult agreed to wear a GPS tracking device for seven days whenever they left their home, and their geographical locations were recorded every 15 seconds. The sample for this dissertation was selected from the mobile tracking group, and the analysis is based on their GPS tracking data (not the travel and mileage logs).

Study Subject Selection

One of research objectives of this dissertation is to examine whether and in what ways trip chaining patterns change from Wave I to Wave III. Therefore, the sample selection criteria for the study is that the individual should participate in the GPS tracking group in all three waves and should take at least one trip during each wave. There are 138, 100 and 80 usable individuals in the GPS mobile tracking groups of Wave I, II and III, respectively. These individuals have at least 8-hour daytime GPS records per day for 3 days in each wave. 58 people participated in all the three waves and 55 of them provided valid weekly travel records. Among them, 27 participants live within 1/2 mile from the Expo Line stations (the "experimental" group), and 28 live further than 1/2 mile from the stations (the "control" group). These 55 participants are selected as the analysis sample for this dissertation. The reason to keep the consistency in samples through all the waves is to eliminate the possibility that the changes in trip chaining patterns are caused by different samples.

The data collection period ranged from September to February for each wave (Table 2-1). Over half (56%) of the participants completed the survey during Wave I in November and December. For Wave II and Wave III, the majority of the participants (71% and 62%) completed the survey in September and October. No participants were monitored during the holiday weeks including Thanksgiving, Christmas and New Years in any of the waves.

Table 2-1 Distribution of Study Subject by Month of GPS Data Collection

	Wave I (2011)		Wave II (2012)		Wave III (2013)	
Month of Collection	**Count**	**%**	**Count**	**%**	**Count**	**%**
Sep. – Oct.	23	41.82%	39	70.91%	34	61.82%
Nov. – Dec.	31	56.36%	14	25.45%	21	31.18%
Jan. – Feb.	1	1.82%	2	3.64%	0	0
Total	55	100%	55	100%	55	100%

Database Design and Construction

Relational Database Structure

A relational database with multiple linked data files was developed to accommodate the complex nature of travel information. The database contains four components: (1) participants' demographic and social-economic data, (2) trip and tour files with basic statistics aggregated from the original GPS trace data, (3) the original trace files in which each row represents one spatial-temporal stamp of a trip, and (4) location files for the activity locations visited by the participants during the survey period, public transit stops/stations and city-wide point of interests. These tables are linked together using the unique traveler ID, trip ID and location ID.

In the original Expo study, a randomly-generated four-digit household ID (HID) is assigned to each participant. Using HID, we can access the participant's demographic characteristics, including gender, age, educational level, employment status, household income, and travel preference. Each trip is assigned a unique ten-digit trip id (TID) in the format of *hhhhhwwttt*, where *hhhhh* = participant id, *ww* = wave id, and *ttt* = trip id. For instance, *HF5C3W2T04* represents trip

04 by participant F5C3 during wave II. Via TID, we can reach the trip's trajectory data with geographical coordinates and time stamps recorded for every 15 seconds. Furthermore, each place visited by the participants during the three waves is assigned a unique location ID (LID). Using LID, we can access the information for each place, including the geographical longitude/latitude, location type (home, transit stop or regular place) and other features, such as land use and network accessibility.

Raw GPS Data Processing

The most tedious and the most important work in the construction of the Expo Tour database is to classify of GPS data into periods in which each participant stayed at locations or traveled (by travel mode). The raw data exported from GPS device included four variables: timestamp, longitude, latitude and speed. The GPS device recorded its wearer's location every 15 seconds and the tracking lasted 7 days 24 hours a day for each wave. Therefore, each participant's data file contains approximately 40,000 records. The devices, however, do not directly record the travel status and travel mode associated with each time point. Hence, a two-step data classification procedure is implemented.

The first step is an automated procedure to identify a rough estimate of the travel status of each point. The algorithm includes two parts. First, all the points are sorted into two categories, staying and moving, based on the speed associated with the points. Then, clusters of points representing the locations where participants visited or occupied are identified if sequential GPS points exceeding a 2-minute duration are clustered within a circle of 20-meter diameter (Houston et al., 2014). If a cluster of points is identified, all points within the specified threshold distance (20 meters) to the cluster center with speed lower than 2 mph are labeled as a location ("LC") indicating the participant was staying at this location. On the other hand, if a series of points show a linear pattern, then the points are labeled as a travel status code indicating the participant was moving. The travel status code is assigned based on modes. If the corresponding speed is 2 - 6 mph (a typical walking speed), the point would be labeled as "WLK" (Walking). If the speed is over 6 mph, the point was labeled as "VDR" (Vehicle Driving). This first phase of classification could only distinguish between staying, walking or moving in a vehicle at a relatively low accuracy level.

In the second part of the classification process, research assistants reviewed all GPS tracking files in ArcGIS and manually corrected and sub-categorized the status of each GPS point using series of decision rules. The travel modes, such as by bus, by train/light rail and by vehicle, are determined and labeled as “VBS” (Vehicle Bus), “VTR” (Vehicle Train) and “VDR” (Vehicle Drive), respectively. Though the moving speeds are often similar (~20-50mph), the three different motorized travel modes can be distinguished by examining the location patterns of previous and subsequent location points. For instance, if both trip head and tail involve a period of walking and a short staying at certain location near the road as opposed to in the center of a block, then researchers verify the locations. If the transition location is identified as a transit stop and the subsequent route matches a transit route, the trip is corrected as a bus (“VBS”) or light rail (“VTR”) trip. The walking mode is sub-categorized into two types based on the purpose of walking. If the walking starts/ends at a transit stop/station, these points are labeled as "W2T" (Walking to Transit). If both the origin and destination points of the walking trip appear to be the activity locations, the walking points are labeled as "WLK" (Walking). Figure 2-4 shows an example of a typical GPS trajectory map with the travel status classified.

Points misclassified during in the first the classification phase are corrected during this second phase of the classification process. For instance, vehicle stopping periods (such as waiting for the traffic signal at intersections, or a traffic jam on a freeway) could have been misclassified in the first phase as a cluster and a location or destination are corrected to be “VST” (Vehicle Stopping). However, one limitation of the classification procedure is that we cannot tell whether the GPS device wearer is traveling alone or not. This is a major drawback particularly for vehicle trip study considering it is useful to know whether the traveler is driver or passenger. Furthermore, we cannot make inferences on the interaction between household members.

Admittedly, some potential classification errors may exist. One error source derives from the geo-coding process. About 15 voluntary assistants were involved in the manual hand coding part over a three-year period. To minimize discrepancies, each assistant was individually instructed on the status classification process by a senior researcher before they start hand coding. Despite training, the accuracy of the travel status identification might still suffer from inconsistent personal decisions. Therefore, after coded by voluntary assistants, all the trace files were reviewed carefully by the author of this dissertation to ensure consistency of coding assumptions and to correct obvious mistakes and discrepancies. Nevertheless, the status identification of a travel trajectory is still a subjective result; but the bias is considered as relatively more consistent across travel profiles. Another possible source of potential error is the positional accuracy of the GPS signal. In general, the signal quality is very good and the precision level is around 10 meters. For example, it is visually distinguishable in ArcGIS whether a participants was located on the northbound or southbound of the road. "Bouncing" points that are off the main travel path exist in some traces and these points usually were seemed to suggest a participant traveled at an unreasonably high speed (e.g. 150 mph). In addition, the transmission of the GPS signal could be disturbed by tall buildings and tunnels. The tracking precision in the Los Angeles downtown area was somewhat limited. Sometimes the GPS tracking device would lose signal for several minutes, even hours. The decision for how to handle above two cases in the analysis sample is discussed in the section of Missing Data Handling.

Although some minor discrepancies may exist in the analysis dataset during periods of transition from being a given location and initiating travel, the travel mode classification during periods of travel is considered as very reliable. The difference between walk, bicycle, vehicle, bus and train travel has been carefully reviewed and verified by the hand-coders and researchers. First, walking and bicycling have a much lower speed. The consecutive points are closer to each other when they are plotted on map. Second, train/light rail trip has a speed much higher than bus and more stable than private vehicle, and it starts/ends at rail stations. Bus trips usually follow and are followed by a walking trip and show periodic stopping-moving patterns in the middle of road segments. Private vehicle trips do not involve lots of walking at two trip ends and typically do not have frequent stops during the trip.

As previously discussed, one major disadvantage of the analysis dataset is that the activity purpose associated with each trip is not recorded on the supplemental paper travel logs, which only tabulated today daily trips by mode and did not record trip or location specific information. Given this limitation, the relationship of trip purpose to trip and tour complexity patterns is not examined in this dissertation.

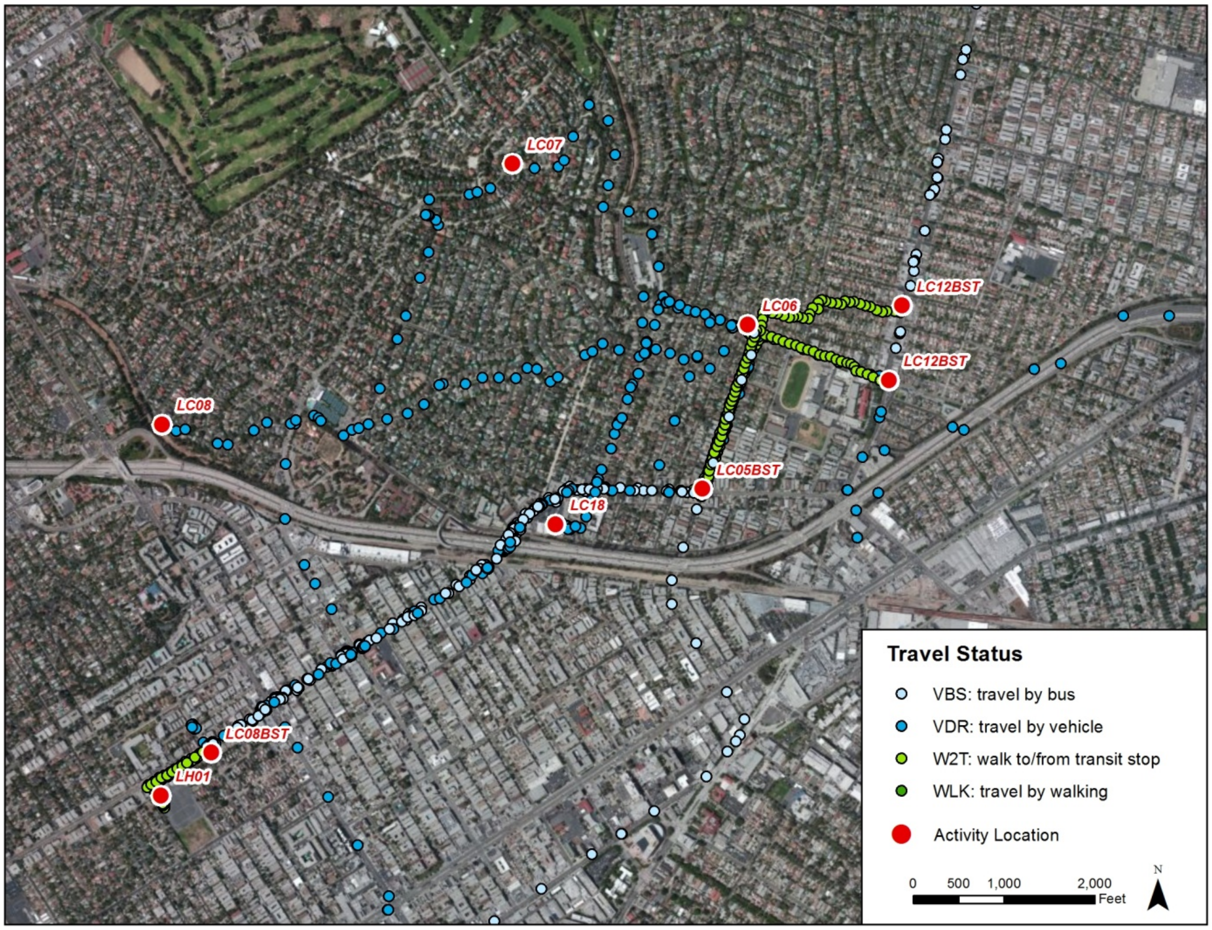


Figure 2-4 Illustration of Status Classified Travel Trajectory

Definitions and Issues in Trip Identification

This section clarifies how the trip chaining terminologies have been covered in Chapter 1 were operationalized in the analysis. Firstly, the characteristics of the destination location of a trip is defined as the location's corresponding census block level, which is the smallest geographical unit with demographic data. Two sequential locations or stops within the same census block are melted into one trip with one destination, even when there is observable moving trajectory from one part of the block to another part. This rule not only assures data availability for each destination but also simplifies some situations such as shopping activity at a large block. Secondly, tours as defined for this analysis always start and end at home while individual trip segments always start at one destination and end at another.

The procedure to determine whether a stop qualifies as a "destination" is as follows. If the idle time between two trips exceeds 3 minutes, the location where this idle period occurs is considered as a destination. If the idle time is shorter than 3 minutes, a case-by-case decision is applied to determine whether it should be classified as a destination. Generally, such a shorter idle period is classified as a "destination" if it affects the routing arrangement, which is determined by visual inspection. For instance, trips to pick-up/drop-off someone usually only include a 1-minute stop at the pick-up/drop-off location. Such stops serve a legitimate travel purpose and often require a distinct detour that is distinguishable using GIS visualization. Most of the previous studies applied a strict 3-minute rule, which could result in an underestimation of the complexity of trip chaining behavior. In addition, the location of a mode change (such as W2T to VBE) is not considered as destination.

Several ambiguous yet not uncommon situations are encountered during the process of determining whether a trip or a tour is valid for this study, and are summarized in Table 2-2. The last two scenarios described in Table 2-2 involve missing data (discussed in the next section). In addition, if a traveler has too many missing data or too many invalid tours during any wave of the Expo study, this traveler is removed from the analysis dataset. Due to this reason, 3 of the 58 individuals who participated in all three waves are excluded from further analysis.

Table 2-2 Summary of Ambiguous Scenarios and Decisions in Data Processing

Scenario	Decision
Visiting gas station	Labeled as a trip
Suspicious working trip where travel could be related with the work/business, usually associated with abnormally high trip frequency and repeated for more than one day	Kept in the dataset
Travel speed is too high for walking but too low for vehicle travel	Labeled as biking trip
Abnormally long stopping at a bus stop proximity in the middle of a bus trip	Kept as part of the current bus trip
Missing or drifting data points during a trip	Kept if two trip ends can be clearly identified; otherwise removed
Missing or invalid trips within a tour	Tour, including other valid trips in the tour, is removed

<u>Missing data handling</u>

Missing GPS data points occurs often in the sample data and for many participants. Missing data can be caused by actions of participants such as turning off the device or movement in built environments (such as concrete and steel structures) that obstruct GPS device communication with an adequate number of satellites, by temporary lapses of GPS instruments, or by the research team who post-processed the data. Periods of missing data can be categorized into 5 cases based on causation and consequence (Table 2-3).

Case A: Missing data periods during travel may result in trip destinations being missed by data processor. In this case, tour estimates would be retained but the number of legs for the tour would be underestimated. Given these instances can not be identified, it is hard to estimate about how many such cases exist in the data. A special case is the bus transition. A bus transfer could be missed if the transfer occurred at the same drop-off location as the previous segment and the time gap was short. This could lead to an underestimation of tour complexity for transit-based tours.

Case B: Trips with bad GPS signals are removed if the number of points "drifting around" a probable path (due to GPS positional errors caused by factors such as an obstructed satellite connection) is too high to compute reasonable trip statistics (such as average speed, etc.). The tour including this type of missing data was removed. This could result in an underestimation of a person's trip and tour frequency. The frequency of such case can be tracked.

Case C: Incomplete trips are those for which there is not a clear departure/arrival time/location, and trips containing a missing trace segment such that the entire route could not be visually inferred. These cases usually include a clear time gap in the GPS data. In these cases, both the trip and tour are removed from the analysis.

Case D: Complete trips with either ends located out side of LA county are excluded. We know how many such trips are there. In these cases, both the trip and tour are removed from the analysis.

Case E.1 If two consecutive and complete trips have unmatched origin and destination, we know there was a missing trip but cannot identify when and how many this situation occures. In these cases, both the trip and tour are removed from the analysis.

Case E.2 If the trace points before and after missing GPS data are at the same non-home or home location, we do not know what happen in between, but if a trip or tour occurred during this time it will not be identified in the data.

Table 2-3 Consequence of Missing Data

Case ID	Aware of trip missing	Know count	Trip removed	Aware of tour missing	Tour removed
A	No	No	-	-	-
B	Yes	Yes	Yes	Yes	Yes
C	Yes	Yes	Yes	Yes	Yes
D	Yes	Yes	Yes	Yes	Yes
E.1	Yes	No	-	Yes	Yes
E.2	No	No	-	No	-

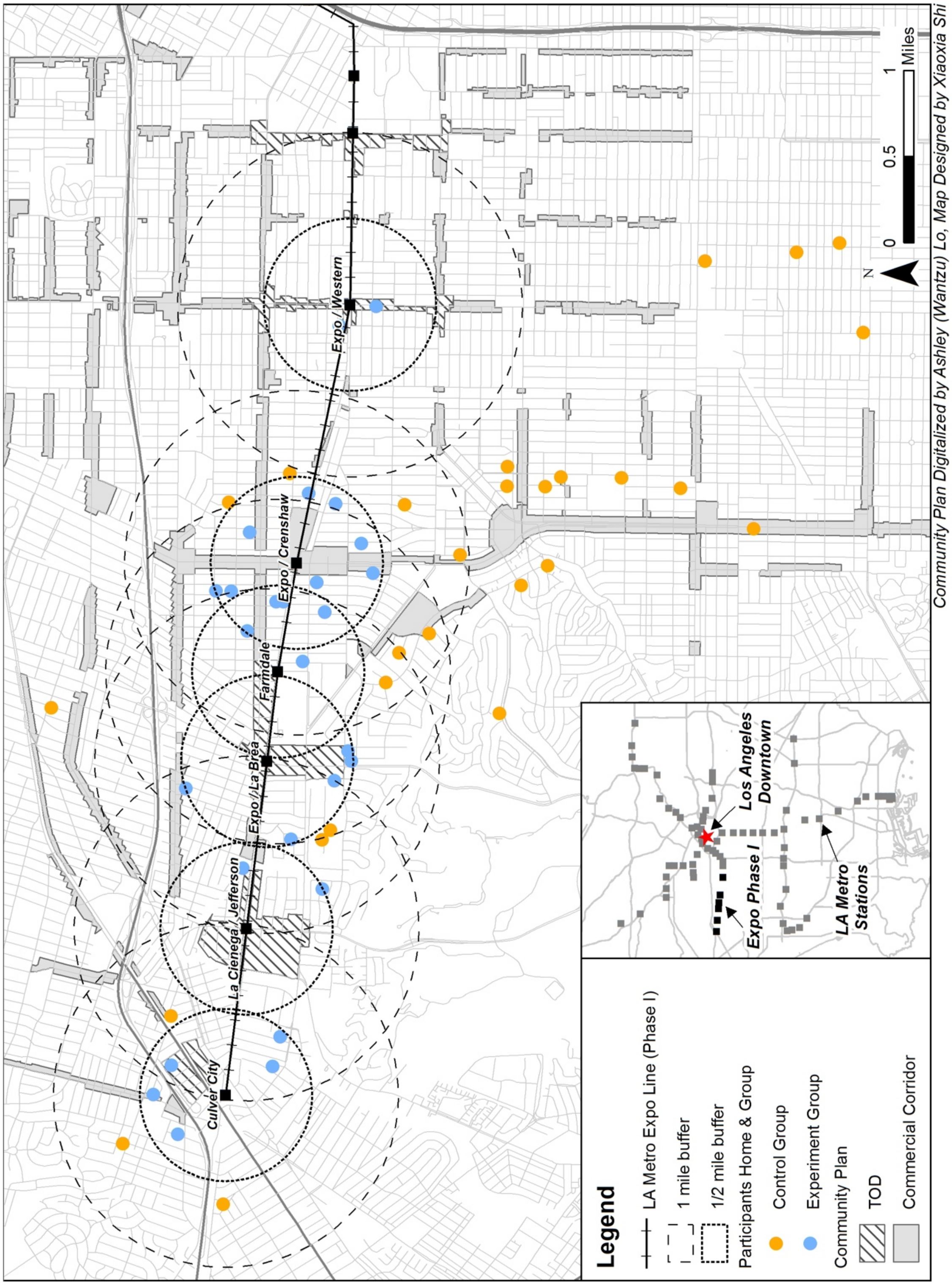


Figure 2-5 Study Subjects Approximated Residential Location

Descriptive Statistics

This section provides an overview of travel patterns of the study sample in the order of three levels: traveler, travel pattern, and activity location, and provides a comparison with 2000 California Household Travel Survey (CHTS) data at the trip level.

Traveler

<u>Sample Composition</u>

The selected participants' approximate residential locations are shown in Figure 2-5. Their social-demographic data is summarized in Table 2-4. 27 participants resided within ½ mile of an Expo Line station and are defined as the “experimental” group because they were the most likely to be affected by the introduction of the “treatment” (the new Expo Line Service). 28 Participants resided beyond ½ mile of and Expo station in areas with comparable built environment characteristics, and are defined as the “control” group. See Boarnet et al. (2013) for additional detail on the quasi-experimental research design of the Expo Line study.

The personal variables, such as gender, age, ethnicity and educational level, are based on the original data collected during Wave I, 2011. At an individual level, 38 (69%) of the sampled individuals are female; 30 (55%) are between 20 to 54 years old; 35 (64%) are African-American and 12 (22%) are white; 24 (44%) hold a Bachelor’s degree or a higher degree; 32 (58%) are employed full-time or part-time. Comparing to the county's social-demographic statistics (US Census Bureau, 2011), the study samples has a higher composition of women, African-Americans, people above middle-age and people with advanced degrees. It suggests the findings of this study might be biased by the composition of the samples and should not be generalized to other areas.

Table 2-4 Social-demographical Statistics of Study Subjects

	Sample (Wave I, 2011)		Los Angeles County (2011)
	Count	Percent	Percent
Grand Total	55	100%	
Gender			
Female	38	69%	51%
Male	17	31%	49%
Age			
20-34	8	15%	23%
35-54	22	40%	28%
55-64	13	24%	11%
65+	12	21%	11%
Ethnicity			
Asian	5	9%	13%
Black	35	4%	8%
Hispanic	1	2%	48%
White	12	22%	28%
other/multi	2	4%	3%
Education			
less than 12th grade	3	5%	24% [a]
High school or equivalent	3	5%	21% [a]
some college or associate	23	42%	26% [a]
bachelor	15	28%	19% [a]
post graduate	9	16%	10% [a]
NA	2	4%	-
Employment/Study Status			
Full-time employed	16	29%	52% [b]
Part-time employed	16	29%	
Full-time student	1	2%	48% [b]
Part-time student	1	2%	
Not employed/student	21	38%	

[a] *Percentage of population over 25-years-old.*
[b] *Percentage of population over 20-years-old.*
Data source: U.S. Census Bureau, 2011 American Community Survey 1-Year Estimates

Table 2-5 Household Demographical Statistics of Study Subjects

	Sample (Wave I, 2011)		Los Angeles County (2011)
	Count	**Percent**	**Percent**
Grand Total	55	100%	
Household size			
1	23	42%	26%
2	14	26%	27%
3	8	15%	16%
4+	10	18%	31%
Car ownership			
No Licensed Driver (LD)	5	9%	10%
At least 1 LD			
No car	5	9%	
< 1 car per LD	9	16%	90%
>= 1 car per LD	36	65%	
Family member under 18			
0	41	75%	64%
1	9	16%	37%
2+	5	9%	

Data source: U.S. Census Bureau, 2011 American Community Survey 1-Year Estimates

At the household level (Table 2-5), the percentage of single-person family households (42%) is much higher than the county average (26%); 10 (18%) participants have a family of size of 4 or more, less than county average (31%). A larger household size means a stronger interdependency on travel arrangement among family members. For vehicle accessibility, 10 (18%) participants are not licensed to drive or do not have family-owned vehicle; 50 out of the 55 participants are a licensed driver (82%), 36 of whom live in a household with more than one car per driver. A vehicle ownership per licensed driver rate over 1 indicates the individual's travel arrangement will likely not be constrained by vehicle availability. The car ownership rate in the sample is lower than county average (82% v.s. 90%). The presence of young family members under age 18 implies the household has the travel demand for school and associated trips. 75% (41) of the participating households do not have children under age 18. The percentage is about 10% higher than the county average (64%). 5 participants have two or more young children in their families. In summary, the study sample contains more single-person households, fewer participants with families, and a lower car ownership than the county average.

Participants and Expo Stations

Expo Line Phase I includes 12 stations, 10 of which are newly constructed and 2 are of which were existing stations that have already served other Metro lines for over ten years (the Pico and 7th St./Metro Center stations). In addition, the neighborhoods that are near 4 of the 10 new stations (LATTC/Ortho Institute, Jefferson/USC, Expo Park/USC, and Expo/Vermont stations, are excluded from the Expo Study due to possible bias caused by college student's non-representative travel patterns. Therefore, the 55 participants analyzed in this dissertation were sampled from areas in proximity to the 6 new stations on the western portion of the Phase 1 line segment (Figure 2-6). The Culver City Station represents the western edge of study area while the Expo/Western Station represents the eastern edge of study area. The Expo/Crenshaw Station is located along the north-south Crenshaw Corridor, which includes many commercial and civic destinations and is expected to be a major attraction for daily activities.

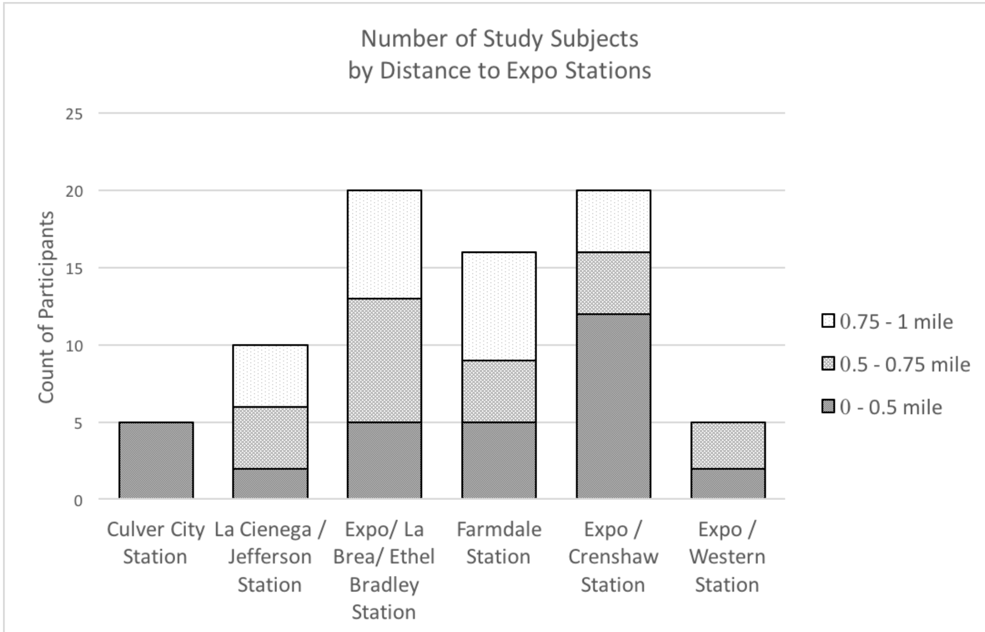


Figure 2-6 Number of Study Subjects by Distance to Expo Stations

As presented in Figure 2-6, the number of participants who live in the immediate proximity of each study area station varies. For a 0.5-mile buffer, Expo/Crenshaw Station can be reached by 12 sample participants, much higher than for other stations. For a 0.75-buffer, Expo/Crenshaw Station still has the largest number of participants (16) but is closely followed by Expo/La Brea Stations which has 13. The same number of sample participants, 20, live within 1-mile buffer to either of these two stations. Farmdale Station ranks third in terms of number of participants who live nearby. Moreover, all the five participants closest to Culver City Station are within 0.5-mile of the station. Note that none of the stations have more than half of the total sample participants within 1 mile.

From the perspective of participants, some of them are close to more than one station, which suggests greater flexibility in accessing transit system. Among the 27 participants who live within a 0.5-mile buffer of any Expo new stations, four of them reside in buffer areas overlapped by two stations. More participants can be connected to the stations and more are served by more than one station, if we expand the buffer size. For instance, for 0.75-mile buffer, 16 people reside near two or three stations. However, there are 16 participants that are not close to any of the new Expo stations (at least 1 mile away) (Table 2-6 and Table 2-7).

Table 2-6 Distribution of Study Subjects by Distance to Expo Stations from Home, N=55

Distance to Expo Station	Count	Percent
<= 0.5 mile	27	49.09%
> 0.5 mile	28	50.91%
Total	55	100%

Table 2-7 Distribution of Study Subjects by Number of Nearby Station at Home, N=55

Number of Nearby Station	Frequency of participants		
	0.5 mile buffer	0.75 mile buffer	1 mile buffer
0	28	18	16
1	23	21	12
2	4	15	17
3	0	1	10

Travel Pattern

Travel Frequency, Distance and Duration

Table 2-8 provides a summary of travel statistics for the study sample aggregated for all three waves by sample subgroups. Overall, the sample of 55 takes 4,210 trips and made 1,122 tours during the weeks surveyed in Wave I, Wave II, and Wave III. Note that the actual number of trips and tours is higher because some are removed due to missing GPS trace data (described above). On average, each person takes 0.97 tours, 3.64 trips, spends 48.95 minutes on travel and travels 12.71 miles per day. Female participants chain 0.3 more trips per tour and travel 2 miles more than male participants on a daily basis. However, the tour frequency and travel duration are similar by gender. Employment status apparently seems related to travel behavior. Employed participants or those who are students, regardless full-time or part-time, travel more frequently (1.03 vs. 0.88) and take longer trips than unemployed people (15 vs. 9 miles; 56 vs. 38 minutes). These participants also chain trips at a slightly higher rate. Participants living with household members under age 18 show higher travel demand and more complex chaining patterns than participants who live alone or only with adults. Although single participants chain a little more, they have the lowest daily tour frequency, travel distance and travel duration among three groups. Driver license and accessibility to vehicle are positively related with travel frequency, travel magnitude and trip chaining. A licensed driver in the sample with a household vehicle on average travels 8 miles more than a non-driver (14 vs. 6 miles), but he/she does not have such a great difference from other participants for daily travel duration perhaps due to the lower speed efficiency of the chosen travel mode among non-drivers, such as walking or bus. In Wave II and Wave III, participants who live within ½ mile from any Expo stations have lower daily travel demand and magnitude, but they chain as much as people who live farther from an Expo station.

Table 2-8 Travel Statistics by Population Segmentation, Wave I – III, N=55

				Daily average per person		
	total # of trips	**total # of tours**	**avg. # of legs/tour**	**tour frequency (#)**	**travel distance (mile)**	**travel time (minute)**
Overall	4,210	1,122	3.75	0.97	12.71	48.95
Gender						
Female	2,974	776	3.83	0.97	13.38	49.05
Male	1,236	346	3.57	0.97	11.21	48.72
Employment/study status						
Employed/student	2,813	732	3.84	1.03	14.96	55.60
Not employed/student	1,397	390	3.58	0.88	9.06	38.19
Household size						
Single family	1,484	396	3.75	0.82	9.87	42.66
All-adult family	1,416	391	3.62	1.03	12.79	49.77
Family with under 18	1,310	335	3.91	1.14	17.25	58.24
Vehicle accessibility						
No car/driver license	500	146	3.42	0.70	5.93	40.48
Own license and car	3,710	976	3.80	1.03	14.21	50.83
Group						
Experiment	1,952	520	3.75	0.92	11.63	48.73
Control	2,258	602	3.75	1.02	13.75	49.16

Mode Usage

Analysis of mode choice at a trip-level shows that mode choice deviates across population groups (Table 2-9). Overall, 82% trips taken are by vehicle, with 12% by non-motorized (mainly walking) and 6% by public transit (mainly bus). Each participant tends to take 3 vehicle trips, 0.4 non-motorized trips and 0.2 public transit trips per day. Broken down into population groups, female participants show higher reliance on vehicle travel than male participants, as they take 1 more vehicle trip than males and are much less likely to use non-motorized mode. Employed or student participants have more dependency on motorized travel modes, such as vehicle, bus or rail, while unemployed participants take more non-motorized trips. Vehicle dependency is highest for people who live in a household with members under age 18. This group takes about 4 vehicle trips and 0.4 non-vehicle trips per day. On the other hand, participants in single-person households make 2.3 vehicle trips and 0.8 non-vehicle trips per day. This is probably due to the fact that a vehicle can provide greater flexibility and convenience for households with children. As expected, vehicle accessibility not only impacts the overall travel demand but also affects mode choice. Participants without a driver license or participants with no car take 1 non-motorized trip per day while participants with vehicle access only take 0.3 non-motorized trips per day. The daily vehicle and transit usage is similar for participants without access to car. Participants who live near an Expo Line station show slightly higher daily usage of public transit and lower vehicle usage. During later time periods, this group makes as many non-motorized trips as participants who live farther away from an Expo Line station.

Table 2-9 Trip Level Mode Usage by Population Segmentation, Wave I – III, N=55

	Vehicle		**Public Transit**		**Non-motorized**	
	Total #	**Daily avg.**	**Total #**	**Daily avg.**	**Total #**	**Daily avg.**
Overall	3463	3.00	236	0.20	511	0.44
Gender						
Female	2650	3.32	115	0.14	209	0.26
Male	813	2.28	121	0.34	302	0.85
Employment/study status						
Employed/student	2470	3.46	159	0.22	184	0.26
Not employed/student	993	2.25	77	0.17	327	0.74
Household size						
Single family	1112	2.30	125	0.26	247	0.51
All-adult family	1155	3.06	81	0.21	180	0.48
Family with under 18	1196	4.07	30	0.10	84	0.29
Vehicle accessibility						
No car/driver license	153	0.73	145	0.69	202	0.96
Own license and car	3310	3.50	91	0.10	309	0.33
Group						
Experiment	1531	2.70	180	0.32	241	0.43
Control	1932	3.29	56	0.10	270	0.46

Destinations

Land Use

The land use discussed here refers to the primary land use at census block level. Land use information is downloaded from SCAG (2009). The primary land use type is the one occupies the largest area in the census block. The broad range of the “commercial and services” is not sensitive enough to distinguish land use types whose popularity depend on days of the week. Hence, “commercial and services” is further disaggregated to a set of sub-types including “general office use”, “educational”, “retail and commercial”, and “other commercial and service”. Land use types such as “public facilities”, “special use facilities”, and “military installations” are categorized into “other commercial and service”. Excluding home-return trips, commercial or residential oriented areas (or census blocks) are the dominate land use types for destinations. Within the category of commercial and service, blocks with general office use, educational, and retail and commercial account for only one third of the destinations, with the other two-thirds being in blocks with "other commercial and service" land use types. The destination land use patterns between weekday and weekend do not differ significantly. The major difference is that destinations in blocks with a primary use of educational and open space are visited slightly at a higher rate during weekdays. Meanwhile, blocks dominated by industrial or "other" land use are destinations more frequently on weekends. Surprisingly, blocks with predominately business use (general office use) only comprise about 1% of all trip destinations over three waves (Table 2-10).

Table 2-10 Distribution of Trips Based on Primary Land Use at Destination

Primary Land Use Type	Weekday		Weekend	
	count	percent	count	percent
Residential	816	34.99%	290	38.36%
Other commercial and service	639	27.40%	208	27.51%
Other	197	8.45%	71	9.39%
Retail and commercial	185	7.93%	59	7.80%
Industrial	183	7.85%	76	10.05%
Educational	177	7.59%	17	2.25%
Open space and recreation	104	4.46%	28	3.70%
General office use	31	1.33%	7	0.93%
Total	2332	100.00%	756	100.00%

Chapter 3 Tour Complexity and Mode Choice

Introduction

Complex trip chaining is a common phenomenon in urban settings. Trip chaining can be generally defined as a set of consecutive trip segments (Primerano et al., 2007). A better understanding of the trip chaining behavior is needed to improve travel demand models, transportation policy and planning, and travel behavior research about activity and tour making. Activity-based and tour-based studies have shown trip chaining is a complex subject and that multiple dimensions must be considered. However, previous research has mostly treated tours as either simple or complex based on a rather simple classification of whether the tour has more than one destination chained. The conventional approach shows little consideration about other dimensions of a tour, such as the spatial relationship between trip ends, the routing arrangement of a tour, and the overall environment of the chained destinations. In addition, research on the association between tour complexity, mode choice and land use has obtained mixed findings. Some studies show public transit tours and car-based tours are different in nature (Primerano et al., 2007). Other studies argue that complex tours will increase people's reliance on automobile (Hensher & Reyes, 2006). Evidence has shown that inconsistencies in the research may largely be due to the over-simplified classification method for tour complexity (Currie & Delbosc, 2011; Harding et al., 2015; Ho & Mulley, 2013). Giving the advantage of the GPS trajectory data, this chapter explores the relationship between tour complexity, mode choice and land use by extending the tour complexity classification scheme into multiple dimensions. The core research questions being investigated here are how tour complexity is affected by various factors and how tour complexity is associated with mode choice.

This chapter proposes a new method to characterize the complexity of a tour, which will be referred to as the Tour Complexity Index (TCI) in this and the following chapters. The proposed TCI not only considers the number of chained trips per tour, but also takes into account the spatial relationship of the destinations, the sequence of connected destinations, and the land use type of destinations. In this chapter, I examine the dependency of tour complexity on a group of variables using mixed effect generalized regression models. It compares and contrasts the results with previous studies that adopt the more simplistic traditional definition of tour complexity. The proposed TCI provides insights into the association between tour complexity with mode choice, land use features, tour attributes and individual characteristics. The results also shed light on why the existing tour complexity research has reached mixed and often contradictory findings.

This chapter is organized as followed. It starts with a review of travel behavior research regarding the concept of tour complexity. Following a brief review of tour terminologies from previous studies, it then introduces the proposed TCI concept with hypothetical examples, the research questions, data and methodology. The results section presents the descriptive and modeling analysis. Finally, the findings are summarized with a discussion about limitation and future works.

Literature Review

A tour (chained trips) in most research is referred to as a travel process from and returning to home. It can be very simple with one destination close to home, or be extremely complex with several destinations chained and involving multiple travel modes. Analyzing travel behavior at a tour-level provides a better understanding of the complexity of daily travel patterns beyond the basic insights provided by analysis of aggregate daily trips or travel distance. For example, it helps to understand the high frequency of discretionary activities during peak hours given workers tend to link non-work activities with the commute trip. If we look at the commute trip in isolation, discretionary activities appear unreasonable since commuters face strict time constraints during peak hours. Tour complexity is a sub-topic in the trip chaining literature which seeks to understand the relationship between trip chaining behavior and transportation issues. For instance, as a trip chain with multiple stops can reduce the total distance traveled and time cost for the revealed travel demand, it is also can induce extra travel by relaxing cost constraints. In addition, it is important to clarify the relationship between complex tour and public transit usage. If complex travel demand was shown to make more hesitate to take public transit, policy makers should consider how to make the transit system more suitable for complex travel.

A few papers have sought to identify trends in trip chaining and tour complexity. But such research has not revealed consistent insights in tour complexity over years. Levinson & Kumar (1995) find an increase in trip chaining in working trips between 1968 and 1988. The study in Washington D.C. showed that the percentage of chained trips increased by 10 times and 3 times for home-work and work-home trips, respectively (1.5% vs. 15%; 9% vs. 31%). On the other hand, Currie & Delbosc (2011) find a slight decrease in tour complexity in Melbourne area between 1994 to 1999. Furthermore, the argument about whether tour complexity is increasing is often related to the debate that whether complex tours increase reliance on automobiles because such a trend could cause long-term pressure on road network.

Even though the relationship of tour complexity and external factors has not been extensively studied, available studies show consensus that a broad range of factors impacts tour complexity. Household and individual characteristics including age, gender, employment status, household income, accessibility to private vehicle and household structure are found to be associated with tour complexity. Tour attributes such as travel mode, destination choice, time of travel, trip purpose also affect trip chaining behavior (Currie & Delbosc, 2011; de Abreu e Silva et al., 2014; Harding et al., 2015; Ho & Mulley, 2015; Primerano et al., 2007).

The majority of tour complexity studies focus on the relationship among number of chained trips, mode choice and land use, but they have resulted in inconsistent findings. Many studies report that complex tours tend to be vehicle-based. A study using 1992 Sydney data by Hensher & Reyes (2006) find that transit usage has decreased as tours become more complex. This study argued that public transit was not suitable for complex tours because of inflexibility. On the other side, a Melbourne metropolitan study based on data from 1994 to 1999 finds that train and tram have the highest average number of trips per tour (3.37 and 3.24), followed by car-driver (3.07) and bus (2.81) (Currie & Delbosc, 2011). Earlier, Primerano et al. (2007) points out that the nature of trip chaining by vehicle and by transit is different. The author argued that we cannot rule out the role of public transit in complex travel due to flexibility. The author suggests that public transit can support complex tour if mixed and dense activity destinations are located around transit facilities. This argument is later supported by a Swiss-based study (Harding et al., 2015) and another Sydney-based study (Ho & Mulley, 2013). Moreover, Ye et al. (2007) examined the causality between mode choice and tour complexity. By comparing three models with different casual assumptions, the researchers find that the model assuming tour complexity dominates mode choice fits the data best, followed by the model assuming tour complexity and mode choices are made simultaneously, while the model assuming the choice mode occurs prior to the choice for tour complexity has the lowest goodness-of-fit.

The relationship between tour complexity and land use has been investigated and results are also mixed. In a Seattle study, Krizek (2003) investigates individual trip chaining patterns before and after a home relocation and finds that greater neighborhood accessibility lead to more frequent but less-complex tours. Meanwhile, Maat & Timmermans (2006), using samples from central and urbanized regions of Netherlands, have found that higher density areas are associated with more tours overall and more-complex tours. A Portland-based study concludes that better accessibility to bus is associated with more trip chaining (Greenwald & Mcnally, 2008).

Several factors may cause the inconsistency in the reviewed literatures. First, the study subjects and trip/tour types analyzed vary across studies. Some studies have focused on work tours and others on non-work tours. Second, studies use different categorizations of transportation modes. Some research compares motorized v.s. non-motorized mode, and studies which analyze public transit modes vary in classification. Some studies group all transit service as one type while some studies analyze tram, train and bus separately. Third, the method to quantify explanatory variables impacts the modeling results. Some use binary indicators; some use continuous measures such as population density; and some use participant-defined measures. The variation is often high, especially in studies analyzing the influence of land use characteristics. Studies also vary in terms of the geographic definition of land use factors. Most studies examine the influence of near-residence land use patterns but not patterns near other destinations and activity locations.

One of the most common areas of debate in available research is how to classify tour complexity. A binary classification scheme defines a tour as either simple or complex based on number of chained trips. If a tour has two trips, it is classified as a simple tour; if a tour has three or more trips and thus at least two destinations, it is classified as a complex tour. Several recent studies suggest this binary method is limited because it assumes homogeneity among tours with two or more stops. Moreover, it ignores the relationship between trip ends and other tour attributes. A few studies have experimented with an extended measurement for tour complexity. Currie and Delbosc (2011) propose a new approach by treating tours as "more or less" complex depending on number of trips per tour. They have found tours involving public transportation, especially for non work based tours, were more complex than tours by car. Another Australian study based in Sydney further classifies tours not only based on number of chained trips, but also by the spatial relationship between trip ends (Ho & Mulley, 2013). In this study, there are three categories for tours: SPSD (single purpose single destination), MPSD (multiple purpose single destination), and MPMD (multiple purpose multiple destination). It identifies two consecutive activities performed at locations within a walking distance (800m) as a single destination, and consecutive activities undertaken at different locations greater than 800m apart as multiple destinations. The study has found that mode choice is not merely determined by the number of chained trips, but is also influenced by the spatial distribution of destinations. For instance, for a work tour, MPSD tours are more likely to be transit-based and MPMP tours are more likely to be car-based. This study also applies a model with only number of trips for comparison with the new method. It finds that for maintenance and discretionary tours, the results are reversed with number of trips as the only complexity assessment. Without consideration of the trip end's spatial distribution, transit tours are found to be more complex. However, when spatial relationship is included in the tour definition (by grouping destinations within 800m as a single destination), transit tours are likely to be less complex than car tours.

Harding et al. (2015) has followed the tour complexity definition by Ho & Mulley (2013) and extended it in two ways. First, the researchers disaggregate SPSD trips based on travel distance. If the trip is shorter than 800m, it is referred to as SPSD-short; if a trip is longer than 800m, it is identified as SPSD-long. The study has found that the mode share for SPSD tours changes based on distance. Car is the primary mode for SPSD-long tours (61.3% of this category) followed by public transit, while walking is the primary mode for SPSD-short tours (80.5% of this category) followed by bike. Then, the researchers use clustering analysis to classify locations (both home and destination) into four urban types based on population, employment, land use mix and transit facility density. The locations are then classified as type 1 (low density, low mix and poor transit access) to type 4 (most urban, high land use mix, high density and transit accessibility). The location type at home or destination is then used to interact with the tour types. It concludes that residents of urban clusters tend to take MPSD tours and SPSD-short tours. This finding is consistent with Krizek (2003). A destination at a low density environment is most likely to be in a SPSD-long tour and MPMP tour. A destination at an urban environment has higher chance to be involved in SPSD-long tour and MPSD tour.

These three studies which have utilized more detailed tour classification methods (Currie & Delbosc, 2011; Harding et al., 2015; Ho & Mulley, 2013) all supported the argument that the traditional binary classification method is insufficient to study the relationship between tour complexity and other factors. However, there are several aspects that still deserve research attention. First, given Ho & Mulley (2013) and Harding et al. (2015) have stressed the importance of examining the influence of activity clusters, it would be useful to investigate the relative influence of activity cluster centered around home versus around non-home destinations. Second, none of the tour complexity research considers the visiting sequence or frequency of chained destinations. This may due to the difficulty in extracting complete tour-based information from travel survey or dairy data. In addition, previous studies have not considered the influence of the non-primary destinations along a tour. Most studies have focused on the urban environment either at home or at the major tour destination where the primary activity occurs.

Definitions

This section summarizes key trip chaining terminology used in this chapter. A “tour” is a series of consecutive trips that starts and ends at home. Some research refers “tour” as a “trip chain.” In this dissertation, “trip” is limited to a travel process with activities achieved and utility generated at its destination. Hence, a recreational walking trip that starts and ends at home is not considered as a “trip” because the travel process per se is the activity. A stop for travel mode exchange is not considered a destination neither. Meanwhile, the term “trip” is exchangeable with “chained trips”, “(tour) segment” and “leg” in this dissertation. The identification of “activity location” is rule-based. The details are provided in Chapter 2. In short, it is the place where the traveler either has stopped for at least 3 minutes, or has stopped for less than 3 minutes for an identifiable purpose (such as dropping off or picking up someone) based on a manual review of trip routing and the GPS trace. “Activity location” is defined at census block level, which is the smallest geographical unit for which socio-economic data is available. It is also referred to as “(intermediate) stop”, “(chained) location”, or “place”. The set of consecutive activities which occur within the same census block are considered as a bundle, or the same destination. Thus, movement between sequential “stops” within the same census block is not counted as a trip. The above definitions are consistent with the majority of the tour complexity literature so that the results is comparable (Currie & Delbosc, 2011; Harding et al., 2015; Ho & Mulley, 2013).

Research Objectives

The goals of this chapter is to:

- Compare the proposed TCI with the traditional binary classification of tour complexity
- Investigate how tour complexity is affected by travel-related factors
- Examine the relationship between complex tour and public transit usage
- Explore how activity locations influence tour complexity and mode choice

Methodology

Data

The tour-level data analyzed in this chapter includes Wave I data from the Expo tour dataset described in Chapter 2. The population and employment data at census block level, and the road TIGER Lines/shapefile are retrieved from U.S. Census Bureau (2011). The 2011 American Community Survey 1-Year Estimates data is used. Transit information, such as stop location and service frequency, is computed from the LA Metro GTFS (General Transit Feed Specification) for the 2013 October service period. The 2008 existing land use data is downloaded from Southern California Association of Governments (2009).

Tour Complexity Index

The proposed Tour Complexity Index (TCI) is comprised of four indices, described below:

- Segment Index: number of trip segments chained in a tour
- Cluster Index: whether any trip segment is shorter than 1/2 miles
- Efficiency Index: whether the route carried out is equivalent as the shortest path that connects all the destinations with each destination is visited exactly once

- Diversity Index: whether there more than one dominant land use types across destinations

Segment Index

The Segment Index provides the basic information about a tour: how many activity locations the traveler has combined into the tour. It is defined as the number of chained trips in a tour (3-1). It is designed as a counterpart to the traditional binary classification method of tour complexity. Instead of referring a tour as either simple or complex (based on number of trips), the proposed TCI directly quantifies the number of trips chained in the tour. For instance, for tours, {home – school – work – home} and {home – laundry – school – work – school – restaurant – home}, TCI describes them as "a tour with three trips" and "a tour with six trips", rather than calling both "complex tour". The minimum value of the Segment Index is 2 (a tour formed by one outbound and one inbound trip).

$$TCI \mid Segment\ Index = \#\ of\ trips \quad (3\text{-}1)$$

Cluster Index

The Cluster Index is a binary indicator reflecting the spatial relationship among chained stops (3-2). If the tour contains at least one pair of sequential destinations that are with ½ mile (~800m) distance, then its Cluster Index is 1. The threshold of ½ mile distance follows Harding et al. (2015) and Ho & Mulley (2013), which choose this distance because it is a typical walking distance (about 10 minutes walking for an adult). A participant's home can be part of a cluster. Order matters such that a tour cannot be classified as having a cluster if two points being within 0.5 miles of each other but are not sequential. For example, the hypothetical examples of Tour B and D in Figure 3-1 have the same chained locations. Tour B would have a Cluster Index of 1 and Tour D would have a Cluster Index of 0 due to the different visiting sequence. When two locations close to each other are visited together in sequence, it suggests they maybe involve in the tour planning process as a bundle which makes the tour spatially less complex.

$$TCI \mid Cluster\ Index = \begin{cases} 1, there\ is\ at\ least\ one\ cluster \\ 0, otherwise \end{cases} \quad (3\text{-}2)$$

Efficiency Index

Efficiency Index is a binary indicator which indicates how participant arranges the route to visit all the locations in a tour. The Efficiency Index is equal to 1 if the actual route is identical to the shortest path connecting all the locations; otherwise, the index is equal to 0 (3-3). In general, there are several types of tours resulting an Efficiency Index of 0. The first case is when the locations are visited repeatedly in a tour, which indicates the presence of an anchor point other than home (Figure 3-1, Tour B). The second case is a visiting order constrained by the interdependency between the locations. For example, a parent wants to take his/her child to the book store. He/she picks up child from school then visits the book store together, even though visiting the book store first would have resulted in a shorter total distance. Essentially, these two cases reflect the complexity determined by the activities. The third case of inefficient routing may due to an impromptu travel demand which could merely reflect a participant's inefficient route arrangement (Figure 3-1, Tour D).

$$TCI \mid Efficiency\ Index = \begin{cases} 1, & if\ route\ is\ shortest\ path \\ 0, & otherwise \end{cases} \quad (3\text{-}3)$$

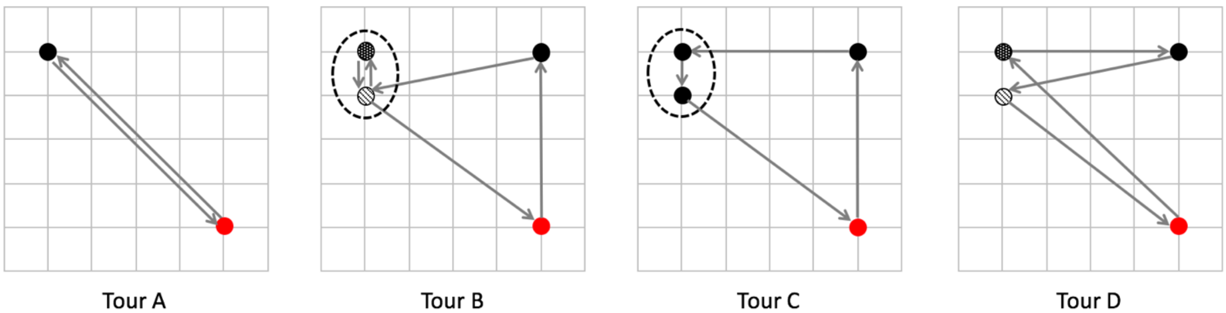
I have considered other methods to quantify the spatial relationship among multiple locations, including convex hull and standard deviational ellipse (SDE), which are more widely used (Perchoux et al., 2014). There are two reasons that they are not adopted. First, the emphasis of the current research is about complexity, not about size or direction. If the chained locations are connected through a shortest path, we consider the tour is efficient in routing arrangement, regardless of the number of stops or travel distance (Figure 3-1, Tour A and C). Second, neither the convex hull nor the SDE methods are applicable for classifying situations such as a single-destination tour with only have one outbound and one inbound trip (Figure 3-1, Tour A). The major advantages of the definition used in TCI is that it can be applied to a single-destination case and it accounts for the travel sequence. At the same time, it avoids overlapping with other indices in TCI, particularly the Segment Index (the number of trips per tour).

Diversity Index

Just as the Segment Index quantifies the complexity relating to the quantity of chained activity locations, the Diversity Index is more of a qualitative measurement that shows the complexity relates to the destinations visited in a tour. It is a binary index of whether there is deviance in the activity location, indirectly measured using dominant land use type of the destination's census block (3-4). A Diversity Index of 0 means the series of chained locations have the same dominant block-level land use type (Figure 3-1, Tour A and C); an index of 1 indicates the series locations are located in blocks with different dominant land use types, which further suggesting a change in the activity context and type (Figure 3-1, Tour B and D).

$$TCI \mid Diversity\ Index = \begin{cases} 1, & LU_i \neq LU_j \quad i, j = 1, \dots, k;\ i \neq j \\ 0, & LU_1 = \cdots = LU_k;\ k = nubmer\ of\ locations \end{cases} \tag{3-4}$$

Diversity Index is designed to be the counterpart of the set of trip purpose in a tour. Information regarding trip purpose, however, is not available in Expo GPS dataset, so that we cannot measure diversity using stated trip purpose. Land use type has been found to be closely related with travel behavior (de Abreu e Silva et al., 2014) and therefore I use the dominant land use of a destination to approximate a trip's purpose and the corresponding level of tour diversity.



		Tour A	Tour B	Tour C	Tour D
Proposed TCI	segment	2	5	4	4
	cluster	0	1	1	0
	efficiency	1	0	1	0
	diversity	0	1	0	1
Traditional classification		simple	complex	complex	complex

1. land use at location is the dominant land use type of a census block associated with the location

Figure 3-1 Illustration of Tour Complexity Index

Analytical Approach

This analysis first utilizes descriptive statistics to explore the relationship between the proposed TCI and other factors, such as travel mode, land use and participant characteristics. Methods include contingency tables and clustering. Then, two groups of mixed effects generalized regression models are specified to first study the impact of external factors on TCI and to second investigate the influence of TCI on transit usage. In both descriptive approach and modeling approach, the traditional binary classification method of defining trip complexity is also applied to compare and contrast the two classification schemes.

Models

Mixed effect generalized linear models are constructed to understand how travel modes and other travel attributes affect tour complexity and how tour complexity influences mode choice. Binomial distributions are analyzed for binary response variables, e.g. whether a tour follows a shortest route; Poisson distributions are analyzed for count variables, e.g. how many trips are chained in a tour. AIC is used for model comparisons since the tested models are not nested. Lower AIC indicates a better model.

The first group of models addresses the first and second research question: how is the result differ under the traditional binary classification of tour complexity and under the proposed TCI; how is tour complexity is affected by travel, personal, and destination characteristics. TCI estimates are treated as response variables. I model separately whether the tour will chain more than one trips (Segment Index), will have a cluster of chained stops within walking-distance (Cluster Index), will have a compact (Efficient Index) route connecting all the stops, or will involve multiple dominant land use types (Diversity Index). The independent variables include travel mode, tour and trip attributes, participant characteristics, and home and destination features. A model of factors associated with a tour complexity based on a simplistic binary definition is included as a comparison between the proposed TCI method and the traditional method.

The second group of models is to investigate the first, third and fourth research questions: how do results vary between the proposed TCI compared to the traditional binary classification of complexity; and what is the relationship between public transit and tour complexity; how activity locations influence tour complexity. This group of models treats mode choice as the response variable to estimate the probability about whether a tour utilizes public transit, given TCI and other related variables. The first two models use traditional tour complexity classification schemes. One includes a binary complex tour indicator; another uses a count variable as number of trips chained in the tour. The other three models are TCI models and are specified as follows:

1. A TCI model including the four proposed indices: number of chained trips, indicator of cluster, indicator of efficient route, and indicator of diverse land use types
2. A TCI model that splits the cluster indicator into two variables: cluster at home or cluster at non-home destinations
3. A TCI model that is the same as the first TCI model in TCI variables, but using furthest destination attributes instead of main destination attributes (the one with longest activity duration)

Independent Variables

Participant characteristics include gender, age, employment/school enrollment status, household income, car ownership, vehicle per driver license, presence of children under 18. Tour and trip attributes include departure time, day of week, total travel distance and destination land use types. Origin and destination characteristics are designed following the Density-Design-Diversity (3-D) effects introduced by Cervero & Kockelman (1997).

For density, population per acre and urbanization status are computed for activity locations at census block level. Census-block population is aggregated from the census track population data of 2007 – 2011 American Community Survey 5-Year Estimates, U.S. Census Bureau. The size of census block equals the land area plus water area obtained with the census block shapefile downloaded from the 2011 U.S. Census Bureau's TIGER Line/Shapefiles. Urbanization status is the proportion of a census block containing land use information based on 2008 existing land use data obtained from Southern California Association of Governments. The total area of a census block is calculated using ArcGIS "calculate geometry" function.

First, it is necessary to clarify the concept of "diversity" used as the exploratory variable. The term "diversity" in the 3-D effects mainly refers to the degree of environmental mixture such as a mixture of land use types at a specific location. It is separated concept from the Diversity Index of TCI with the latter refers to a change of dominant land use across all chained locations. Land use mixture is estimated using an entropy index to quantify diversity at destination level. The employer and resident ratio at the census block level is computed but excluded from models due to a large number of blocks with an unreasonable number of employees. The data source of census block employment is the Longitudinal Employer-Household Dynamics (LEHD) Origin-Destination Employment Statistics (LODES) dataset. The relevant variable, population mix, is constructed based on the total block employment for the year of 2011 (U.S. Census Bureau, 2011).

As for design, accessibility is reflected in the following variables: distance to the nearest high-frequency bus stops, and distance to the nearest freeway entrance. They are computed using "empty space distance", one of the classical methods to quantify the interaction between spatial locations (Baddeley, 2008). It is defined as the Euclidean distance between a specific location, u, to the nearest point in a point cluster X (3-5).

$$d(u) = min_i \|u - x_i\|, \qquad x_i \in X \qquad \textit{(3-5)}$$

The criterion of high-frequency is a service window shorter than 6 minutes during morning peak. The bus service frequency per bus stop is extracted from Los Angeles Metro GTFS 2013 October service period. In addition to the accessibility, distance to downtown Los Angeles is computed to control for a general urban setting. As destinations are closer to downtown, the area may be denser, have better road and transit accessibility and higher diversity in land use. On the other hand, the further away from downtown, the destination area is expected to be less dense, dispersive and to have limited transit accessibility.

To address the fourth research question regarding the role of destinations in mode choice, two types of destinations are studied. The first is defined by the main tour destination ("M"), which is the location in which the participant spends the longest activity duration. The duration is aggregated if a location is visited more than once during a tour. The second type of destination is defined as the furthest destination from home with distance computed as spherical distance ("F").

The complete list of the variables and descriptive statistics are provided in Table 3-1. For binary variables, the mean is computed as the ratio of the true value to total number of observations. Its standard deviation is the square root of mean times one minus mean divided by the number of observations. The statistics for participants are computed using a total of 55.

Table 3-1 Descriptive Statistics of Independent Variable

Variable Name	Description	Mean	s.d.
Traveler Characteristics			
Male	Male (0/1)	0.309	0.062
Age	Age	52.163	14.19
High income	Household annual income >= $75,000 (0/1)	0.182	0.052
Own car	Own car (0/1)	0.818	0.052
Share car	Share car with family (0/1)	0.255	0.059
Worker/Student	Employed or student (0/1)	0.618	0.066
Fulltime worker/student	Fulltime (0/1)	0309	0.062
Young children	Presence of household children under 18 (0/1)	0.255	0.059
Single-person family	Single-person family (0/1)	0.418	0.067
Large family	Family size > 4 (0/1)	0.182	0.052
Tour & Trip Attributes			
Weekend	Weekend (0/1)	0.303	0.023
Peak hour	Peak hour (0/1)	0.511	0.025
Travel distance	Total travel distance in miles	11.208	10.492
Location Features			
Density			
Pop. density (M)	Number of residents per acre at main stop (M) or at furthest stop (F)	10.584	15.585
Pop. density (F)		11.088	15.794
Urbanization (M)	% of area with designated land use in total area at main stop (M) or at furthest stop (F)	0.803	0.113
Urbanization (F)		0.796	0.115
Diversity			
LU mix (M)	Land use mixture measured in entropy index at main stop (M) or at furthest stop (F)	0.387	0.399
LU mix (F)		0.379	0.402
Design			
Dist. bus (H)	Distance to nearest high-frequency bus stop in mile from home (H), main stop (M) or furthest stop (F)	0.768	0.427
Dist. bus (M)		0.858	0.904
Dist. bus (F)		0.919	0.919
Dist. freeway (H)	Distance to nearest freeway entrance (mile) from home (H), main stop (M) or furthest stop (F)	1.119	0.716
Dist. freeway (M)		0.832	0.669
Dist. freeway (F)		0.804	0.675
Rail density (M)	Rail kernel density at main stop (M) or at furthest stop (F)	0.030	0.113
Rail density (F)		0.029	0.112
Dist. DT LA (M)	Distance to downtown Los Angeles (miles) from main stop (M) or furthest stop (F)	7.772	3.640
Dist. DT LA (F)		7.970	3.686

Results

Descriptive Analysis

Distribution of tours by TCI

The analysis in this chapter is based on the Wave 1 GPS traces of the 55 participants in the study sample and include a total of 409 valid tours. Among them, 36.2% of tours are simple tours with two trips (single destination); 28.6% of tours have three trips (two destinations); 13.5% of tours have four trips; 21.7% of tours have five or more trips (Table 3-2 and

Table 3-3). There is a tour with 16 trip segments where the trips may be part of the participant's job (such as bus driver or delivery person). However, it is not removed from the analysis due to lack of evidence. Applying the binary classification method of tour complexity, "complex" tours (>= 2 trips) account for 63.8% of all the tours; and about 66% of these "complex" tours have chained two or three destinations (Table 3-2). The average number of trips chained in a tour is 3.6 much higher than the empirical results from previous studies. A Melbourne-based study found that, from 1994 to 1999, the average trip legs per tour was 2.9 and less than half of the tours were complex (Currie & Delbosc, 2011). The difference may due to the different living styles between two geographical regions and different time frames. It is also possibly due to the sampling method where the data in Currie & Delbosc (2011) has a a wider population coverage (randomly sample 5,000 Melbourne households per year). Moreover, Currie & Delbosc (2011) uses travel diary while this study is based on GPS data. The trip rate obtained from travel diary has been found to be averagely lower than that from GPS data (Houston et al., 2014). On the other hand, the result of this study could be inflated because of the possible inclusion of trips that are part of the traveler's job.

Table 3-2 TCI Statistics of Expo Wave I tours

TCI	mean	[min., max.]
Segment	3.614	[2, 16]
	# of 0	**# of 1**
Cluster	242	167
Efficiency	82	327
Diversity	192	217

Table 3-3 Distribution of Expo Wave I Tours by TCI Segment Index

Frequency	TCI			
	Segment	**Cluster**	**Efficiency**	**Diversity**
148	2	0.18	1	0
117	3	0.42	1	0.71
55	4	0.49	0.71	0.84
31	5	0.52	0.48	0.97
21	6	0.76	0.23	1
13	7	0.69	0.23	1
8	8	0.88	0	1
6	9	1	0	1
2	10	1	0	1
3	11	1	0	1
1	12	1	0	1
2	13	1	0	1
1	14	1	0	1
1	16	1	0	1

Table 3-4 Contingency Table for TCI Values

		Diversity Index = 0			**Diversity Index = 1**			
		Efficiency Index			**Efficiency Index**			
		0	1	subtotal	0	1	subtotal	row sum
Cluster Index	0	1	135	136	27	79	106	242
	1	2	54	56	52	59	111	167
	col. sum	3	189	192	79	138	217	*Total: 409*

Table 3-3 shows the distribution of tours disaggregated by TCI and

Table 3-4 is the contingency table with Segment Index aggregated. In Wave I, 167 out of 409 tours (40.8%) have clustered locations within the tour. Further examination shows only 26 of these 167 tours with clusters are short-distance tour with one destination near home (<= 0.5 mile). It also means the majority of the single-destination tours (148 - 26 = 122, 82.4%) are long distance. The percentage of long-distance single-destination tours is higher than that from a Swiss-based study where 76.1% single-destination tours are long distance (Harding et al., 2015). A total of 327 tours have followed shortest path, of which 189 have a Diversity Index of 0 and 138 have a Diversity Index of 1. However, if we exclude those tours with one or two destinations that have an Efficiency Index as 1 by default, only 18 (12.5%) of the multi-stop tours are "compact". For the Diversity Index, about half of the tours have the same dominant land use setting across stops (192, 47.2%). The likelihood of Cluster Index as 1 and Diversity Index as 1 both increase as the Segment Index increases, meanwhile the Efficiency Index decreases.

Table 3-5 Distribution of Expo Wave I Tours by Mode Usage with TCI Statistics

	Frequency		**TCI (mean)**			
	Count	**%**	**Segment**	**Cluster**	**Efficiency**	**Diversity**
Grand total	409	100%	3.62	0.41	0.80	0.53
One mode	*369*	*90.22%*				
Vehicle	326	79.71%	3.57	0.31	0.80	0.54
Transit	11	2.69%	2.18	0.00	1.00	0.09
Non-motorized	32	7.82%	2.66	0.97	0.97	0.19
Mixed mode	*40*	*9.78%*				
Vehicle + transit	4	0.98%	5.68	0.91	0.50	0.82
Vehicle + non-motorized	22	5.38%	4.25	0.50	0.50	0.75
Transit + non-motorized	13	3.18%	4.46	1.00	0.69	0.85
Transit + vehicle + non-motorized	1	0.24%	5.00	1.00	1.00	1.00

TCI and travel mode

Table 3-5 and Figure 3-2 shows how travel mode interact with TCI. 90% of the tours are dominated by a single mode and only 29 (7%) tours involve transit. For parsimony, the tour mode in the following analysis refers to the primary mode of the tour. The primary mode is determined based on the hierarchical order: transit, vehicle, and non-motorized. So for a tour used both vehicle and transit, it is referred to as a transit tour.

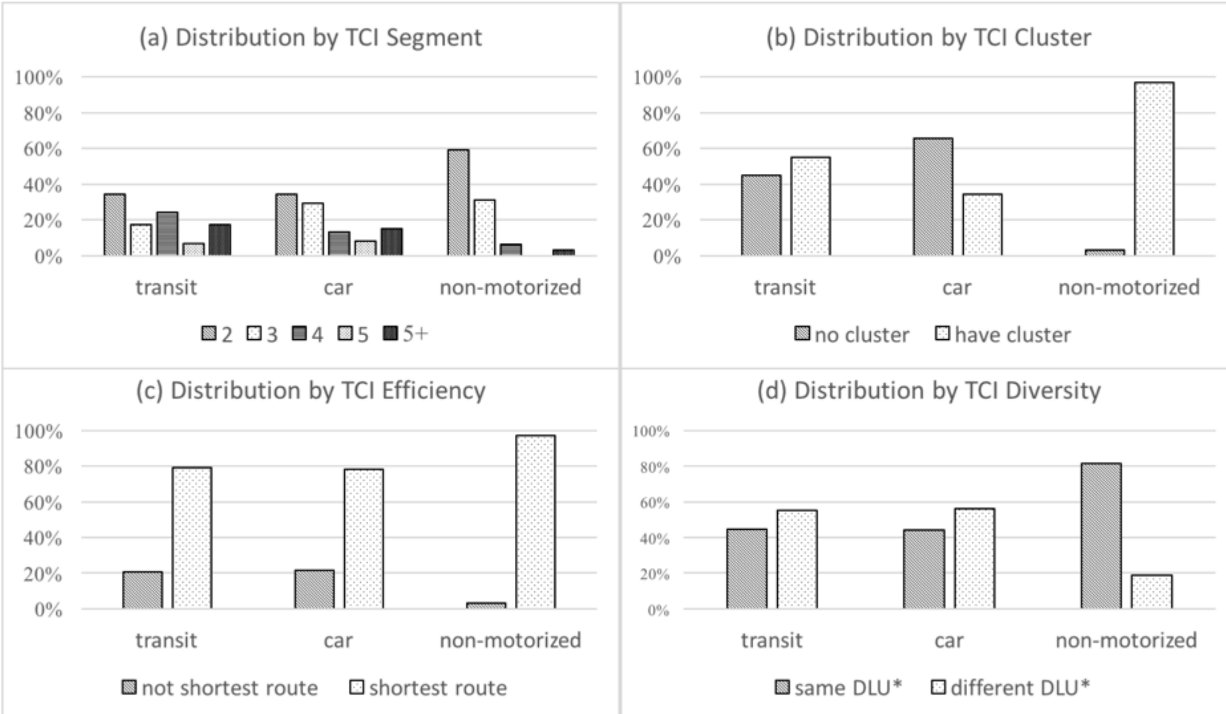
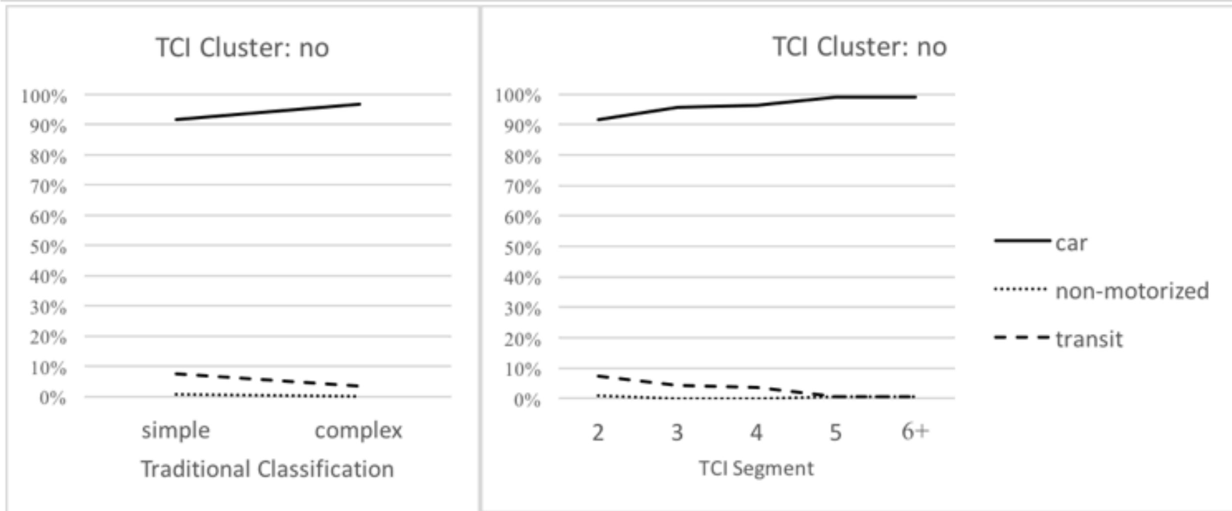


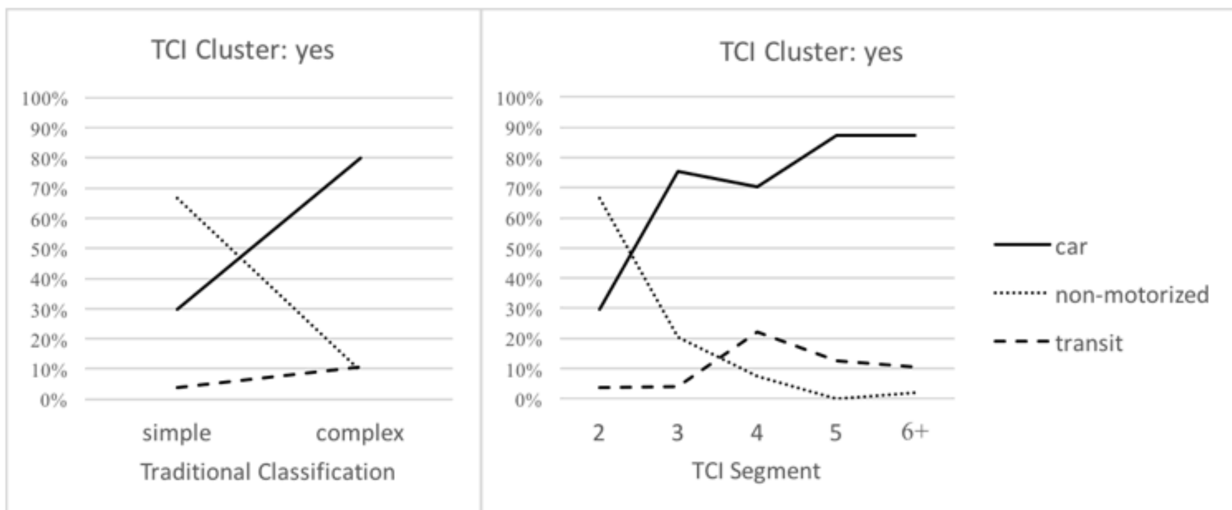
Figure 3-2 Shares by TCI Across Travel Modes

Figure 3-2 shows how TCI and its subcomponents are distributed across travel modes. The distribution of tours by number of segments is similar in car-based tour and transit-based tours. This implies the number of tour segments is not a barrier of using transit. At the same time, tours with non-motorized modes seem to involve fewer chained trip segments. Car-based tours are less likely and non-motorized tours are most likely to involve location clustering. Transit-based tours have a slightly higher rate of involving location clustering than car tours. The three tour modes show similar trends in terms of route efficiency, but non-motorized tours are dominated by route-efficient tours. Since the distance cost tends to be higher for non-motorized tours, participants likely seek to minimize the distance for a walking or biking trips. Meanwhile, transit and car tours show an almost identical patterns in terms of diversity. This suggests that complexity in routing is not a barrier to choosing transit. Lastly, transit and car-based tours both have higher likelihood to have chained destinations with mixed dominant land use, while non-motorized tours have larger chance to be dominated by a single land use type. Although non-motorized tours can have multiple trip segments, their overall geographical range is smaller and so it is more likely that these chained locations have similar urban setting.

Figure 3-3 depicts how tour mode share is related to the number of trips. Panel (a) is based on the traditional binary classification scheme of complexity where tours are labeled as either simple or complex based on whether more than one destination are chained; Panel (b) is based on TCI Segment Index which equals to the number of segments without further simplification. When there is no cluster in the tour, two classification schemes show identical patterns. That is, the share of vehicle tours increases and the share of non-vehicle tours decreases when the number of segments increases. However, if a tour contains clustered destinations, the classification scheme that reports the exact number of trips (Panel (b)) reveals more information than the binary classification scheme (Panel (a)). First, both shows that non-motorized mode dominates the short-distance (1/2 miles) single-destination tours, followed by vehicle and then transit. Then, when number of trips increases, Panel (a) tells us that the share of both vehicle and transit mode increase monotonically while Panel (b) shows that the transit usage actually declines with number of chained trips exceeds 5. This suggests that the choice to take public transit may be discouraged when a traveler plans to visit a larger number of destinations. This result partially supports the argument by Ho & Mulley (2013) that dense activity centers encourage transit use. However, based on Ho & Mulley's study, the share of vehicle mode decreases as number of segment increases when clusters present which does not agree with the result found here.



(a) by binary classification



(b) by TCI segment

Figure 3-3 Mode Share by Segment and Cluster Indices

Clustering

Instead of adopting a tour topology from other studies, I employ a clustering algorithm to explore the tour patterns. There are two popular clustering algorithms, K-means and K-medoids. The concept of K-means is to minimize the within-group variance by assigning groups to the data points after a set of group centers is found. K-medoids is a generalized version of K-means. The major difference is that K-medoids uses an observed data point as the group center and is more robust to outliers. Meanwhile, K-medoids can apply any dissimilarity metric in addition to Euclidean distance. Considering the dataset contains a few extreme cases and a mixture of variable types (ordinal and binary), I adopted the K-medoids for the analysis.

Table 3-6 Tour Clustering by K-medoids Algorithm by TCI

Group (count)	Description	Group mean/mode			
		Segment	Cluster	Efficiency	Diversity
A (135)	Single-stop long distance tours	2.10	0	1	0
B (56)	Single-stop or two-stop tours with clustered and similar destinations	2.70	1	1	0
C (79)	Multi-stop tours with dispersed and diverse destinations	3.43	0	1	1
D (59)	Multi-stop tours with clustered and diverse destinations	3.92	1	1	1
E (28)	Multi-stop tours with dispersed and diverse destinations, and inefficient route	5.14	0	0	1
F (52)	Multi-stop tours with clustered and diverse destinations, and inefficient route	7.64	1	0	1

For K-medoids clustering, Gower's similarity is used to compute a dissimilarity matrix with Segment Index treated as an ordinal variable and the other three indices treated as binary. The criterion used for selecting the optimal number of clusters is average silhouette. As a result, the 409 tours are divided into six clusters (Table 3-6, ordered by number of chained trips). This result suggests there are six common types of tours: 1) single-stop tours with the destination exceeding a typical walking-distance as ½ miles (Cluster A); 2) slightly more complex tours with at least one walkable trip and similar types of destinations (Cluster B); 3) multi-stop tours with dispersed and different types of destinations (Cluster C); 4) multi-stop tours with at least one destination cluster and different types of destinations (Cluster D); 5) multi-stop tours with dispersed destinations, diverse land use settings and an inefficient routing arrangement (Cluster E); and 6) a multi-stop tour with at least one destination cluster, diverse land use settings and an inefficient routing arrangement (Cluster F).

The most common tour type is Cluster A, a long distance tour with a single destination. It comprises 33% of the Wave I tours. Other tour types are almost evenly distributed, each comprising on average 13.4%. Based on the combination of Cluster, Efficiency and Diversity Index, we can see the six clusters form three pairs: Clusters A and B, Clusters C and D, and Clusters E and F. The first pair, Cluster A and B, account for 46.7% of Wave I tours. Most of the tours in these two types have one or two destinations. They are both have efficient shortest-path routes and the tour destinations have same dominant land use types. The difference between Cluster A and B is that Cluster A is dominated by single-stop and long distance tours, while Cluster B is a composition of single-stop tours whose destination is close to home and two-stop tours with at least one pair of destinations close to each other. Cluster C and D are mainly multi-stop tours with more than one type of destination and segments chained into the shortest-path route. The difference between C and D is whether the tour includes destination clusters. Together, they account for 33.7% of the tours in Wave I. Finally, Cluster E and F are tours with more chained locations that are sequenced in an inefficient, non-direct way. The chained locations are very likely to be in areas with different dominant land use types. As above, the major difference between Cluster E and F is the presence of a destination cluster. Note that for each pair of clusters, the ones with a cluster have higher average number of trips, which suggests dense activity places are associated with more complex tours in terms of trip segments (Clusters B, D and F). Meanwhile, tours with a higher number of trips have a greater chance of having multiple types of locations chained rather than a single type (Clusters C, D, E and F). Tours with a five or more trips are less likely to have an efficient route and are more likely to have an anchor point or zigzagging route (Clusters E and F).

Multivariable Analysis

<u>Influence of travel factors on tour complexity</u>

The impact on the four indices of TCI from the independent variables are different in terms of direction, magnitude and significance level (Table 3-7, Models 2-5). The comparison between the four TCI models (Table 3-7, Model 2-5) and the model with the more simplistic binary classification of a complex tour (Table 3-7, Model 1) suggests that treating tour complexity as a binary variable may overlook important factors explaining the association between tour complexity and external factors.

Tours by vehicle have a strong association with the probability that the tour includes a different primary land use types across destinations (Model 5) and has a higher number of chained trips in a tour (Model 2). Tours by vehicle are also associated with a reduction in a tour's probability of connecting stops using the most efficient shortest-path route by 90% (Model 4). Tours with a non-motorized travel mode are found to have a statistically significant and positive association with the number of segments, clustering, and land use diversity at destinations (Models 2, 3 and 5), they are also significantly and negatively related with an efficient routing arrangement (Model 4). Tours with a public transit mode, however, are not significantly related with any of the TCI outcomes, which may partially be due to limited number of participant transit-based tours. If a tour is initiated during a peak hour, it is less likely to have cluster(s) (Model 3). Both travel distance and the existence of commercial and service type destinations have a positive relationship with number of chained trips, cluster and diverse land use at destinations for a tour, and reduces the chance a tour follows the most efficient or compact route.

Being male is positively associated with an increased likelihood of having tours with at least one cluster and having a higher household income is associated with a reduced likelihood of choosing destinations that are clustered. Car ownership is associated with a lower likelihood of a tour with destination diversity. This may because participants with a car tend to make less-chained and longer distance tours. Interestingly, participants who need to share a household vehicle with other family members are associated with a higher likelihood of having a tour with a more compact, efficient route. Employed or student participants, regardless of full-time or part-time, have a likelihood of making chained trips 1.2 times that of other participants and they are 75% less likely to have a compact route and almost 3 times as likely to have different land use types across stops. Full-time work or school was not significantly related to TCI components, but it reduces the chance of complex tour based on the simple binary definition of tour complexity (Model 1). A participant who lives in a family with four or more people is 73% less likely to organize the tour in a compact way but is 156% more likely to have a diverse land uses across tour destinations. Adults with children under 18 are associated with a significantly lower probability of having clustered destinations, likely because they tend to choose destinations based on children's needs rather than based on distance.

The characteristics of destinations in a tour seem to influence on tour patterns choice more than that the characteristics of residential locations. Similar conclusions can be found in studies in non-U.S. urban settings (Maat & Timmermans, 2009). The variables for the main destination characteristics explain the TCI components better than the variables for the furthest destination characteristics. The degree of urbanization and land use mixture at main tour destinations encourages the overall tour-level diversity. If the main destination is far away from a high-frequency bus stop, then the tour is less likely to have fewer trips chained, less likely to have clustered and diverse destinations overall. One additional unit of railway kernel density at main destination cuts the likelihood of the number of chained stops in a tour by almost half. However, it greatly increases the likelihood of having a compact route. If the main destination is far away from downtown Los Angeles, the tour has a lower rate of chained trips and higher probability of the shortest efficient route. The above suggests that transit accessibility encourages more chained trips, more complex routing arrangements and more diversity among activity locations. In addition, neither of the two variables related to residential accessibility are significantly related with the TCI components.

Relationship between transit mode choice and tour complexity

Models 6-10 investigate the influence of the TCI components on the probability a traveler takes public transportation in at least one segment of a tour. Result is shown in Table 3-8. Overall, the model with four TCIs and which includes characteristics of the furthest destination (versus the home destination) fits the data best (Model 10). All three TCI models (Models 8-10) perform better than non-TCI models (Models 6-7). The model treating tour complexity using a binary classification method fits the data most poorly (Model 6). In addition, although Model 9 is slightly worse than Model 8, it shows that non-home clustered destinations play an important role in transit mode choice, rather than clustered destinations near home.

The two models without TCIs have conflicting results regarding the relationship between public transit usage and tour complexity which solely depends on number of trips. The model treating tour complexity as binary suggests that a complex tour is 22% less likely to involve public transit (Model 6). To the contrary, the model treating tour complexity as a count variable suggests that the likelihood that transit usage will increase by 38.5% for each additional chained destination.

The conflict is consistent with the contradictory conclusions from existing tour complexity studies. When other TCI indices are controlled for (Models 8-10), the coefficients for number of segments are positive but none of them are statistically significant, which further suggests that the number of trips may not be an sufficient or efficient explanatory variable for mode choice. This pattern supports Primerano et al. (2007)'s argument that the nature of car tours and transit tours is different and cannot be explained simply using number of trips.

The existence of cluster, especially the existence of cluster near a destination away from the home environment, is positively and significantly related with the usage of public transit. In Model 9, the binary cluster variable is split into two separate cluster indicators. The variable for near-home clusters means at least one destination is within walking-distance (0.5 mile) of the participant's residential location; the variable for non-home destinations means at least one pair of destinations beyond walking distance of the participant's home is within walking-distance. Note that: 1) if one destination is close to both home and another location, then both indicators are equal to 1; 2) if two destinations are close to each other but not visited consecutively, then they are not classified as a cluster because they are not perceived as a cluster for the traveler. Model 9 suggests the existence of cluster at a non-home destination is positively associated with the likelihood a participant will use public transit use, while the binary for a home-based cluster is not statistically significant and its magnitude is much smaller.

Table 3-7 Estimation Results of Tour Complexity Models

Variable Name	Traditional		Proposed TCI		
	Model 1	Model 2	Model 3	Model 4	Model 5
	Complex tour (0/1)	Segment (count)	Cluster (0/1)	Efficiency (0/1)	Diversity (0/1)
	e^{β}(p-value)[1]	e^{β}(p-value)	e^{β}(p-value)	e^{β}(p-value)	e^{β}(p-value)
Tour & trip					
Car	48.014***	1.616***	5.730	0.107**	30.494***
Transit	2.302	1.017	0.873	0.928	1.465
Non-motorized	20.678***	1.599***	177.186***	0.157***	9.030***
Peak hour	1.092	0.966	0.621*	1.151	1.244
Weekend	1.841**	1.043	0.827	0.675	1.329
Travel distance	1.084***	1.030***	1.037**	0.871***	1.103***
Comm. at Dest.	7.988***	1.343***	1.943**	0.172***	15.390***
Traveler					
Male	1.116	1.035	2.287**	1.217	1.390
Age	1.004	1.002	0.984	0.987	0.993
High income	0.860	0.923	0.401**	1.943	0.976
Own car	0.160**	0.890	0.661	2.531	0.162**
Share car	0.540	0.870	0.623	3.187**	0.635
Worker/Student	3.273***	1.209**	1.335	0.257**	2.726**
Fulltime Work or School	0.406**	0.873	0.900	2.113	0.682
Single family	1.650	1.050	0.990	0.777	0.986
Large family	3.732**	1.166	0.822	0.265*	2.561*
Young children	0.571	0.902	0.367**	2.013	0.547
Home & destination					
Pop. density (M)	0.996	0.999	0.986	1.003	0.982*
Urbanization (M)	3.412	0.720	1.916	4.719	13.078**
LU mix (M)	1.677	1.100	1.878*	0.627	2.417**
Dist. bus (M)	0.590***	0.930**	0.632**	1.296	0.676**
Rail density (M)	0.120	0.439***	0.163	36.035**	0.185
Dist. freeway (M)	0.671*	1.007	1.660**	0.570**	0.559**
Dist. LADT (M)	1.052	0.979*	0.969	1.180**	1.028
Dist. bus (H)	0.513	0.936	1.466	2.173	0.799
Dist. freeway (H)	1.597*	0.998	1.288	1.233	1.461
Constant	0.001***	1.760*	0.048	195.604**	0.001***
AIC (d.f.)	426.853 (28)	244.005 (28)	439.628 (28)	329.607 (28)	416.265 (28)

*Notes: 1. The values reported are change of odds ratio for logistic model and change of rate for Poisson model as response to one unit increases in the corresponding variable, with other factors fixed; significance level: 0.01***; 0.05**; 0.1**

Table 3-8 Estimation Results of Tour-level Mode Choice Models

Variable Name	Without TCI		With TCI		
	Model 6	Model 7	Model 8	Model 9	Model 10
	e^{β}(p-value)[1]	e^{β}(p-value)	e^{β}(p-value)	e^{β}(p-value)	e^{β}(p-value)
Tour complexity					
Complex tour (0/1)	0.787	-	-	-	-
# of chained trips	-	1.385***	1.260	1.251	1.198
Cluster (0/1)	-	-	15.121***	-	15.362***
Cluster at home (0/1)	-	-	-	1.129	-
Cluster at dest. (0/1)	-	-	-	12.605***	-
Compact route (0/1)	-	-	0.874	0.818	0.816
Mixed env. (0/1)	-	-	0.163**	0.170**	0.152**
Tour & trip					
Peak hour	1.427	1.484	1.708	1.766	1.434
Weekend	1.101	1.208	1.353	1.042	1.606
Travel distance	0.969	0.909**	0.937	0.910**	0.957
Comm. at Dest.	0.833	0.597	1.084	0.898	0.979
Traveler					
Male	25.761***	30.026***	17.708***	24.254***	17.007***
Age	1.034	1.040	1.048	1.032	1.047
High income	0.203	0.181	0.466	0.306	0.400
Own car	0.001***	0.001***	0.001***	<0.001***	0.000***
Share car	0.501	0.586	0.631	0.756	0.647
Worker/Student	3.383	3.105	2.743	3.093	2.346
Fulltime Work or School	0.879	1.196	0.859	1.028	0.760
Single family	3.028	3.334	2.521	3.476	2.400
Large family	2.660	2.426	3.182	2.209	2.874
Young children	2.688	3.709	4.804	3.477	6.226
Home & destination					
Pop. density (M)	1.007	1.010	1.016	1.015	
Pop. density (F)					1.022
Urbanization (M)	0.190	0.323	1.386	1.457	
Urbanization (F)					0.545
LU mix (M)	2.132	1.573	1.449	1.384	
LU mix (F)					3.187*
Dist. bus (H)	0.417	0.427	0.276	0.372	0.248
Dist. bus (M)	0.967	1.135	1.018	1.028	
Dist. bus (F)					0.818
Dist. freeway (H)	0.670	0.674	0.521	0.593	0.427
Dist. freeway (M)	2.009**	2.352**	1.899*	2.503**	
Dist. freeway (F)					1.937*
Rail density (M)	19.815	111.199*	57.729	353.520*	
Rail density (F)					23.468
Dist. LA dt (M)	1.063	1.132	1.171	1.199	
Dist. LA dt (F)	-	-	-	-	1.159
Constant	1.071	-0.908	-2.531	-1.502	0.314
AIC (d.f.)	260.171 (26)	253.347 (26)	235.600 (29)	240.658 (30)	231.321 (29)
Model comparison	M5 > M3 > M4 > M2 > M1				

*Notes: 1. The values reported are change of odds ratio as response to one unit increases in the corresponding variable, with other factors fixed; significance level: 0.01***; 0.05**; 0.1**

Discussion and Conclusions

The results in this chapter add to the understanding of the connection between tour complexity and other travel factors. It also suggests reasons why previous research has obtained conflicting results. First, both descriptive and modeling analysis suggests that the proposed TCI suits tour-level study better than the traditional binary classification method, regardless of whether the study is about tour complexity per se or is about other travel aspects with tour complexity as an exploratory variable. The tour complexity models provide a better understanding about the travel features which affect tour complexity by separately examining the influence of 4 separate TCI components. The mode choice models suggest that the spatial clustering of destinations and the land use diversity at destinations are both important explanatory variables, rather than the widely-used count of chained trips. Compared to the few tour complexity studies which considered the spatial pattern of destinations (Harding et al., 2015; Ho & Mulley, 2013), the inclusion of the diverse land use indicator provides more statistically and practically significant information. Second, we find that the total travel distance and whether there is commercial and service land use at destinations significantly influence tour complexity. Individual characteristics, such as gender, income, employment/enrollment status, family size and presence of children, are also associated with tour complexity. Third, we find that vehicle and non-motorized travel modes have stronger connection with tour complexity than public transit. This suggests the discussion about mode choice and tour pattern or complexity should focus more on these two modes. However, the existing literature tends to group non-motorized modes with transit modes, or tends to solely focus on public transit. Last but not least, we discover that destinations beyond the home environment have more impact on tour complexity and mode choice than near-residence destinations. More specifically, the characteristics of destination with the longest activity duration explains tour complexity better, while the characteristics of destination that is farthest from the residential location explains mode choice better.

Limitations and future work

Firstly, the study has not validated the proposed TCI from a theoretical perspective. All the results are obtained via empirical analysis. Further effort could be made to look at the tour complexity theoretically. Secondly, the study does not analyze the interaction between the TCI components. A multivariate analysis that treats the four indices of TCI simultaneously as a set of response variables would provide a deeper understanding about how the external factors affect the TCI. The major reason for not applying multivariate analysis is the limited sample size. The size of 409 tours is small comparing to other tour-based studies. Meanwhile, the tours are undertaken by 55 travelers, meaning there is dependency among observations. So the complexity of possible statistical models is severely constrained after adding a correlation matrix to control for the dependency. One potential step further is to test the performance of TCI on a larger dataset, such as the state-wide household travel survey. Furthermore, although the exploratory on dataset without trip purpose information is worthwhile, a dataset including trip purpose could facilitate the analysis about tour complexity and the nature of the chained activities. Another direction that could be further investigated is the characteristics of clustered destinations since a cluster at non-home destinations is found to be greatly associated with public transit usage. More knowledge in this area could help policy makers and transportation planners to better promote transit and reduce car reliance.

Chapter 4 A Longitudinal Analysis of Daily Travel Variability

Introduction

Many urban travel behavior studies and travel demand models are developed based on of the single-day travel survey. Besides the fact that one-day surveys cost less than multiday surveys, the underlying assumption here is: 1) most urban residents have habitual travel patterns, and that 2) by randomly choosing individuals/households and randomly observing weekdays, the resulting sample is representative to the population overtime (Pas & Sundar, 1995). Based on these assumptions, it is sufficient to establish theories and to form models based on single-day travel survey data. There has been continuous questioning about this assumption since 1980s. Many studies have investigated the day-to-day travel variability and found that the amount of deviation is too high to be compensated for by the random sampling mechanism (Hanson & Huff, 1982; Pas & Sundar, 1995; Raux et al., 2016; Stopher & Zhang, 2011). Moreover, most single-day surveys are restricted to weekdays. Questions have also been raised regarding the length of the cycle of human travel patterns, i.e. whether travel behavior repeats on a daily, weekly, or monthly basis. It is important to clarify these issues because they have significant implications on travel survey design. For example, it has been suggested that three or four days represent the optimal length for data collection, and that the widely-used one-day travel survey should be extended, and that costs can be reduced by shortening week-long surveys (Pas & Sundar, 1995; Stopher et al., 2008). Further, the variability in travel is sensitive to model estimation and result interpretation. Research has found that the modeling residual is associated with the length of the repetition cycle used in the model (Raux et al., 2016).

Given the fact that many variability-related studies have confirmed the existence of day-to-day variability in various aspects of travel behavior such as trip generation, time allocation, destination choice, mode usage and daily pattern, it would be helpful to examine the variability in complex trip chaining behavior. Using a longitudinal GPS-derived tour dataset, this chapter provides an empirical assessment of the daily and weekly variability from the perspective of tour complexity and mode choice. Meanwhile, given the experimental nature of the dataset, a before-and-after analysis is performed to examine whether proximity to transit facility would affect the variability in complex trip chaining behavior.

The chapter is organized as follows. First, it provides an overview of research on travel pattern variability, and variability in trip chaining in particular. Second, it presents a discussion of the analytical methods and dataset. The measurement of variability in trip complexity and the TCI components is introduced using an illustrative example developed from the dataset. Third, the results show the components of the variation in tour complexity, the influence of a new transit service on individual's trip chaining variability, and the association between social-demographical factors and variability in tour complexity. Finally, conclusion and discussion is provided.

Literature Review

Travel demand forecasting based on one-day travel diary data may be biased due to unobserved day-to-day variability. As urban travel behavior is a complicate topic, it is impossible to provide complete analytical consideration of all the aspects in a single investigation of the variability. Previous studies have focused on a various aspects of variability, including travel frequency, time allocation, destination choice, mode usage and trip chaining. Most studies examine multiple subjects either separately or simultaneously (Hanson & Huff, 1982; Kitamura & Van Der Hoorn, 1987; Pas & Sundar, 1995; Raux et al., 2016). A few studies focus on one subject when the subject per se is multi-dimensional, such as in the case of trip chaining (Stopher & Zhang, 2011). Based on a 35-day travel survey (collected in Uppsala, Sweden, in 1971), Hanson & Huff (1982) concluded that a considerable amount of day-to-day variability exists across population groups by examining multiple travel aspect jointly. Kitamura & Van Der Hoorn (1987) specifically analyzed the travel scheduling regularity on a weekly basis using a 1984 nationwide Dutch panel dataset with repeated measurements on the same group of participants. This study confirms a systematic variability exists across weeks in terms of scheduling and trip rates. On the other hand, this study did not find significant differences between the two waves of repeated measurements taken five months apart, which indicates that variability is stable over time. Consistently, many other researches have also found that the day-to-day variance in travel patterns should not be neglected and that a multi-day travel survey at least one-week long is strongly recommended for development of models and for a more accurate understanding of travel behavior (Ettema & Lippe, 2009; Hanson & Huff, 1986, 1988; Schlich & Axhausen, 2003).

Few researchers have studied the variability from the stand point of trip chaining behavior. The limited number of research in trip chaining pattern variability may because two joint reasons: 1) the difficulty in identifying and extracting trip chaining behavior from travel diary, and 2) the scarcity of longitudinal travel datasets. Stopher & Zhang (2011) found that little repetition of tours occurs over a course of a week. The study is based on a multi-day and multi-year GPS dataset collected in Adelaide, Australia, in 2005, 2006 and 2007. The researcher first classifies home-based tours according to the number of stops in the tour (one as simple, and more than one stop as complex) and each stop's trip purpose (work, education, shopping, or other). Then, the repetitiveness of tour types is examined from a frequency perspective with repetition defined as jointly identical tour types and one of the tour attributes (i.e. total travel distance, total travel time, total activity duration, and total tour duration). The result shows a low level of repetition for both simple and complex tours. However, the repetition might be underestimated given the fact that the trip purpose used in the analysis is inferred using a program rather than reported by travelers.

Many studies further investigate the decomposition of daily travel variability. A widely-used framework, first introduced by Pas (1987), suggests that variation should be disaggregated into two parts: 1) interpersonal, which refers to differences between different individuals; and 2) intrapersonal, which means the deviation observed for the same person. Both components are further divided into a systematic subcomponent which can be explained and a random subcomponent which cannot be explained. For interpersonal variability, the systematic subcomponent is associated with individual characteristics; for intrapersonal variability, the subcomponent is related with temporal factors such as time of day and day of the week and personal social-demographical status. Further, the statistical method, analysis of variance (ANOVA), is often used for such analysis. Pas (1987) reports that intrapersonal variability explains 50% of the total variability in trip rate (based on data collected in Reading, England, in 1973). In a later study based on data from Seattle area, collected in 1989, the author finds a percentage of 38% of the total variability in trip rates is due to intrapersonal variability. The deviation between individuals is significant even for people from the same household. The Seattle study not only examines variability at the person-level but also analyzes the interaction between individuals within two-person households. The study is based on a survey where each family member above 16-year-old reports three consecutive days' travel activity and one day's activity for members below age 16. They found that 73% of the within-household variability can be explained by the between-individual subcomponent while the within-household subcomponent constitutes 62% of the total variation (Pas & Sundar, 1995).

A more recent study by Raux et al. (2016) is conducted using the same framework based on a 2008 travel survey from the area of Ghent, Belgium. The study examines the daily trip purpose sequence which is an important factor in the trip chaining field. The researcher creates strings of activities classified by purpose and uses the Sequential Alignment Method (SAM) to quantify pair-wise dissimilarity. Analysis of variation components is then applied and the study finds that: first, the total variation is lowest during weekends; second, over 70% of the variability is interpersonal variability and the rest of the variation is intrapersonal variability, day-to-day periodical deviation, and unexplained. The study also analyzes the variability in travel frequency and time allocation, and it finds intrapersonal variability contributes as much as or more than interpersonal variability. Findings suggest trip purpose patterns are relatively more habitual at an individual level than other travel attributes.

Previous research has also examined the determinants of intrapersonal variability. The level of intrapersonal variability is often hypothesized as differing across population groups due to different personal and household characteristics. However, little support for this hypothesis exists. Hanson & Huff (1982) found that two population segments, 1) non-employed married females who have pre-school age children and full-time employed husbands, and 2) these women's full-time employed husbands, do not differ significantly in level of variability or routine repetition. The study expected to see more repetition and higher spatial concentration in the travel of husbands due to their employment status, and more irregularity and spatial dispersion in non-employed wives because their travel needs were hypothesized to be dominated by household errands and childcare. However, their result suggests a similar degree of variability between the husbands and wife pairs. More specifically, the researchers did not find more concentrated destinations for the husband group. It suggests the limited power of socio-demographic characteristics in explaining travel variability. Kitamura & Van Der Hoorn (1987) also find no evidence that travel variability in trip frequency and time allocation is related with personal characteristics. However, the hypothesis of intrapersonal variability is validated in an England-based study from the perspective of trip frequency. The author concludes that the variability in trip rates is likely to be higher for people who have lower social status, have family related constrains, or have family constrains and low out-of-home activity demands (Pas & Koppelman, 1986). Nevertheless, note that all of these studies were conducted based on datasets collected many years ago on other continents than the Expo Study. Raux et al. (2016), using a more recent dataset, found that gender, age, occupation, and family size are significantly influential on intrapersonal variability in terms of daily trip frequency, activity scheduling and sequencing of travel purpose. Males, young students, single persons, or persons in a family without children are more likely to have a lower level of intrapersonal variability. Similar with Hanson & Huff (1982), Raux et al. (2016) found that employment does not show a significant effect on intrapersonal daily variability.

Earlier research on day-to-day travel variability explored approaches in variability measurement. To measure variability is equivalent to measuring similarity and thus, the general goal is to create a measurement sensitive enough to permit meaningful comparison between any pair of analysis units (e.g., a home-based tour, a sequence of one day's activity, etc.). Several different approaches have been attempted and it is necessary to discuss them to provide context for selecting the approach used in this chapter. The following discussion assumes the analysis unit is daily travel-activity pattern. The most basic method is to generate a vector consisting of aggregated characteristics for each day, such as number of trips, number of home-based tours, total length of travel duration, and activity duration. Then a simple identification of match or mismatch is made for any two days. This approach is good for descriptive comparisons but does provide a numerical measurement of similarity. This method is applied in Stopher & Zhang (2011). The second approach, originally developed by Gower (1971) in the field of biology and introduced into travel analysis research by Pas (1980), assigns a scoring system to travel attributes at a stop or destination level and sums up scores based on match/mismatch between two sequences of stops. The advantage of this method is that it can integrate levels of priorities according to the research goals and reflect the hierarchical level in the scores. The third method is developed from a two-dimensional representation of travel patterns as a time-space path. Clustering of patterns can be achieved using this method. The limitation of this method is it focuses on spatial-temporal information but not other attributes, such as travel mode. The fourth approach, proposed by Hanson & Huff (1982) can be considered as an operational extension of the second approach. The approach is a stop-based sequential measurement where a pattern is represented by a series of ordered stops (trip links) with each stop being assigned a class based on its link characteristics. Then, two patterns are compared based on attributes such as number of stops, order, and classes. Many recent research on travel variability use variations based on this fourth approach (Raux et al., 2016; C. Wilson, 2008). It is important to reemphasize that although discrepancy exists in the variability definition, measurement and research focus, a common conclusion is that a considerable amount of intrapersonal daily variability in travel patterns exists.

Research Goal and Question

The reviewed research that has focused on variability in trip chaining behavior classifies tours based on the number of legs or trips in each tour and based on the trip purpose associated with each stop within a tour (Raux et al., 2016). The previous chapter indicates tour patterns for the study sample vary at many spatial and temporal aspects. Analysis that just focuses on the number of trips per tour or uses the binary classification of tour complexity (either single or complex, depends on whether more than two trips chained in the tour), may be misleading. Moreover, GPS trajectory data is becoming more readily available nowadays, but passively recorded data, like GPS data, usually does not contain trip purpose information. This chapter contributes to the trip chaining literature by developing and testing an analysis procedure to examine trip chaining variability without reliance on trip purpose information. The Tour Complexity Index (TCI), comprised of Segment Index, Cluster Index, Efficiency Index, and Diversity Index, is applied here to characterize trip chaining behavior. See Chapter 3 for a detailed introduction of the construction of TCI. More specifically, the goals and questions that guide this chapter are as follows:

- To demonstrate the application of TCI in longitudinal analysis of travel behavior
- To examine changes in tour complexity on a daily and weekly basis
- To test the general hypothesis that variability differs across population segments
- To investigate the impact of a new transit rail line service on travel variability

Dataset and Methodology

Data and Descriptive Statistics

This chapter uses data from each of the three waves of longitudinal GPS data collected for a sample of 55 adult participants of the Expo Line Study. See Chapter 2 for a description of the data collection methods, data features and general statistics. The uniqueness of this dataset for the study in travel variability lies in that it allows for pattern comparison not only over a short period but also over three years. As pointed out who Stopher et al. (2008), who reviewed the household travel survey results around the world since 1970s, the collection of longitudinal travel survey data with multi-day and multi-year travel information for the same individuals or households is rare (13 studies out of hundreds). Although their review is not exhaustive and more datasets should become available after the review, longitudinal datasets continue to only comprise a small percentage of all travel datasets. Although the sample is small, analysis of the Expo tour dataset will contribute to improving the understanding in travel variability. This section provides a summary of the Expo Study GPS data from three perspectives that are related with the research goals of this chapter: descriptive travel statistics, tour complexity statistics and mode choice across waves.

Table 4-1 Travel Statistics of Expo Tour Dataset, Wave I - III

	Wave I	**Wave II**	**Wave III**	**Overall**
total # of trips	1,478	1,321	1,411	4,210
total # of rail trips	4	6	11	21
total # of bus trips	25	27	26	78
total # of tours	409	359	354	1,122
mean # of trips per tour	3.61	3.68	3.99	3.75
mean # of tours per day	1.06	0.93	0.92	0.97
total distance traveled per day (mile)	11.91	13.04	13.18	12.71
total VMT per day (mile)	10.68	12.13	11.96	11.59
total time on travel per day (minute)	50.73	45.61	50.51	48.95
total time on vehicle trip (minute)	38.73	35.30	36.38	36.80

There are a total of 4,210 trips and 1,122 tours in the Expo tour dataset through all three waves (Table 4-1). Only a limited number of trips are undertaken by public transit (bus or rail transit) and there is only a modest rise in rail usage after the opening of Expo Line (after Wave I). On average across the three waves, participants make 0.97 tour per day and travel 12.71 miles for 48.95 minutes, within which 11.59 miles and 36.8 minutes are by vehicle. The number of tours declined after Wave I, but more trips are taken and more mileages are traveled during Wave III than Wave II, suggesting an increase in trip chaining.

Types of Tour Complexity and Frequency

Table 4-2 provides summary statistics of the TCI components for each wave of data collection. The overall number of segment per tour is 3.75; 44% of tours have clustered destinations; 78% of tours are chained and configured using efficient or compact routing; and 53% of tours involve a diverse land use environment across destinations. Each of the three binary component indices includes a reasonable share of all the tours. These averages do not show drastic changes across waves, but we can see an increase in the number of segments per tour and the percentage of tours which include clustered destinations, although the total number of tours per wave declines by 13.4%. Details about TCI are provided in Chapter 3.

Table 4-2 Mean Value of TCI Components for Expo Tour Dataset, Wave I - III

	Wave I	**Wave II**	**Wave III**	**Overall**
Segment (average)	3.61	3.68	3.99	3.75
Cluster (% of 1)	0.41	0.46	0.47	0.44
Efficiency (% of 1)	0.80	0.77	0.78	0.78
Diversity (% of 1)	0.53	0.52	0.55	0.53
Number of Tours	409	359	354	1,122

Next, tours are classified into types based on the TCI components. For parsimony, the Segment Index of TCI is defined by converting the segment count variable into a categorical form defined by the following three levels: 1) “single destination” for tours with only one destination; 2) “simple chain” for tours with two or three destinations; 3) “complex chain” for tours with four or more destinations. By definition, all single destination tours are efficient in routing and do not have diversity at destination (Segment = “single”, Efficiency = 1, Diversity = 0). Hence, there are 18 possible combinations of TCI values in total, 2 for single destination, 8 for simple chain and 8 for complex chain (Table 4-3). In the following analysis, each combination of the TCI components is referred to as a tour complexity type. The sample-wide occurrences of each tour complexity type per wave are summarized in Table 4-3.

Table 4-3 Frequency of Tours by Tour Complexity

Tour Complexity Index				Wave I	Wave II	Wave III	All	
Segment	**Cluster**	**Div.**	**Eff.**	**Count**	**Count**	**Count**	**Count**	**%**
Single destination	0	0	1	121	93	97	311	27.7%
	1	0	1	27	45	33	105	9.4%
						Subtotal	*416*	*37.1%*
Simple chained (2-3 dest.)	0	0	0	1	2	1	4	0.4%
	0	0	1	14	13	11	38	3.4%
	0	1	0	10	9	2	21	1.9%
	0	1	1	71	48	51	170	15.2%
	1	0	0	2	2	1	5	0.4%
	1	0	1	26	14	14	54	4.8%
	1	1	0	3	4	10	17	1.5%
	1	1	1	45	44	38	127	11.3%
						Subtotal	*436*	*38.9%*
Complex chained (4+ dest.)	0	0	0	0	1	1	2	0.2%
	0	0	1	0	0	0	0	0.0%
	0	1	0	17	22	21	60	5.3%
	0	1	1	8	5	5	18	1.6%
	1	0	0	0	1	1	2	0.2%
	1	0	1	1	1	0	2	0.2%
	1	1	0	49	43	42	134	11.9%
	1	1	1	14	12	26	52	4.6%
						Subtotal	*270*	*24.1%*
Total				409	359	354	1,122	100%

First of all, the Segment Index results in a good split of the tours, as each of the three levels includes a reasonable share of tours. Single destination tours and simple chained tours each comprise a share of about 38% and complex chained tours comprise 24%. The percentage of single destination tours is similar with that reported by Raux et al. (2016) who found single destination tours comprised 41% of their observed tours. Both of these estimates of the percentage of single destination tours, however, are lower than that reported by an Australia-based trip chaining study which found that single destination tours comprised 53% of work tours and 64% of non-work tours (Currie & Delbosc, 2011).

Breaking down the analysis to the 18 tour complexity types, there are six types that each comprises a share over 5% of all tours in all waves of data collection, two from each of the three levels of the Segment Index. Within each level, one of these dominant tour complexity types has clustered destinations (Cluster = 1) and the other one does not contain clustered destinations (Cluster = 0). This suggests that the status of tours (defined by the Cluster Index) is the second most effective criteria for characterize tours after the Segment Index. The two dominant types of simple chained tours both have a Diversity Index of 1 and an Efficiency Index of 1. This indicates simple chained tours are more likely to have efficient routing, and to include destinations in areas with different dominant land use types. On the other hand, the two dominant types for complex chained tours both have a Diversity Index of 1 and an Efficiency Index of 0. This means complex chained tours tend to have repeated visits or a zigzagging route arrangement. This seems reasonable since as more destinations are chained, it becomes a greater challenge to ensure an efficient route arrangement but more likely a tour will involve multiple types of destinations. Overall, there seems to a good mixture of tours in the study sample in terms of tour complexity type.

In addition, not all tour complexity types appear in the Expo tour dataset or in each wave. The type of {complex chain, no cluster, efficient routing, and single land use at destinations} is missing in all three waves, and the types with {complex chain, single land use at destinations} have a very low occurrence rate. However, it is necessary to clarify that although it appears that the Segment Index and Cluster Index appear sufficient to classify tours in the Expo tour dataset, the importance of Diversity Index and Efficiency Index cannot be dismissed. The Expo tour dataset includes a limited sample size and representativeness, and an application of TCI on a larger scale dataset or another population may reveal that all four components play an important role.

Mode Usage

Table 4-4 groups the 55 participants based on whether they used a travel mode at least once during the surveyed weeks. With three mode types, there are seven possible combinations of mode usage, including three single usage types (only one mode has been used during the week) and four mixed types (two or three modes have been used during the week). The arrow next to the count indicates the trend compared to the previous wave. The mode is flagged for a traveler as long as it has been used at least once during the observation week in each wave, regardless of the total frequency or whether it is a primary mode of a trip. Transit mode refers to both bus or rail. Non-motorized mode refers to walking or biking. Though walking and biking are coded separately during data processing, very few biking trips occur so they are merged with walking into one mode, "non-motorized", for the convenience of analysis. Likewise, bus and rail are merged as "transit". The walking period to access a transit station or vehicle is not considered as a separate trip.

Vehicle usage is dominant in the study sample (Table 4-4). A large portion of people only used vehicle and did not undertake even a single trip during the week using transit, walking or biking. However, such vehicle reliance steadily declined from 51% to 45% and to 38% of participants for waves I, II, and III, respectively. During this same time, the percentage of participants who mixed vehicle with transit or walking/biking has increased slightly. No participant choose transit as the only travel mode, but 11 participants (20%) mixed transit with other modes during the survey week for Wave I and Wave II and this number increased to 15 participants (27.3%) in Wave III.

Table 4-4 Percentage of Study Subjects based on Mode Usage

Mode Choice	**Wave I**		**Wave II**		**Wave III**		**All**	
	Count	**%**	**Count**	**%**	**Count**	**%**	**Count**	**%**
Single								
Vehicle	28	50.9%	25 ↓	45.5%	21 ↓	38.2%	10	18.2%
Transit	0	0.0%	0	0.0%	0	0.0%	0	0.0%
Non-motorized (NMT)	0	0.0%	1 ↑	1.8%	1	1.8%	0	0.0%
Mixed								
Vehicle + Transit	1	1.8%	2 ↑	3.6%	3 ↑	5.5%	1	1.8%
Vehicle + NMT	16	29.1%	18 ↑	32.7%	18	32.7%	25	45.5%
Transit + NMT	5	9.1%	5	9.1%	5	9.1%	3	5.5%
Transit + Vehicle + NMT	5	9.1%	4 ↓	7.3%	7 ↑	12.7%	16	29.1%
Total	55	100%	55	100%	55	100%	55	100%

Interestingly, the statistics with all three waves merged are quite different from those disaggregated by wave (Table 4-4, "All" column). For instance, the number of participants who used a vehicle as the only mode for travel through the three weeks reduces to 10 participants (18%), while the number of vehicle users who have taken at least one transit trip or walking/biking trip increases to 42 participants (76%)! The number of participants using transit at least once across the entire three waves of data collection doubled from 11 participants to 20 participants. The number of non-vehicle users are almost identical during each wave (5 for Wave I, and 6 for Wave II and III), but it reduces to 3 if we look at the number aggregately for all waves. These deviations suggest the mode choice varies across waves at individual level. and that we might expect to see high intrapersonal variation in mode choice.

Scoring of Daily and Weekly Variability in Tour Activity

The goal of this chapter is to analyze the variability in trip chaining behavior over a day and a week, with tours classified by TCI and travel mode. The procedure to quantify variability follows the fourth general approach for analyzing day-to-day travel variability discussed in the Literature Review section above which has been utilized by Hanson & Huff (1982), Raux et al. (2016), Wilson (2008). It uses the Sequential Alignment Method (SAM) to quantify the difference between any of two strings, each formed by tour complexity types undertaken during one day. Below, I discuss an example based on a three-week travel record to demonstrate how the daily and weekly travel variability is measured and how to interpret the numbers. Table 4-7 shows a hypothetical one-week-long travel record at a tour level, with the tours characterized based on the Tour Complex Index (TCI) components and travel mode. The coding representation is listed in Table 4-5 and Table 4-6. The record is same as the Week 3 record in Table 4-8. Table 4-8 presents a hypothetical three-week-long travel record, listing each day's sequence of home-based touring activity and each day's variability score.

In the first step (Table 4-7), each tour is characterized using TCI and then assigned a code to jointly represent the tour type and tour travel mode. For instance, a simple chained tour with clustered destination, multiple land use environments and chained via a shortest path, is a type J tour. Such a tour is then labeled as "J1" if by rail, "J2" if by bus, "J3" if by vehicle and "J4" if by NMT. See Table 4-5 and Table 4-6 for code dictionary. This procedure could be applied to tour type and mode choice separately if interested in analyzing the variability for a single factor. It can also be extended by replacing or adding other travel attributes such as distance, duration, and departure time. The analysis here focuses on the joint choice of tour complexity type and travel mode. Note that one should be cautious about not only how the variability is defined and measured, but also the definitions of other important aspects of travel pattern matter such as departure and arrival times. The level of variability and length of repetition cycles resulting from the analysis depend on these classifications (Hanson & Huff, 1982). The major reason that travel mode is integrated into the classification scheme is because the Expo Study sought to assess the impact of a change in the study area's transit service supply system. As we are primarily interested in the variability or repetition in complex trip chaining behavior, the proposed measurement is chosen to be able to capture switches in travel mode as well to analyze the association between tour complexity and mode choice.

In the second step (Table 4-8), the tour-level records are flattened into daily sequences by creating a string of tours made during each day. For example, if there is no travel on a given day day, a code of "S0" is assigned to that day for the purpose of similarity comparisons. Then, the pairwise dissimilarity between daily sequences are computed. The variability of a certain day is determined based on its aggregated dissimilarity with other days. The dissimilarity is measured using one-dimensional SAM, a dynamic programming technique originally designed to compare protein or DNA sequences. It has been shown to be effective in comparing travel patterns (Joh et al., 2001; Wilson, 1998). The goal is to find the optimal (least) number of three basic operations, i.e., insertion, deletion and substitution, to equalize two sequences. The more operations needed to equalize two sequences, the higher dissimilarity between them. For instance, consider two tour sequence, "F3, A4, F3" and "J3, F3, A1". To equalize these two, the operations are: 1) delete first "F3" in the former sequence (that is, "~~F3,~~ A4, F3"); 2) substitute "A4" with "J3" (that is, "~~A4~~ (J3), F3"); and 3) insert "A1" in the end (that is "J3, F3, (A1)"). Thus, the least number of operations is 3 to equalize the sequences. Note that the operations, insertion and deletion always come as a pair and is often referred to as *indel* (a term from molecular biology for the insertion or the deletion of bases in the DNA of an organism).

The transformation from number of operations to a numerical dissimilarity measure is a more complicated step. The distance between two sequences usually is generalized to a function of a number of operations (4-1). The operations could be weighted differently, for example, as the cost of *indel* defined to be more than substitution. In this study, the default scoring system is used: cost of *indel* and substitution all equal to 1. Hence, the distance between two daily sequences directly equals the total number of operations (4-2). The variability of a day compared to the rest of the week is defined as the summation of its distance to other days within the same week for the same person (4-3). Moreover, multi-dimension SAM could be used to add in more travel dimensions (i.e., to create separate sequences for tour complexity, tour mode, activity duration, etc. per day), and to match two groups of sequences simultaneously. Although the multi-dimension approach is more comprehensive, the technical solution is very complex and hard to justify (Joh et al., 2002). Therefore, a one dimension SAM is applied here. In addition, the weight of tour complexity type and mode choice are the same. The distance between "J2" and "A2" (different tour types, same mode) is the same as the distance between "J2" and "J3" (same tour type, different modes). The tour type and mode choice are treated as nominal factors, which means any two tour types or any two modes has the same degree of dissimilarity.

$$d_{ij} = f(n_{insert}, n_{delete}, n_{substitue}) \tag{4-1}$$

$$d_{ij} = n_{insert} + n_{delete} + n_{substitue} \tag{4-2}$$

$$V_{ik} = \sum_{j=1;j\neq i}^{J} d_{ijk} \tag{4-3}$$

where

d_{ij} = distance between daily tour sequence *i* and *j*

V_{ik} = variability of day *i* in week (wave) *k*

The three-week example in Table 4-8 is designed to represent three levels of weekly variability. For Week 1, the travel pattern is quite simple. The person is characterized as a full-time worker who goes to work, has lunch at a place close to the work location, has minimum (zero) out-of-home, non-work activity once he/she returns home, makes personal trips on Saturday and stays at home on Sunday. In terms of tour complexity type classifications, he/she performs the same type of tour in same travel mode once per day from Monday to Friday (type “J3”, simple chain, cluster, diverse destination environment, efficient route, by vehicle). He/she makes two tours on Saturday, one type J and one type A (single destination, long distance), both in vehicle (“J3, A3”). For Sunday, he/she stays at home (“S0”). As a result, the overall variability is low (3.43). The daily variability is low as well, with most individual days having a variability level of 2 and higher during weekends (7). For Week 2, same person makes the same trips as Week 1 but uses different travel modes. The person in week 2 uses rail on Monday, bus on Tuesday and car for the rest of the days. As a result, the weekly average variability increases by 2.57 and variability for individual days increases as well, especially for Monday and Tuesday. For Week 3, there is no clear travel pattern across days as every day with out-of-home activity has different sequences. In this case, the overall variability (12.57) and daily variability are both much higher than the those for Week 1 and Week2.

Three general rules can be observed from this hypothetical example:

1. When there are more tours in one day, there is a higher level of variability relative to other days in the week, since it requires more operations to equalize the day with other days
2. The variability of one day is dependent on other days in the week, because the dissimilarity measure if mutually applied on both sequences; that is, if one one day there are several tours undertaken but the other days all have an identical sequence with each other, the scores for these matching days will be inflated by the one day with more tours

3. When there are more different types of tours spread in a week, then that week has a higher overall variability because every daily sequence requires a certain number of operations to match with other sequences in that week

Table 4-5 Tour Type Code

Tour Complexity Index				Tour Code
segment	**cluster**	**diversity**	**efficiency**	
Single destination	0	0	1	A
	1	0	1	B
Simple Chain (2-3 dest.)	0	0	0	C
	0	0	1	D
	0	1	0	E
	0	1	1	F
	1	0	0	G
	1	0	1	H
	1	1	0	I
	1	1	1	J
Complex Chain (4+ dest.)	0	0	0	K
	0	0	1	L
	0	1	0	M
	0	1	1	N
	1	0	0	O
	1	0	1	P
	1	1	0	Q
	1	1	1	R
Staying at home				S

Table 4-6 Mode Code

Mode	Mode Code
Rail	1
Bus	2
Vehicle	3
Non-motorized (NMT)	4
Staying at home	0

Table 4-7 Example of Weekly Tour-level Travel Record

	Tour Seq. Number	Tour Complexity Index				Mode	Class Code
		Segment	Cluster	Diversity	Efficiency		
Monday	-	-	-	-	-	-	S0
Tuesday	-	-	-	-	-	-	S0
Wednesday	1	complex	1	1	0	vehicle	Q3
	2	simple	1	1	0	vehicle	I3
Thursday	1	simple	0	1	1	vehicle	F3
	2	single	1	0	1	NMT	B4
	3	simple	0	1	1	vehicle	F3
Friday	-	-	-	-	-	-	S0
Saturday	1	simple	0	1	1	vehicle	F3
Sunday	1	single	1	0	1	vehicle	B3
	2	simple	0	0	1	vehicle	D3
	3	complex	1	1	1	vehicle	R3

Table 4-8 Example of Weekly Day-level Travel Record

Week	Day of Week	Sequence of Tour	Variability	Description
Week 1	Monday	J3	2	• Low variability • Same routine on weekdays • Single travel mode • Mean: 3.43
	Tuesday	J3	2	
	Wednesday	J3	2	
	Thursday	J3	2	
	Friday	J3	2	
	Saturday	A3, J3	7	
	Sunday	S0	7	
Week 2	Monday	J1	7	• Medium variability • Same routine on weekdays • Three modes • Mean: 6.00
	Tuesday	J2	7	
	Wednesday	J3	4	
	Thursday	J3	4	
	Friday	J3	4	
	Saturday	A3, J3	9	
	Sunday	S0	7	
Week 3*	Monday	S0	9	• High variability • No clear pattern • Various tour types • Two modes • Mean: 12.57
	Tuesday	S0	9	
	Wednesday	Q3, I3	14	
	Thursday	F3, A4, F3	18	
	Friday	S0	9	
	Saturday	F3	11	
	Sunday	B3, D3, J3	18	

**Week 3 depicts the same time period presented in Table 4-7.*

Decomposition of Variance in Variability

There are two methods that can be applied here to quantify the relationship between external factors and the travel pattern variability. One is generalized linear regression and the other is analysis of variance (ANOVA). The underlying least square algorithm is exactly the same for the two methods. When the independent variables are all categorical, the linear regression model provides coefficients for each level of the variables while ANOVA treats factor as a whole by providing an overall measure of the variation explained by the factor. For this reason, the linear model is more suitable if the intention is to examine the effects brought by each level within the factor, and ANOVA is more suitable if the intention is to examine the effects of each factor. The research focus of this analysis component is whether factors affect the variability in complex trip chaining. For instance, to assess how much variation among the observed daily variability is introduced by intrapersonal variation, rather than to obtain a specific number about the difference between 2 different participants. For this reason, ANOVA serves to answer the research question better.

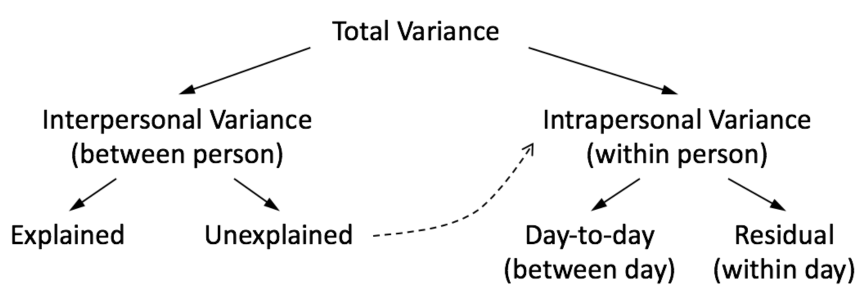


Figure 4-1 Diagram of Travel Behavior Variance Decomposition

Source: Pas (1987)

This study follows the decomposition approach introduced by Pas (1987) which is developed based on ANOVA principle. The concept is illustrated in Figure 4-1 above. The author states that the observed variance of certain travel attributes (e.g., daily trip rates, total travel time, etc.) from a population can be traced to interpersonal variance (difference between individuals) and intrapersonal variance (deviation for one individual). The interpersonal variance is influenced by the social-economical status of the individual as different population groups are expected to have different travel needs. The portion of variance that can be accounted for by person's socio-economic status is called "explained" portion of the variance, and the rest of the variance is referred to as "unexplained". The intrapersonal variance reflects day-to-day differences that are influenced by an individual's activity and travel needs. The simplest example is that the travel needs on weekdays and weekends are different for workers, students, and people whose travel needs are closely influenced by workers and students. Given a longitudinal dataset, this individual day-to-day variance can be quantified while the remainder is termed as "residual" (Pas & Koppelman, 1986; Pas & Sundar, 1995). One issue should be clarified here that in the original study by Pas, the term "variability" is used instead of the term "variance". However, we want to avoid any confusion about the weekly trip chaining variability that will be studied and the "variability" in Pas' work. Hence, the statistical term "variance" is adopted here to refer to Pas' "variability". Essentially, the "variability" discussed in Pas' work is based on the concept of "variance".

Numerically, this concept diagram (Figure 4-1) can be expressed using the following equations. Notice that the unexplained portion of interpersonal variability is passed to intrapersonal variance and is folded into the intrapersonal residual from the modeling perspective.

$$SST = SSBP + SSWP \tag{4-4}$$

$$SSWP = SSBD + SSWD \tag{4-5}$$

$$SST = \sum_{k=1}^{K} \sum_{i=1}^{I} \sum_{j=1}^{J} \left(s_{kij} - \bar{s}\right)^2 \tag{4-6}$$

$$SSBP = \sum_{i=1}^{I} (\bar{s}_i - \bar{s})^2 \tag{4-7}$$

$$SSBD = \sum_{k=1}^{K} \sum_{i=1}^{I} \sum_{j=1}^{J} \left(\bar{s}_{ij} - \bar{s}_i\right)^2 \tag{4-8}$$

$$SSWD = \sum_{k=1}^{K} \sum_{i=1}^{I} \sum_{j=1}^{J} \left(s_{kij} - \bar{s}_{ij}\right)^2 \tag{4-9}$$

where SST means total sum of squares, SSBP means between person sum of squares, SSWP means within person sum of squares, SSBD means between day sum of squares, and SSWD means within day sum of squares; k is the index for week if the data contains records for more one week, i is the individual index, and j is the day-of-week index.

Figure 4-2 provides an example of the variance decomposition framework using daily trip rates. The result is obtained based on Expo Study GPS study sample of 55 individuals' three-week-long travel records. For each individual week and day of week, there are three repeated measurements. The result shows 21% of the variance in daily trip rates is interpersonal (between individuals), 24% is the systematic day-to-day variance within individual while 55% of the total variance is considered as residual that cannot be explained by either social-demographical factors or day-of-week factors.

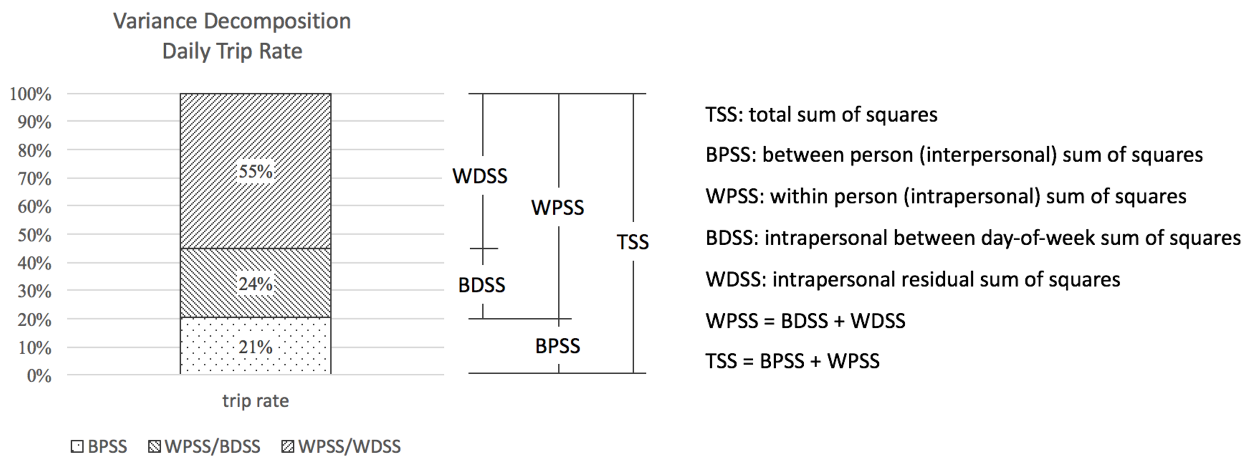


Figure 4-2 Illustration of the Concept of Variance Decomposition

Data source: Expo tour dataset, Wave I - III (n=55)

The method by Pas (1987) has been validated by the author in two studies on different datasets, one with data collected from Reading, England, in 1973 and the other one with data collected from North King County, Washington, in 1989 (Pas, 1987; Pas & Sundar, 1995). However, only a limited number of research has adopted this method, probably due to the difficulty in obtaining longitudinal datasets since the decomposition framework requires repeated measurements of individual patterns over time. In addition, none of the studies targets patterns in the Los Angeles area. Hence, it is important to test whether the method is applicable to a more recent dataset and in a new urban context.

Results

Daily and Weekly Travel Variability

On a daily basis, the daily variability scores range from minimum of 1 and maximum of 40 while the weekly variability ranges from 1.71 to 21.14. The overall average variability is 7.78. As shown in Table 4-9, the percentiles and means of the daily and weekly averages are very similar. As discussed in the Methodology section, the daily variability is dependent on the patterns of other days in the week. It would be more meaningful to compare the variability within the same week. The day of the week with the lowest variability score means the tour-mode classes undertaken during that day have the highest degree of repetition with those for other days during that week. The day of the week with the highest variability score indicates the tour-mode classes for that day are quite different from the classes of other days during that week. Minimum and maximum scores can appear on more than one day during the week. In this case, the pattern for the days with the same lowest score can be considered as the typical pattern for a person. Table 4-10 and Figure 4-4 present the probability that a given day of the week will have the lowest variability (min.) and the probability that a given day of the week will have the highest variability in that week (max.). The probabilities are different cross days. It appears that the middle of the week, Thursday, is least likely to deviate from the rest of the week. On the other hand, the probability of being the most different day of the week is highest on Monday followed by Saturday. These patterns could be because participants conduct most of the errands and non-regular travel demands either during weekend or the beginning of the week.

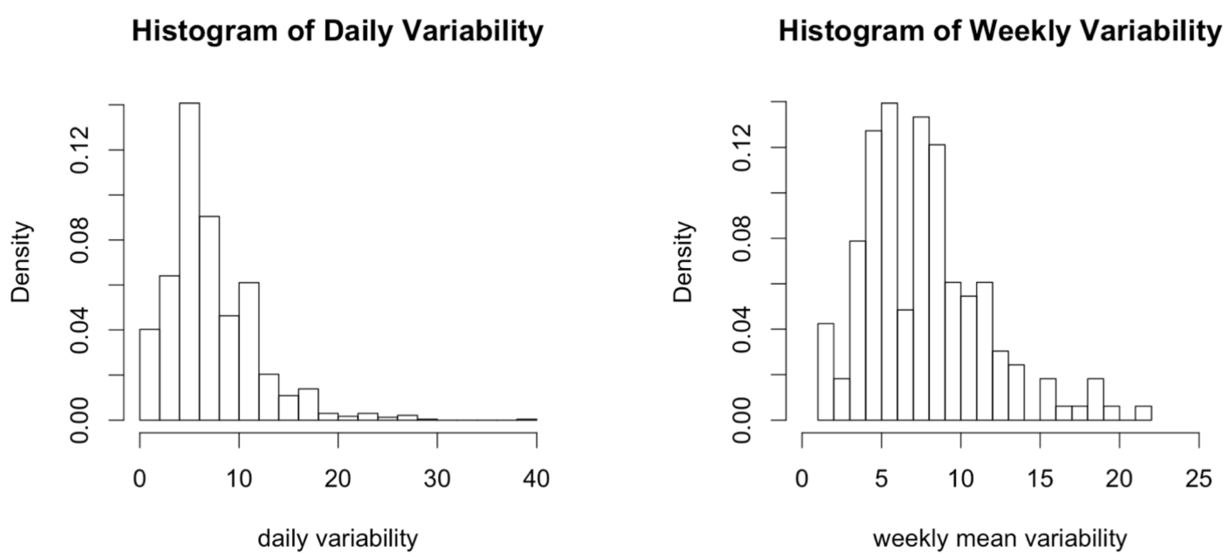


Figure 4-3 Histogram of Variability

Table 4-9 Summary Statistics of Variability

	Min.	Q_1	Median	Mean	Q_3	Max.
daily	1.00	5.00	7.00	7.78	10.00	40.00
weekly avg.	1.71	4.86	7.43	7.78	9.71	21.14

Table 4-10 Probability of Being Min./max. by Day of Week

	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
Prob(min.)	0.42	0.48	0.48	0.52	0.48	0.43	0.42
Prob(max.)	0.30	0.25	0.27	0.22	0.24	0.28	0.24

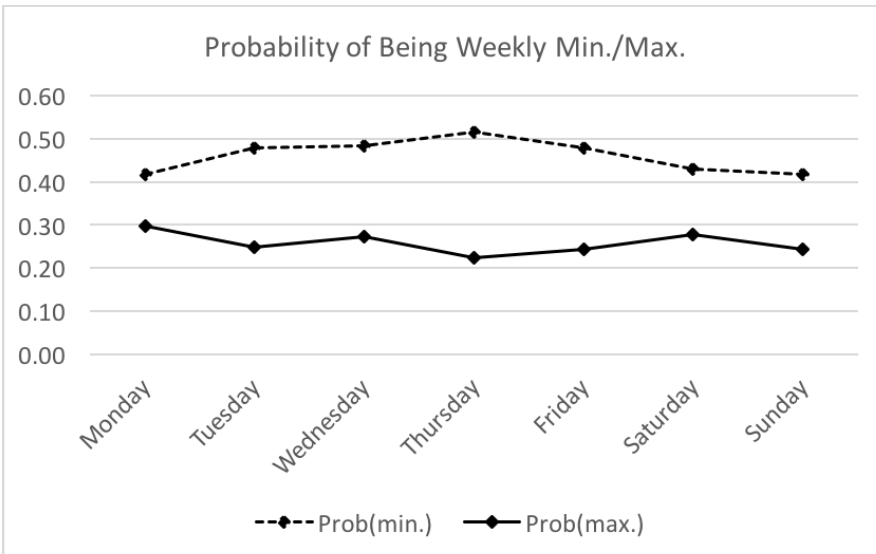


Figure 4-4 Probability of Being Min./max. by Day of Week

Source of Variation

The purpose of the three-way ANOVA analysis (Table 4-11) is to test the interaction and main effect from individual, day of week, and year on the daily pattern variability. The traveler variable is included as a fixed block factor to control for natural variability among people. Wave, a factor with 3 levels, is designed to reduce the bias brought by different survey periods. Day of week is a factor with 7 levels. It is included to control for the periodical change in travel behavior throughout the week. The main effect of intrapersonal deviance (traveler) and its interaction with wave on daily travel variability are both statistically significant at 99% confidence level. They explain 70%

of the total variation in the data. The main effect of wave has significant impact on daily variability as well but only explains 1% of the variation. Day of week does not have strong effect on daily variability based on this dataset, which means the variability is not statistically significantly different by day of week after other factors are controlled for. The lack of association between day of week and daily variability is contradictory to the findings of Kitamura & Van Der Hoorn (1987) and Raux et al. (2016). These studies have found statistically significant difference between weekday and weekend. This disagreement may be due to different time points and geographical regions of data collection.

Table 4-11 Three-way Anova Table for Daily Trip Chaining Pattern Variability

Source	d.f.	SS	MS	F-value	p-value	SS/SST
Traveler	54	9440.46	174.82	23.895	< 0.001***	38.56%
Wave	2	255.52	127.76	17.462	< 0.001***	1.04%
Day of week	1	4.00	4.00	0.547	0.460	0.02%
Traveler : Wave	108	7427.53	68.77	9.400	< 0.001***	30.34%
Traveler : Day of week	54	521.14	9.65	1.319	0.065*	2.13%
Wave : Day of Week	2	5.58	2.79	0.381	0.6832	0.02%
Residuals	933	6826.14	7.32	-	-	27.88%
Total SS (SST)	1154	24480.37	-	-	-	100%

*Significance level: 0.01 ***; 0.05 **; 0.1 **

Travel Variability Before-and-after Expo

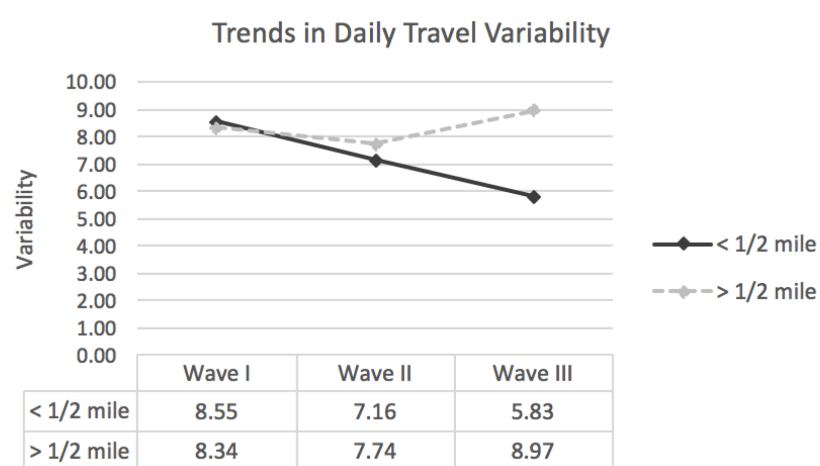


Figure 4-5 Trends in Weekly Mean Variability by Distance to Expo Stations

As Figure 4-5 shown, participants who live in different areas defined by distance to Expo stations exhibit different levels and trends of daily travel variability. During Wave I, people who live within ½ mile from any Expo station in the study area (experiment group) and people who live further away (control group) have almost identical mean daily variability. However, the mean variability for the experiment group steadily declines (black solid line) from Wave I to Wave III, while the mean variability for people in the control group changes through waves but does not show a linear trend (gray dash line). By the time of Wave III, control group has an average variability of 8.97, 3.14 higher than that of experiment group (average 5.83). In this part of the analysis, we use analysis of variance (ANOVA) methods to examine: 1) whether the difference between two groups are significant; 2) whether the change through waves are significant; and 3) whether the variability among individuals within the same distance group is large.

First, a two-way and nested ANOVA is performed to assess the overall main effects and interactions between the spatial-temporal factors and the daily travel variability. The spatial factor is distance to an Expo station from the participant's residential location, with two levels: experiment and control group. The temporal factor is the wave in which the observation occurred, with three levels: Wave I, Wave II, and Wave III. As the previous analysis on source of variance shows, the intrapersonal variability is significant and cannot be ignored. Hence, it is necessary to add traveler to the model as blocks to reduce the variance brought by individual difference. Since each traveler can only be in one distance group, the traveler variable is nested under the distance factor and cannot have an interaction with distance.

The two-way ANOVA model is formulated as follows:

$$d_{aijk} = \mu_{..} + \gamma_a + \alpha_i + (\gamma\alpha)_{ai} + (\gamma\beta)_{a(i)j} + \epsilon_{aijk} \tag{4-10}$$

where

d_{aijk} = alignment distance of day *k* by traveler *j* in area *i* during wave *a*

$\mu_{...}$ = overall mean (per wave)

γ_a = how wave *a* differs from the overall mean, $\mu_{a..} - \mu_{...}$

α_i = how group *i* differs from the overall mean, $\mu_{.i.} - \mu_{...}$

$(\gamma\alpha)_{ai}$ = how group *i* during wave a differs from the overall group mean and wave mean, $\mu_{ai\cdot} - \mu_{a..} - \mu_{.i.} + \mu_{...}$

$(\gamma\beta)_{a(i)j}$ = how traveler *j* at wave *a* differs from the overall traveler mean and wave mean, $\mu_{aij} - \mu_{a..} - \mu_{.(i)j} + \mu_{...}$

and $\Sigma\gamma_a = \Sigma\alpha_i = \Sigma(\gamma\alpha)_{ai} = \Sigma(\gamma\beta)_{a(i)j} = 0$

Table 4-12 Two-way Anova Table for Daily Trip Chaining Variability

Source	d.f.	SS	MS	F value	p-value	SS/SST
Wave	2	255.52	127.76	17.19	< 0.001***	1.04%
Distance to Expo	1	394.25	394.25	53.05	< 0.001***	1.61%
Wave: Dist.	2	590.49	295.25	39.73	< 0.001***	2.41%
Wave: Traveler	159	15883.25	99.89	13.44	< 0.001***	64.88%
Residuals	990	7356.86	7.43	-	-	30.05%
Total SS (SST)	1154	24480.37	-	-	-	100.00%

*Significance level: 0.01 ***; 0.05 **; 0.1 **

The results from the overall model shows the impact from spatial-temporal factors and the traveler variable on the daily travel variability is significant, given all p-values smaller than 0.001 (Table 4-12). However, the percentage of variance explained is quite different across effect components. The interaction between wave and traveler explains 65% of the total variance, followed by the distance factor which explains 5%. It provides strong evidence that although residential location may be associated with the daily travel variability, the difference among people is much more important in explaining the population variation in daily travel patterns. On the other hand, with significant interactions, it is meaningless to discuss the main effect based on this model. So nested a ANOVA model is performed separately for each wave to examine the main effect of distance to Expo station on the daily variability.

The nested anova model is formulated as follows:

$$d_{ijk} = \mu_{..} + \alpha_i + \beta_{(i)j} + \epsilon_{ijk} \tag{4-11}$$

where

d_{ijk} = alignment distance of day k by traveler j in area i

$\mu_{..}$ = overall mean (per wave)

α_i = how group i differs from the overall mean, $\mu_{i.} - \mu_{..}$

$\beta_{(i)j}$ = how traveler j differs from the group mean, $\mu_{ij} - \mu_{i.}$

and $\Sigma\alpha_i = \Sigma\beta_{(i)j} = 0$

The result from the nested ANOVA models (Table 4-13) is consistent with the previous models and the pattern displayed in Figure 4-5. First of all, the intrapersonal difference is still the most powerful component in explaining the variance of daily travel variability among the entire sample. Secondly, the main effect of the distance factor becomes more significant from Wave I to Wave III. The distance factor only covers 0.05% of the total variance and is statistically insignificant during Wave I. The percentage of variation explained increases to 0.4% in Wave II and further increases to 12% in Wave III. During Wave II, the main effect of distance is significant at 95%

confidence level. It becomes statistically significant at 99% confidence level during Wave III. This indicates a systematic impact from the residential environment measured as proximity to an Expo Station on daily travel variability.

Table 4-13 Nested Anova Table for Daily Trip Chaining Variability by Wave

Source	d.f.	SS	MS	F-value	p-value	SS/SST
Wave I						
Distance to Expo	1	4.39	4.39	0.50	0.473	0.05%
Traveler	53	5202.26	98.16	11.22	< 0.001***	61.03%
Residuals	330	2888.29	8.75	-	-	33.88%
Total SS	384	8523.99	-	-	-	
Wave II						
Distance to Expo	1	32.47	32.47	4.70	0.031**	0.39%
Traveler	53	5862.02	110.60	16.02	< 0.001***	70.36%
Residuals	330	2279.14	6.91	-	-	27.36%
Total SS	384	8331.36	-	-	-	
Wave III						
Distance to Expo	1	947.89	947.89	142.87	< 0.001***	11.80%
Traveler	53	4818.97	90.92	13.70	< 0.001***	60.00%
Residuals	330	2189.43	6.63	-	-	27.26%
Total SS	384	8031.78	-	-	-	

*Significance level: 0.01 ***; 0.05 **; 0.1 **

The above models confirm significant relationships between the travel behavioral variability and individual deviance. The next question is how the daily travel variability of the two groups differs per wave and how daily travel variability of the same group changes through waves. Tukey's method is applied to compute the 95% confidence interval of each contrasts. The numerical result is presented in Table 4-14 and visualized in Figure 4-6. Within the experiment group, all intervals do not contain zero. During Wave II and Wave III, there is statistically significant decrease in daily travel variability. As a result, the net decrease between Wave I and III is also significant. Within the control group, Wave II shows a slight decrease in daily travel variability but is followed by a small increase during Wave III. Although the increase in Wave III is statistically significant at 99% level, the overall change is insignificant. Between groups, during Wave I and Wave II, the differences in daily variability from participants in the two distance groups are not significant, but zero is moving towards the boundary. During Wave III, the mean daily travel variability for the group within proximity to Expo stations is much lower than the other group (significant at 99% level). In sum, people in the experiment group were associated with a drop in daily travel variability from Wave I to Wave III after the opening of Expo line. Compared to the control group, the experiment group has a more stable and predictable travel pattern at daily level at the time of Wave III.

Table 4-14 Pair-wise Confidence Interval for Variability Difference by Distance Group

Contrasts	Difference (Variability)	95% C.I.		p-value
		Lower	Upper	
Within Experiment Group				
Wave II – Wave I	-1.386	-2.187	-0.586	< 0.001***
Wave III – Wave I	-2.720	-3.520	-1.919	< 0.001***
Wave III – Wave II	-1.333	-2.134	-0.533	< 0.001***
Within Control Group				
Wave II – Wave I	-0.592	-1.378	0.194	0.199
Wave III – Wave I	0.633	-0.154	1.419	0.347
Wave III – Wave II	1.224	0.438	2.011	< 0.001***
Between Groups: Experiment - Control				
Wave I	0.214	-0.580	1.007	0.970
Wave II	-0.581	-1.374	0.213	0.284
Wave III	-3.139	-3.932	-2.345	< 0.001***

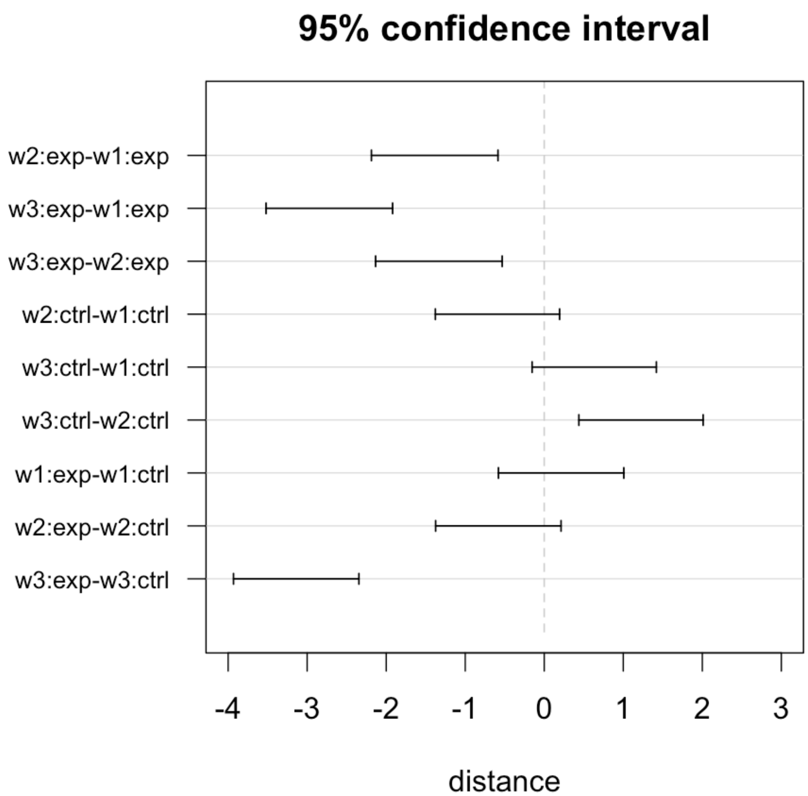


Figure 4-6 Simultaneous Confidence Interval for Variability by Distance to Expo Stations

Travel Variability and Socio-demographic Factors

This section investigates whether there are confounding factors that are associated with the significant decrease in travel variability of the experiment group, socio-demographic factors in particular. Unbalanced two-factor ANOVA is applied and uses Type II Sum of Squares (SS). Type II SS computes the SS explained by the factor given other factors co-exist in the model. The other two types are: 1) Type I SS computes a sequential SS, which means the SS for a factor is conditioned on other factors that enter the model before it but it is independent of the factors after it; 2) Type III SS computes a SS as if the factor is the only variable in the model. Both Type II and Type III SS is order insensitive while order matters in Type I SS. Here, we are interested in the explanatory power of the socio-demographic factors and the distance to Expo factor conditioned on each other, i.e., after the variance from the distance factor is accounted for, whether the socio-demographic factors still contribute a large degree of explanatory power. For this reason, the Type II SS method is more suitable.

The socio-demographic factors examined are: gender, age, car ownership, work/study schedule, household income, household size, and family life cycle. Seven two-way ANOVA models are constructed for each wave of data. In each model, the main effects of socio-demographic factors and the distance factor ("control" and "experiment group") and their interaction are tested against the average weekly travel pattern variability at person level (Table 4-16).

During Wave I, none of the main effects and interactions show statistically significant reductions of the total variation in the mean weekly variability across participants. This overall insignificance is within expectations since the major change analyzed, the opening of Expo Line, had not yet occurred. During Wave II after Expo Line service began, both the main effect of family life cycle and its interaction with the distance factor show some significance in explaining the weekly variability. However, during Wave III, the distance grouping factor shows significant impact on weekly travel pattern variability at the 99% confidence level across all the models. In addition, the household income factor appears to have strong impact on weekly variability. As Figure 4-7 (left panel) shows, experiment group participants who live in a low or medium annual income household (below $75,000) exhibit declined weekly variability from Wave I to Wave III, while their counterparts in the control group show an increase of a similar magnitude. The interaction between the income and distance factors is also strong, which means the effect from household income on weekly variability is different across experiment control groups. The 95% confidence interval using Tukey's method shows that the difference mainly occurs among the medium income participants (Table 4-15). Results suggest 95% confidence that, during Wave III, the medium income experiment group participants have a weekly average variability lower than that of control group by 2.46 and 13.78, with a mean estimate at 8.12. On the other hand, the income effect on weekly variability is not obviously different across control and experiment groups for low income or high income participants.

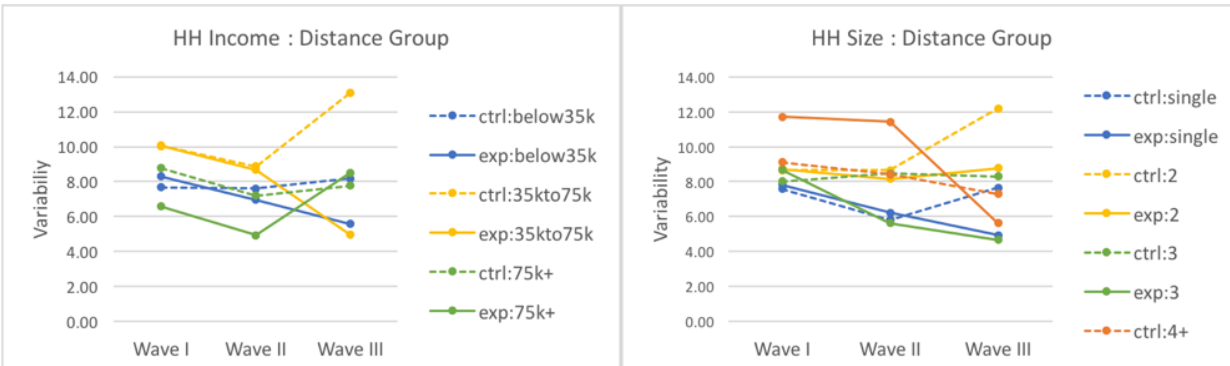


Figure 4-7 Variability Change by Household Factors

In addition to the influence of the household income, the main effect of household size on weekly variability is statistically significant at a 99% level, which indicates that during Wave III, the variance explained by household size cannot be ignored even after the variance associated with the distance factor is controlled for (Table 4-16). However, the interaction between household size and distance factor is not significant. This suggests that the impact of household size on weekly variability is not very different across experiment and control groups. Figure 4-7 (right panel) shows that from Wave I to Wave III, experiment group participants who live alone or in a family larger than 3 persons experience a drop in their travel variability at week level. On the other hand, experiment group participants who live in a two-person household do not show significant change from Wave I to Wave III; control participants with 2 household members make a significant increase in weekly travel variability. By applying Tukey's method, we can see that a household size of 2 contributes to the significance of the main effect, as the three pairs involving size of 2 are statistically significant while other pairs all cover zero.

Table 4-15 Pair-wise Confidence Interval for Variability Difference by Household Factors

Contrasts	Difference (Variability)	95% C.I. Lower	95% C.I. Upper	p-value
Household Income : Distance (interaction)				
low:exp - low:ctrl	-2.604	-6.121	0.912	0.258
med:exp - med:ctrl	-8.121	-13.780	-2.462	0.001***
high:exp - high:ctrl	0.738	-5.670	7.146	0.999
Household Size (main effect)				
1 – 4+	-0.924	-4.223	2.374	0.878
3 – 4+	0.129	-4.002	4.259	1.000
2 – 4+	3.914	0.309	7.520	0.029**
3 - 1	1.053	-2.521	4.627	0.861
2 - 1	4.838	1.887	7.790	< 0.001***
2 - 3	3.786	-0.074	7.645	0.056*

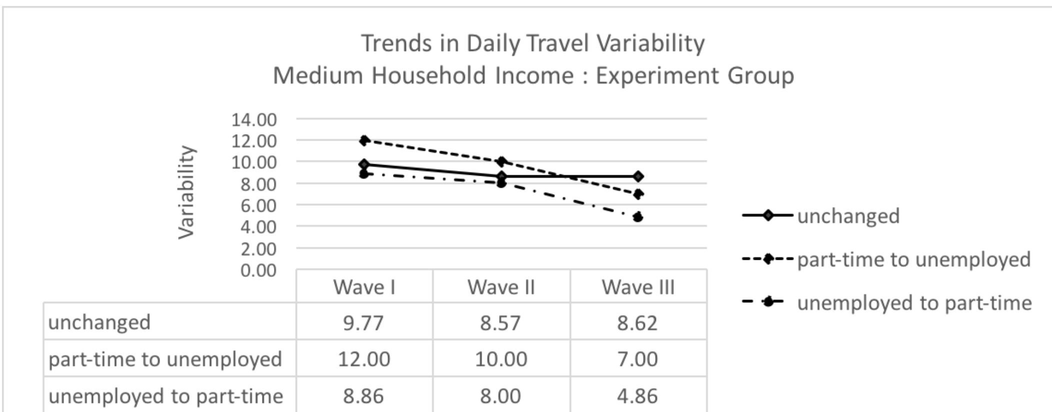


Figure 4-8 Change of Weekly Variability by Change of Demographical Status

Since it is a relatively small dataset, it is highly possible that the change in variability across waves could be driven by changes for a few participants. Therefore, we examined patterns for those participants who live in a medium-income family with a location within ½ mile to Expo stations since this group appears to be the major contributor to the variability reduction of experiment group. There are 8 participants in this particular group. From Wave I to Wave II, none of the 8 participants have changed social-demographical status. From Wave II to Wave III, two of them switched from part-time employed to unemployed and one of them changed from unemployed to part-time employed. Meanwhile, other individual level characteristics have not been changed except for age. As shown in Figure 4-x, the reduction in variability mainly occurs for the three participants whose employment status changed. We would expect to see an opposite pattern for the two opposite directions of change (i.e. from unemployed to part-time employed vs. from part-time employed to unemployed), but they are all connected with a reduction in variability. This suggests that employment status and living environment may interact together and influence the travel variability. It is important to point out that these averages are based on a total of 8 participants. Hence, further validation via a larger dataset would be necessary to establish the connection between employment status and travel variability level.

Table 4-16 Two-way Anova Table

	Wave I				Wave II				Wave III			
Source	**SS**	**d.f.**	**F val.**	**p-value**	**SS**	**d.f.**	**F val.**	**p-value**	**SS**	**d.f.**	**F val.**	**p-value**
gender	5.08	1	0.353	0.555	2.14	1	0.134	0.716	5.53	1	0.418	0.521
expo	0.53	1	0.037	0.848	4.81	1	0.301	0.586	133.84	1	10.112	0.003***
gender:expo	5.12	1	0.356	0.553	20.06	1	1.255	0.268	7.85	1	0.593	0.445
Residuals	732.95	51	-	-	815.25	51	-	-	675.04	51	-	-
age group	35.61	3	0.818	0.490	27.1	3	0.595	0.622	71.96	3	1.851	0.151
expo	0.87	1	0.060	0.808	2.44	1	0.161	0.690	129.98	1	10.030	0.003***
age group:expo	25.92	3	0.596	0.621	96.69	3	2.123	0.110	7.41	3	0.191	0.902
Residuals	681.62	47	-	-	713.66	47	-	-	609.06	47	-	-
vehicle	82.4	2	3.151	0.052*	8.96	2	0.290	0.749	24.68	2	0.914	0.408
expo	6.58	1	0.503	0.482	2.35	1	0.152	0.698	113.31	1	8.391	0.006***
vehicle:expo	20.03	2	0.766	0.470	71.51	2	2.314	0.110	2.06	2	0.076	0.927
Residuals	640.71	49	-	-	756.98	49	-	-	661.69	49	-	-
schedule	11.69	2	0.399	0.673	55.79	2	1.778	0.180	14.74	2	0.539	0.587
expo	0.39	1	0.027	0.871	5.85	1	0.373	0.544	131.98	1	9.647	0.003***
schedule:expo	13.26	2	0.452	0.639	13.06	2	0.416	0.662	3.31	2	0.121	0.886
Residuals	718.19	49	-	-	768.61	49	-	-	670.38	49	-	-
household income	43.14	2	1.542	0.224	40.88	2	1.268	0.291	24.86	2	1.109	0.338
expo	0.02	1	0.001	0.971	9.41	1	0.584	0.449	143.79	1	12.832	0.001***
income:expo	14.41	2	0.515	0.601	6.45	2	0.200	0.819	114.51	2	5.110	0.010***
Residuals	685.59	49	-	-	790.12	49	-	-	549.06	49	-	-
household size	38.11	3	0.859	0.469	86.43	3	1.893	0.144	180.85	3	5.639	0.002***
expo	5.49	1	0.371	0.545	0.16	1	0.011	0.918	103.72	1	9.703	0.003***
size:expo	10.04	3	0.226	0.878	35.81	3	0.785	0.509	5.15	3	0.161	0.922
Residuals	694.99	47	-	-	715.21	47	-	-	502.44	47	-	-

*Significance level: 0.01 ***; 0.05 **; 0.1 **

Conclusion and Discussion

This chapter examines the significance of intrapersonal variability in trip chaining behavior from the perspective of tour complexity type and mode choice. The study uses a tour-level dataset developed from multiday individual GPS trajectories. First, the analysis confirms with the previous research that intrapersonal variability accounts for a major portion of the total observed variation in trip chaining behavior. More specifically, the result shows the existence of the day-to-day within-person variability in tour type sequence, by jointly considering tour complexity and travel mode. In short, the findings of this chapter have confirmed and extended the previous research. One of the purposes of this analysis component is to demonstrate the application of TCI in complex trip chaining analysis. In the method utilized here, a home-based tour is classified based on tour complexity type and primary mode. Then, a string is formed by the tours occurred during the same day by the same person. Travel patterns on any two days are compared using the string for each and quantified via SAM. The result validates this analysis framework and the effectiveness of TCI in characterizing trip chaining behavior. Further, the comparison in variability before and after the opening of a rail line across treatment groups shows the variability keeps declining in the group closer to the new transit facility. No statistically significant association has been found between the declining and socio-demographic characteristics so the factor of distance to Expo stations seems to be the best explanation of the observed change.

Limitations and Future Work

Although the sample of this study is small, the detailed longitudinal travel and activity data provide a platform for developing and demonstrating the TCI framework for assessing tour complexity.

Despite this contribution, some limitations exist. This study examines the variability of trip chaining behavior joint with mode choice, but other dimensions of variability in urban travel behavior exist that could be examined using the TCI approach in the future, including variation in day-to-day departure and return time, daily travel duration and route choice. Meanwhile, mode choice and intrapersonal variability in trip chaining may be correlated. For instance, it is possible that people with regular travel demand (e.g. commute) would be more likely to choose public transit; people who have more impromptu travel needs would be more likely to rely on private vehicles. An exploration on this issue would be helpful for a better understanding in individual daily travel variability.

Chapter 5 Recognition of Representative Travel-Activity Patterns Based on Tour Complexity

Introduction

Chapter 3 utilized the Tour Complexity Index (TCI) to examine tour complexity at tour level and Chapter 4 utilized TCI to examine intrapersonal day-to-day variability in tour complexity within a week. This chapter provides new perspectives on tour complexity by utilizing longitudinal Expo Study GPS data to recognize groupings of similar time-activity patterns for out-of-home activity based on TCI classifications and social-demographical factors. Given the condition that no pre-determined labels are assigned to travelers, the task being addressed in this chapter is essentially an unsupervised clustering problem. The analysis utilizes a generalized analysis of variance (ANOVA) method known as discrepancy analysis to embed social-demographical factors into a clustering process. Results provide insights into how social-demographical factors influence weekly travel patterns and how factors interact with each other, and provides a new analytical tool for identifying representative time-activity patterns.

As discussed in Chapter 4, travel behavior analysis based only on a single-day travel survey per individual/household could be biased due to daily, weekly, to monthly variability (Raux et al., 2016) because many activities, particularly maintenance activities, do not occur on a daily basis. Hence, travel survey data collection for longer periods, such as a week, likely provide more representative data. This chapter continues to take the advantage of the longitudinal feature of Expo Study GPS tour dataset and examines week-long patterns for both weekdays and weekends. Most of the literature on representative pattern recognition focuses on features of daily activities and few have added features including location and other travel attributes (Joh et al, 2002; Kim, 2014; Saneinejad & Roorda, 2009; Wilson, 2001, 2008). Admittedly, large-scale passively recorded trajectory data (GPS data, mobile phone data, vehicle data and etc.) tend to have the limitation that they lack of activity information such as trip purpose and household trip interdependency information (who is traveling together, how many per trip, etc.). These data, however, are much easier and cheaper to obtain compared to traditional travel survey data. In addition, such data information often contain more than one-day of travel information which can be beneficial to improve transportation research. Hence, development of methods that can analyze GPS travel information without information regarding activity context and trip details will support the development and improvement of activity-based modeling in an era of big data.

Literature Review

This section includes a brief discussion of travel-activity pattern recognition and a review of previous relevant studies. Classification of travel-activity patterns has been a major branch of transportation research. It sets up the foundation for other theoretical and empirical research. It also provides applicable guidelines for model development, especially regarding how to integrate social-demographical factors into modeling process. However, it is a challenging task to simultaneously capture the full complexity of travel-activity patterns from all dimensions. A review of the related literature reveals that two common approaches have been applied in travel pattern studies. The first approach is to analyze the linkage between traveler characteristics and travel behavior via a series of econometric and discrete choice models. Travel-activity behavior is decomposed into a set of cross-sectional factors, including activity type and frequency, trip rate, trip chaining characteristics, travel mode, destination, and travel and activity durations. (Hanson & Huff, 1986; Jun Ma & Goulias, 1997). Most of the current operational activity-based travel demand modeling system are designed in this way. The second general approach is to study activity patterns as a single entity. Early activity-based travel behavior research included some exploratory studies from this perspective (Recker et al., 1985). The research often starts by discussing a multi-dimensional representation of a series of travel-activity events such as Hägerstrand's trajectory approach (Hagerstrand, 1970). Dimension reduction and clustering methods are used to group observed patterns into typical patterns. In a few studies, subsequent analysis has been applied to test the association between exploratory variables (mostly social demographical variables) and identified clusters. This post-clustering analysis methods include descriptive analysis, ANOVA, and contingency tables (Recker et al., 1985; Saneinejad & Roorda, 2009; Wilson, 2001).

One of the fundamental differences between these two approaches is whether to associate the covariates with decomposed travel behavior dimensions separately, or to associate the external factors with the pattern being treated as a whole. The current analysis adopts the second approach in that it regards the travel-activity pattern as a feasible unit for analysis. The current approach is unique, however in that (1) it avoids the dimension reduction step by applying discrepancy analysis and (2) it is the first study, to the author's knowledge, to analyze travel patterns by combining trip chaining and Sequential Alignment Methods (SAM). In addition, the trip chaining data used here do not include traveler input on activity types and trip details so that it reduces overall data collection burden and increases overall data reliability.

Sequential Alignment Method in Travel-Activity Analysis

Sequential Alignment Methods (SAM) is an emerging technique in pattern study based on the second approach described above. A summary table is provided below for selected studies using SAM to identify travel-activity patterns (Table 5-1). First introduced into the travel behavior field by Wilson (1998), SAM was originally applied in the field of biology to study DNA and protein patterns. The researchers in the travel behavior field have found that events recorded in the travel surveys or general activity diaries can be represented as a sequence (or "trajectory"), which is coded as a list of numeric or alphabetic strings. The code can be defined with a large degree of flexibility to represent travel-activity events or status within a specific time window. For instance, it could be used to identify patterns in activity contexts (i.e. eating, working, school and etc.), travel modes, locations, or even person present during an activity. Theoretically, the number of possible sets of information that can be coded into a sequence is unlimited.

The degree of flexibility brings complexity (Figure 5-1). Dimensionality is one of the major concerns of the SAM approach. The sequence can be maintained as uni-dimensional by coding the events and status jointly; or, it can be expanded to be multi-dimensional by creating a series of sequences with aligned time windows. Many controversial issues are raised when characterizing and analyzing travel behavior via multi-dimensional sequences, including weighting, operational boundaries, and computational cost (Joh et al., 2001). This may explain the fact that most of the reviewed studies have chosen to use the uni-dimensional SAM approach (Kim, 2014; Wilson, 2008). Nevertheless, the uni-dimensional approach fails to fully capture the interdependency between multiple attributes. Another important concern about analyzing sequential human activity patterns is whether to include activity duration. That is, whether the unit piece of a string should be equivalent as one event, or should it be equivalent as a time window. The former results in mismatching in length between observations when two observations have a different number of events; the latter method results in a uniform length among observations which is determined by the size of time window (e.g., 5 minutes, half hour, one hour) and total duration analyzed (e.g., one day, one week, one month). Both approaches have been applied in the reviewed studies (Table 5-1).

Uni-dimension

(1) One attribute, short format

A	D	B	C

(2) Three attributes, short format

Aa1	Da1	Bc0	Cb0

(3) Three attributes, long format

Aa1	Da1	Bc0	Bc0	Bc0	Cb0	Cb0	Cb0

Multi-dimension

activity	A	D	B	B	B	C	C	C
location	a	a	c	c	c	b	b	b
other person	1	1	0	0	0	0	0	0

Activity Attributes and Codes

Code	Description
Activity Type	
A	domestic work
B	education
C	working
D	eating
Location	
a	Zone A
b	Zone B
c	Zone C
Other person present	
0	No
1	Yes

Figure 5-1 Illustration of Travel-activity Sequence

As discussed above the representation of travel-activity is the first step of SAM. The most critical step in SAM is the second step which includes the measurement of similarity and distance/dissimilarity between observations. The measurement is important because it directly affects the clustering results. The most common measurement approach is developed based on a dynamic programming solution known as Levenshtein distance (Kruskal, 1983; Needleman & Wunsch, 1970). The concept is to seek an optimum number of operations that can equalize two strings with the optimality determined by a function of operations and their corresponding costs. There are three types of operations: insert, delete and substitute. Since insert and delete always come as a pair, they are referred to as *indel* in most studies. So essentially, there are only two operations: *indel* and substitute. The debate lies in the cost of the operations. If different costs are applied to the two operations additional questions include: (1) how to define and justify the difference in costs and (2) whether to apply the same cost scheme to all the events and positions. Most of the reviewed studies have used the default mismatch cost approach, which treats any operation in any position between any two levels with the same cost. Only one relevant study has explored different costs (Wilson, 2001). In this study, the substitution between employment and domestic work costs less than that between employment and eating. The first *indel* position costs less than subsequence *indel*. However, the author admitted that this cost assumption is purely based on the author's personal impression about how close two types of activities are related.

Table 5-1 Travel-activity Analysis using Sequential Alignment Method

Study	Sequence Construction			Sample size	Subsequent Analysis
	Attribute(s)	**Length (interval)**	**Equal length**		
Wilson (1998)	activity	1 day (5 min.)	Both	18	-
Wilson (2001)	activity, location, person presented	1 day (30 min.)	No	248	tree-structured clustering; descriptive analysis; ANOVA
Wilson (2008)	activity, location	1 day (30 min.)	No	368	-
Shoval & Isaacson (2007)	location	less than 1 day (1 min.)	No	40	contingency table
Saneinejad & Roorda (2009)	activity, location	5 weekdays (15 min.)	Yes	282	descriptive analysis
Kim (2014)	activity	1 day (5 min.)	Yes	1000	discrepancy analysis; tree-structured clustering

Wilson performed a series of analysis using SAM on activity pattern clustering. The first study in 1998 is an exploratory study that uses an extremely simple dataset to demonstrate the application of SAM on human behavior. The dataset is collected from residents in Reading, U.K. in 1974. This study applies a single-person multiday scenario and a four-person multiday scenario. For each scenario, an equal length approach and a non-equal length approach are attempted. The study successfully extracted representative patterns from the sample and thus concluded that SAM has great potentiality in behavioral research (Wilson, 1998). The author, in 2001, applied a revised methodology on an activity diary set collected from 248 Canadian women in 1992. This study extends the activity sequencing approach by adding two new factors: location and other people present during the activity. The computation of dissimilarity is modified as well (previously described). Moreover, subsequent analysis is performed to validate the effectiveness of clustering in discriminant population groups (Wilson, 2001). Later, Wilson experiments with the idea of amending the activity sequence by adding geographical attributes so that the resulting representative patterns can be visualized as a Hägerstrand trajectory. The proposed methodology is applied to a dataset containing 368 activity diaries collected from 53 survey respondents. The sequence is constructed using a two-digit code and a 30-minute interval. The first digit is for activity type while the second digit is a zone id representing the geographical location of the activity. Each individual sequence starts at the first activity after waking and ends at sleeping at night so that the sequences are different in length at a daily level. SAM is also used to compute the pair-wise distance assuming the mismatch cost is defined as the sum of activity mismatch cost and location mismatch cost. The dissimilarity matrix is put into a program called ClustalTXY clustering. As the result, the 368 individual sequences are assigned into three primary clusters based on a rule to minimize the within-cluster distance. Three individual sequences with the least squared distance to other within-cluster sequences are then identified as the representative patterns. As for subsequent analysis, the author chooses not to associate socio-demographic factors with the clusters because this extra step requires more statistical inference and testing which were beyond the scope of the study (Wilson, 2008).

Saneinejad & Roorda (2009) investigate weekly routine activity schedules using SAM. A total of 282 individuals provide descriptions about their weekly routine in terms of activities and locations. Unlike Wilson's approach in which location is coded based on geographical coordinates, the location in this study is categorized as either home or non-home, an approach that can be further extended to relative distance to home. Again, the activity sequence is uni-dimensional with each piece containing a two-digit code for the status at a 15-minute time interval. Only *indel* applied for alignment and substitution is excluded, which results in a dissimilarity score that is inflated. A total of 9 clusters with respect to schedules are identified. Descriptive analysis is applied in the end to relate socio-economic characteristics to the schedule clusters. The study finds that, age, gender, family life cycle, employment status, and income level are influential to schedule choice. Note that the study is developed based on perceived routine activities and only for weekdays, which implies its results could suffer from a potential loss of accounting for unplanned activities (Saneinejad & Roorda, 2009).

Kim examines the daily activity patterns of 1,000 individuals living in the Portland metropolitan area (Kim, 2014). The traveler-reported activities are aggregated into 12 types by location type visited (i.e. home, workplace, school, and other). For each person, a sequence is coded based on a single attribute, activity type, at a 5-minute interval starting from 3:00 am to 2:59 pm of the next day. The author then uses default SAM (same cost for *indel* and substitution) and uniform penalty for activity difference to construct a distance matrix. Multivariate discrepancy analysis and regression tree methods are applied to estimate the association between the interpersonal dissimilarity and personal characteristics. Both methods explain approximately 19% of total discrepancy. The study found that that employment status, age, presence of K-12 children at home, and household size play a significant role in daily activity sequence.

In addition to regular travel-activity analysis, SAM has also been applied to activity clustering for tourists. In Shoval and Isaacson's 2008 study, the spatial-temporal trajectories of 40 tourists in a historical city in Israel are analyzed. Their activity locations are categorized at site-level in the historical city and the resolution of the time interval is 1-minute. The researchers run the analysis in the program called ClustalG developed by Wilson. A total of 9 tourist groups are identified based on the similarity of their moving trajectory. Contingency table methods are utilized in subsequent analysis to assess the association between trajectory choice and a set of tourist-related variables (Shoval & Isaacson, 2007).

Further, SAM has been demonstrated to be more effective in travel-activity pattern recognition than alternative dissimilarity measurements developed based on Euclidean distance and signal processing methods (Joh et al., 2001). SAM has three main advantages. First, it maintains the sequential nature of the events and the distance measured is order sensitive. Second, it keeps the interdependency across event and travel attributes. Third, it directly compares strings while alternative methods require feature extractions and lead to indirect comparisons and loss of information.

The studies reviewed have focused on activity patterns and provide insights which could benefit activity-based travel demand analysis. The difference between analyzing activity patterns compared to travel patterns is that a major portion of activities are home-based while travel more explicitly relates to out-of-home activities. This chapter examines weekly travel-activity patterns from a standpoint of out-of-home trip chaining behavior. It has important implications for future travel demand research and the development of activity-based modeling, in that it directly focuses on out-of-home travel-activities. It provides a new perspective that, to author's knowledge, has not been previously explored. It takes important steps to associate travel-activity patterns with population segments in order to provide operational guidance for model development. The reviewed literature examines the relationship between activity patterns and socio-demographic factors in a posterior way, which lacks generality from the perspective of modeling. The current study embeds the socio-demographic factors into the clustering process via a tree-structured regression method based on discrepancy analysis (a generalized method following the principle of ANOVA). Hence, it demonstrates a more feasible implication for model developers.

Research Goals and Questions

In brief, the research goals being addressed in this chapter are:

- Identify representative categories of travel-activity patterns based on weekly travel patterns and TCI measures of trip chaining complexity
- Investigate the personal and family characteristics that influence patterns
- Experiment with the discrepancy analysis and regression tree methods

Data and Methodology

The Wave I portion of the Expo tour dataset is used in this chapter. The details regarding Expo tour dataset, including Expo Line Study design, traveler's socio-demographic characteristics, and general travel statistics are provided in Chapter 2. The Wave I data analyzed in this chapter were collected in 2011 and includes the one-week GPS travel trajectories for 55 individuals who live near the newly constructed Expo light rail line in Los Angeles, California. A total of 55 individual weeks (385 days) are sampled, containing 1,478 valid trips and 409 home-based tours. The socio-demographic variables used to associated with travel pattern are summarized in Table 5-2 below.

Table 5-2 List of Independent Variables

Variable	Description	Levels
gender	Gender	male, female
age	Age group	below 35, 35 to 55, 55 to 65, 65 and above
caracc	Vehicle accessibility	not licensed/no car, own and share car, own car and does not need to share
time	Schedule flexibility	free, part-time, fulltime
income	Household income level	low (<=35k), medium, high (>=75k)
hhsize	Household size	single, two, three, four and above
lifecycle[1]	Family life stage	A, B, C, D, E
dist2expo	Distance from home to Expo stations	within ½ mile; ½ - 1 mile; greater than 1 mile

1. Level of lifecycle: A. young adult (below 35) family with no child, B. family with children under 12, C. family with children above 12, D. older adult (above 35) family with no child, and E. retired person

Representation of Weekly Trip Chaining Pattern

First, the trip chaining behavior continues to be characterized using the Tour Complexity Index (TCI). The first dimension of TCI, Segment Index, is not bounded with an upper limit theoretically. As a result, it leads to numerous combinations of possible TCI values. Therefore, for simplicity, the TCI values are simplified to 18 types aggregated by the Segment Index. The rule for aggregation is provided in Chapter 4 including the code dictionary used (Table 4-5). In short, each home-based tour is assigned a one-digit code based on TCI type. TCI codes are determined based on four tour-level attributes: 1) number of trip legs, 2) whether there is spatial cluster between consecutive stops, 3) whether the route used to chain the stops follows the most direct or shortest path, and 4) whether the stops have different land use settings. For example, a short one-stop tour is coded as “B” and a 5-stop tour of which the destinations are not clustered, chained in an efficient route, have destinations with two types of land use, is coded as “N”.

SAM is used here to compute pairwise dissimilarity/distance between any two trip chaining sequences. The details are covered in Chapter 4. The major difference in this chapter is that before applying SAM, data interpolation is performed for each individual day to accommodate the extended analysis goal as weekly pattern recognition. Interpolation is a common data cleaning technique for time-series data to equalize the length of each observation. Assuming all the sequences have a universal length (L = 24 in our case, representing one level for each hour in a 24-hour day), each tour is mapped to its corresponding time point based on the hour in which it occurs and time points in which no tour occurred are filled using a "gap" code ("S"). For example, if a traveler only performs a type "B" tour during 11:00 to 14:00 of the day, then the sequence of that day is coded as: "S, S, S, S, S, S, S, S, S, S, S, B, B, B, B, S, S, S, S, S, S, S, S, S". In this way, all daily trip chaining sequences are represented by strings with an equal length of 24. The purpose of this transformation is to avoid time distortion in the sequential alignment computation. An important goal of the weekly pattern clustering among different individuals is to capture differences in timing and duration of the trip chaining behavior and so it is critical to equalize the string length to enable such comparison.

After interpolation, each day and week can be viewed as a time-series sequence with a length of 24 coded based on TCI types undertaken at each time interval. Three plot types are developed to visualize the trip chaining sequence as shown in Figure 5-2, Figure 5-3 and Figure 5-4. Figure 5-2 is the "matrix plot" designed to visualize a single week's trip chaining behavior for one participant. The X-axis represents hour of the day from 0:00 to 23:00 while the Y-axis represents day of the week with Monday at bottom and Sunday at top. The color code follows TCI code with the darker color representing more complex tour types and lighter representing simpler tour types. The advantage of matrix plot is that the timing, duration, and type of out-of-home travel behavior of a single traveler can be compared across days within the week. It is good for intrapersonal and interpersonal daily comparison. However, to compare multiple individuals' weekly patterns, the matrix plots need to be either vertically or horizontally stacked. As a result, either the days or hours are not aligned. It then brings inconvenience for visual comparison.

Hence, two additional plots are developed. Both Figure 5-3, “time-series plot”, and Figure 5-4, “distribution plot”, use one row to represent each individual week by reshaping the matrix plot to a wide-format. That is, to flatten the matrix from seven rows into a single row. The difference between the “time-series” and “distribution plot” is the order of the color-coded cells. As the name suggests, the cells are ordered by time in the “time-series plot” with Monday 0:00 to the left and Sunday 23:00 to the right. As there is rarely out-of-home activity during late night or early morning hours, each row looks like a “dashed line”. In addition, all the individual sequences are aligned vertically at time points. The advantage of the “time-series” plot is the convenience in comparing timing across individuals during periods of travel. For the “distribution plot”, the cells are ordered by the complexity of TCI type in a descending order from left to right. Although it breaks the alignment in time points, the “distribution plot” is good for visualizing the total out-home activity duration and overall complexity of the trip chaining behavior. For instance, some individuals have a long colored sequence filled with light color, indicating the person tends to spend a long time out home but has simple trip chaining behavior. Other individuals may have fewer colored cells which are filled with dark color, indicating they spend less time out home but chained trips in a more complex way. In the following analysis of pattern clustering, we mainly use the “time-series plot” and “distribution plot” since the emphasize is on interpersonal comparison.

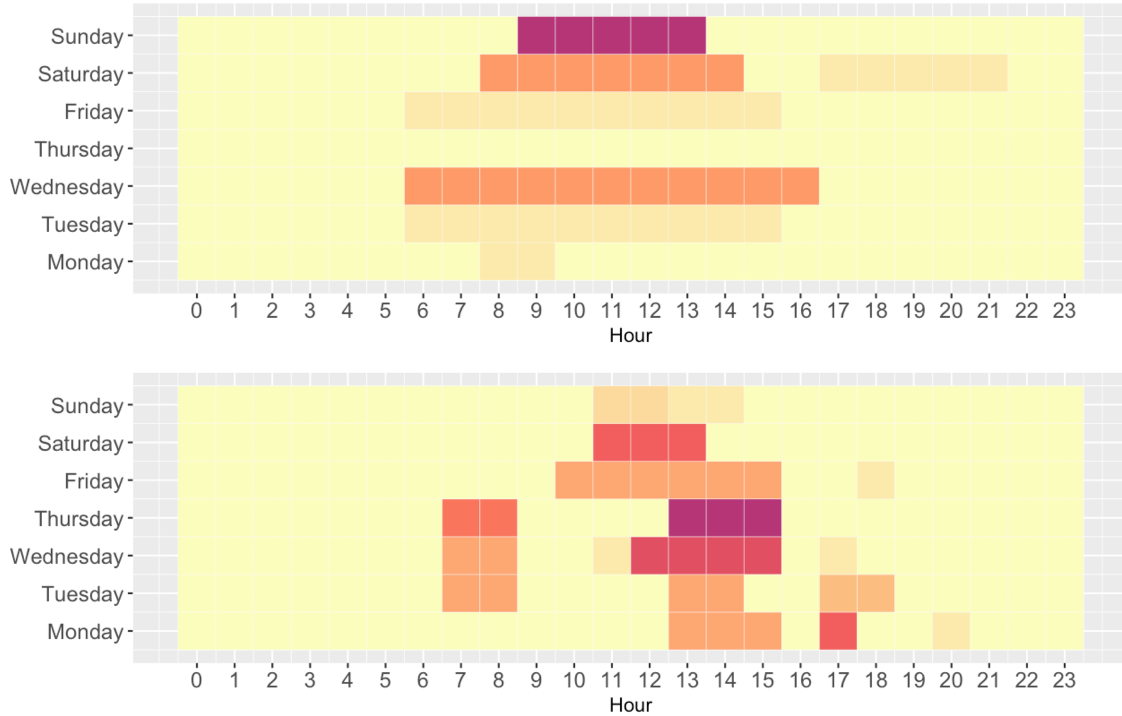


Figure 5-2 Weekly TCI Pattern Plots: Matrix Plot

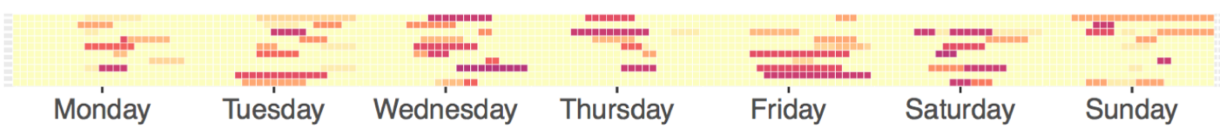


Figure 5-3 Weekly TCI Pattern Plots: Time-series Plot, n=10

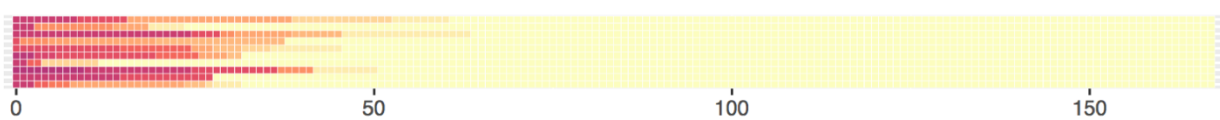


Figure 5-4 Weekly TCI Pattern Plots: Distribution Plot, n=10

Measurement of Dissimilarity

With the representation of weekly trip chaining behavior defined, the next step is to determine how the deviation between each pair of week-long sequences should be measured. A "sequence" is a special case of a "vector" where the order has a temporal meaning. There are many methods to quantify the distance/dissimilarity between two vectors. In the case of SAM, Levenshtein distance, defined as the cost to align two sequences, is particularly suitable for our purpose.

The dissimilarity is first measured at a daily basis as it is not meaningful to align the sequence across days of the week. Levenshtein distance is embedded into SAM to measure the distance between a pair of days (i.e., person A's Monday vs. person B's Monday, person A's Tuesday vs. person B's Tuesday, and so on). As discussed in Chapter 4, a lot of details in SAM remain open for debate when being applied to measure deviation in human behavior. For instance, how to define costs for different equalization operations, how to weight the difference between two tour types and between traveling vs. staying at home, and whether the transitional gap during two tours should be treated differently as terminal gaps with no travel following. The more complicated the methods are, it usually requires a larger and richer dataset to validate the chosen methodology. The dataset analyzed here is relatively small and homogeneous when comparing to typical datasets used for for travel demand model development. Hence, this study follows the most basic settings and focuses more on method demonstration.

The distance is defined as the number of basic operations required to equalize the two sequences:

$$d_{k,ij} = \#\ of\ insertion, deletion, and\ substitution$$

where k refers to the k^th^ day of the week with $k \in \{Monday,\ Tuesday, \ldots, Sunday\}$; i, j is the participant index.

The dissimilarity between i^th^ person and j^th^ person's weekly sequences then equals to the sum of daily dissimilarities through the week:

$$d_{ij} = \sum_{k=1}^{7} d_{k,ij}$$

With the computation being applied to all pairs of participants, a 55 by 55 distance matrix is generated and used as the input for subsequent analysis.

Discrepancy Analysis based on Dissimilarity Matrix

This chapter demonstrates an analysis framework to study the association between interpersonal trip chaining behavior and categorical personal characteristics. Introduced by Studer et al. (2010), discrepancy analysis is developed based on an ANOVA method where the word "discrepancy" is a generalization for "variance" in the ANOVA setting. The method is designed to evaluate the relationship between the deviation across analysis units and their individual attributes. The deviation is expressed as a dissimilarity matrix while the personal attributes are categorical. Originally, this method has been experimented with in a few cases to assess other social science research topics, such as ecosystems and individual life trajectories (Studer et al., 2010). More recently, it has been introduced to travel behavior literature through studies such as the activity diary study by Kim (2014).

As described in Chapter 4, ANOVA is a statistical method to analyze the partition of the total observed variance by splitting it into different sources and assessing the percentage of sources that explained the total variance. The variance is usually noted as "sum of squares", which is defined as:

$$SS = \sum_{i=1}^{n} (y_i - \bar{y})^2 \tag{5-1}$$

It can be proved that SS also can be expressed via pairwise Euclidean distance:

$$SS = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=i+1}^{n} d_{ij}^2 \tag{5-2}$$

where d_{ij} denotes the Euclidean distance (Anderson, 2001).

Anderson (2001) and Studer et al. (2010) indicate the concept of the sum of squares can be applied or generalized to other dissimilarity measures by replacing the squared Euclidean distance d_{ij}^2 with any non-Euclidean measurement at any degree of positive exponent.

$$SS = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=i+1}^{n} d_{ij}^{*\nu} \tag{5-3}$$

The above generalization of SS computing has two implications: 1) the variance can be obtained additively from element-wise distance, without computing the centroid, and 2) the distance between objects is not necessarily measured as Euclidean. The importance lies in that it shows variance-based analysis methods can be applied to more complex objects in which the class center and distance cannot be well defined. These are also the two keys of the discrepancy analysis.

Further, considering the one-way ANOVA, its purpose is to test whether a population grouping factor, as a source of the variance, is able to explain a significant amount of variance in the sample. One-way ANOVA is built on the equation that the total sum of squares (SST) equals to the within group sum of squares (SSW) plus the residual sum of squares (SSB), where SST and SSW can bed computed using the following equation:

$$SST = SSB + SSW \tag{5-4}$$

where

$$SST = \sum_{i=1}^{N} (y_i - \bar{y})^2 \tag{5-5}$$

$$SSW = \sum_{k=1}^{K} \sum_{i=1, i \in level\ k}^{n_k} \left(y_{i,k} - \bar{y}_k\right)^2 \tag{5-6}$$

$$SSB = SST - SSW \tag{5-7}$$

and N = total number of observations; n_k = number of sample in k^{th} level; K = number of levels under grouping factor; $\bar{y}$ = overall sample mean; $\bar{y}_k$= group mean.

It has been found that the equation holds with non-Euclidean distance (Studer et al., 2010). Combining the two generalization principles discussed above, SST* and SSB* can be computed as:

$$SST^{*} = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=i+1}^{N} d_{ij} \tag{5-8}$$

$$SSW^{*} = \sum_{k=1}^{K}\frac{1}{n_k}\sum_{i=1, i \in level\ k}^{n_k}\sum_{j=i+1, j \in level\ k}^{n_k} d_{ij} \tag{5-9}$$

where d_{ij} is a non-Euclidean distance between object i and j.

Therefore, the univariate discrepancy analysis is able to analyze the association between a grouping factor and the dissimilarity between objects which is measured in non-Euclidean distance. By knowing SSW* and SSB*, we can assess the proportion of discrepancy that can be explained by the population segmentation factor. Under the setting of ANOVA, R^2 and F tests are applied to assess the statistical significance of the part of the discrepancy explained by the segmentation. For discrepancy analysis, we adopt the concept of pseudo-R^2 and pseudo-F tests presented in Studer et al. (2010) and shown in (5-10) and (5-11). Permutation test is used to compute p-values with 1,000 iterations to reach the significance level of 5% (Anderson, 2001; Moore et al., 2003; Studer et al., 2010).

$$Pseudo - R^2 = \frac{SSB}{SST} \tag{5-10}$$

$$Pseudo - F = \frac{SSB/(K-1)}{SSW/(N-K)} \tag{5-11}$$

Tree-Structured Analysis

The univariate discrepancy analysis can answer questions such as which factors are strongly associated with the interpersonal dissimilarity. However, it is hard to tell how each level of the factors affects the weekly trip chaining patterns. More specifically, it remains unclear how the clustering patterns are associated with personal characteristics. The regression tree method based on discrepancy analysis is introduced to address this question. Decision trees work in a top-down order described as follows. First, all sequence objects are assigned to one node (root). Then, the algorithm recursively splits objects in each node into sub-groups based on their attributes. The split is determined in a way that the resulting child nodes are as far from other nodes as possible. The procedure is repeated at every new child node until certain preset stopping criteria are met.

The regression-tree is computed using R package "TraMineR" (Studer et al., 2011; Studer et al., 2010). In the tree-structured analysis algorithm designed by Studer et al. (2010), the node-splitting criterion is based on the pseudo-R^2 derived from the univariate discrepancy analysis. At each iteration, all possible grouping factors and the combination of levels within the factors are searched through for the node population. For example, if a factor has three levels (A, B, and C), the combinations of {A} and {B, C}, {A, B} and {C}, and {A, C} and {B}, are computed. The split that achieves the highest pseudo-R^2 value is selected. The stopping criteria applied to this study is 5 as the minimum number of objects and 0.05 as the pseudo-F significance level. The threshold of at least 5 objects at any child node is chosen to reach a balance such that the tree can grow as deep as possible while all the final clusters have a sufficient number of samples ($n>= 5$) to perform meaningful statistical analysis. In addition, the significance level of the F value based on the univariate discrepancy analysis is used as a stopping criteria. Any node stops growing once all the split scenarios result in an insignificant p value for the F score or result in child nodes in which the number of objects falls below a threshold. For large datasets, a more common stopping criteria is the maximum tree depths (the number of parent nodes between a leaf to the root) set to avoid overfitting.

Results

Descriptive Statistics

It is helpful to first examine the overall weekly travel-activity patterns and trip chaining complexity patterns before looking for representative patterns. Three items will be described here: 1) the timing and frequency of daily travel-activity, 2) the allocation of out-of-home time on different tour complexity types, and 3) the overall and within-population-group dissimilarity in weekly trip chaining sequencing.

Timing and frequency of daily travel-activity

As presented in Figure 5-5, we can observe a core active hour for the study sample between 10:00 to 16:00, a period when most of the out-of-home activities are expected to be conducted through the week. It begins with the departure time of first tour of the day and ends with the return hour of the last tour of the day. Note that the in-home time period between two tours is included so that the total duration of the active hour is longer than the actual time spent on travel and activities. The active hours on weekdays are not very different from that on weekends. It is probably due to the large percentage of unemployed or part-time employed people in the sample. Nevertheless, the average active hour is longer on weekdays (6.22 hours) than weekends (4.75 hours). On average, each participant made 1.5 home-based tours per day on days with travel (Table 5-3). Both trip and tour rates reach their peak on Saturday while they are lowest on Sunday. However, in terms of trip/tour ratio, or legs per tour, Saturday has the lowest average number of trips chained per tour (3.22) while Sunday has the highest average number of trips chained per tour (3.66). This implies the travelers in the sample are most likely to make less chains on Saturday even though they have a highest travel demand on Saturday. Some possible explanations include: first, participants may face less time-constraints on weekends so are less motivated to arrange chained tours; second, the activities on weekends may be more likely to be dispersed spatially and thus difficult to chain; and third, it may require more coordination between involved parties for weekend activities so that it makes trip chaining difficult on weekends.

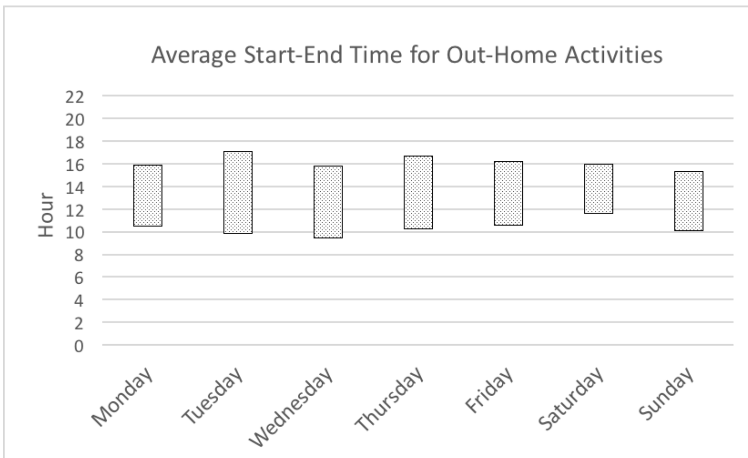


Figure 5-5 Average Start/end Hour by Day of Week, N=55

Table 5-3 Travel Statistics by Day of Week

	Mon.	Tue.	Wed.	Thur.	Fri.	Sat.	Sun.	Overall
Dep. time[a]	10:30	10:00	9:30	10:15	10:30	11:30	10:00	10:15
Ret. Time[a]	16:00	17:00	16:00	16:00	16:15	16:00	15:15	16:00
Duration (hour)	5.40	7.26	6.36	6.43	5.63	4.31	5.19	5.80
Tour rate[b]	1.51	1.54	1.45	1.48	1.47	1.78	1.32	1.51
Trip rate	5.44	5.67	5.30	5.48	5.66	5.72	4.84	5.44
Trips per tour	3.60	3.68	3.64	3.71	3.84	3.22	3.66	3.60

a. average departure time for the first tour of the day and returning time for the last tour of the day
b. days with no travel are excluded for mean calculation

Time allocation by Tour Complexity Types

Table 5-4 and Figure 5-6 present the time allocation of daily travel-activity duration by tour complexity type. That is, how much time a traveler spends on each type of tour with tours characterized by TCI. The time includes both travel time and activity duration. Table 5-4 lists the top eight TCI types undertaken by participants by duration, with the left five columns showing the TCI code and the corresponding TCI values for reference. The eight tour types with the longest duration are a mixture of single-destination, simple chained and complex chained tours. On average, each traveler in the sample spends 3.8 hours outside of home for activities per day, 3 of which are spent at the destination and 0.8 of which are spent traveling from one place to another. In addition, the activity/travel ratio shows the duration of activity for each unit time spent on traveling. The overall activity/travel ratio is 3.53. This means for every 35 minutes at a destination, the traveler is expected to spend 10 minutes traveling.

Table 5-4 Distribution of Tour Complexity Type

TCI Code	Description				Minute per capita per day			Act./travel ratio
	Seg.[a]	Clst.[a]	Div.[a]	Eff.[a]	Total	Travel time	Activity duration	
Q	complex	1	1	0	49.38	10.82	38.56	3.56
A	single	0	0	1	48.51	11.51	37.00	3.21
F	simple	0	1	1	45.11	8.74	36.38	4.16
J	simple	1	1	1	20.82	4.54	16.28	3.59
M	complex	0	1	0	16.50	2.92	13.58	4.65
R	complex	1	1	1	11.60	3.79	7.81	2.06
D	simple	0	0	1	8.80	1.64	7.16	4.37
N	complex	0	1	1	8.66	1.30	7.35	5.65
Other	-	-	-	-	20.52	5.47	15.05	2.75
Total (minute)					229.80	51.00	180.00	211.8

a. Seg. = Segment Index; Clst. = Cluster Index; Div. = Diversity Index; Eff. = Efficiency Index

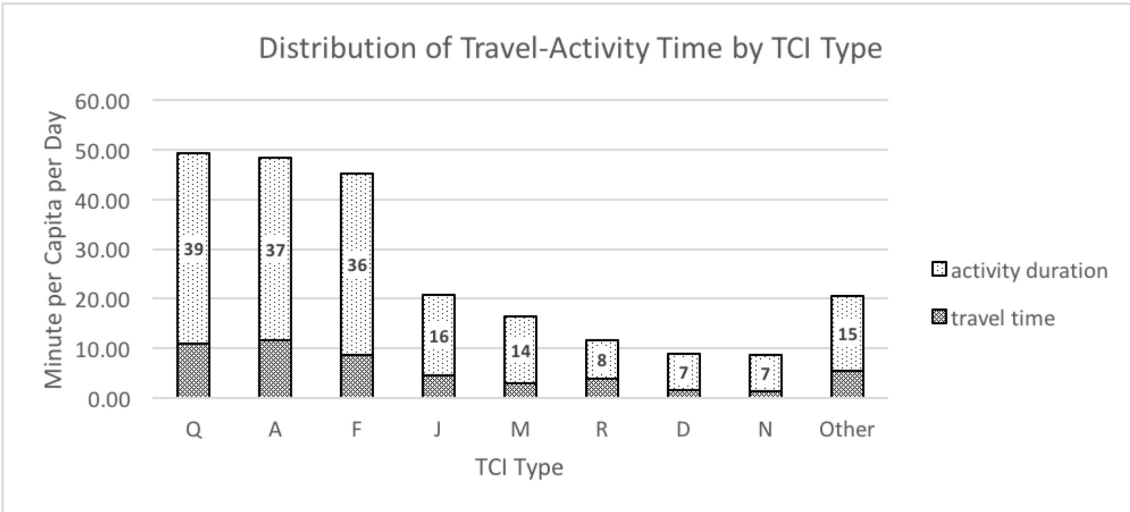


Figure 5-6 Distribution of TCI Type by Travel-activity Duration

In Figure 5-6, each vertical bar represents the total time spent on a type of tour with the TCI code below the bar. Travel time is displayed at the bottom of each column in a darker gray while activity duration is stacked on the upper portion of each column with a lighter fill. Three types of tours, Q, A, and F, dominant out-of-home time. Type Q is a tour type with 4 or more destinations in areas with different primary land use types, with trip segments linked by an inefficient route, and with at least one short trip (<= ½ mile). Type A is a single destination tour with the destination being within ½ mile from home. The distance suggests Type A tour is most likely to be taken via non-motorized travel modes. Type F is a tour type that contains 2 or 3 destinations, with an efficient route following the shortest path between destinations but no trips shorter than ½ mile, and involving destinations with different dominant land use type. Individuals in the sample on average spend about 45-50 minutes on each of these three tour types per day. Stated in another way, Type Q, A, and F tours comprise about 62% of the total out-of-home time.

Dissimilarity in weekly trip chaining sequence

We used a sequential alignment method (SAM) to examine interpersonal deviation in trip chain sequencing. As discussed in the methodology section, a 55 by 55 distance matrix is computed based on SAM, with each cell (d_{ij}) representing the degree of dissimilarity between i^{th} and j^{th} individual in their week-long travel-activity sequence characterized by TCI. The total length of the week-long sequence is 168 (24-hour times 7-day), so the lower and upper boundary of the operation cost is 0, perfectly identical for each hour slot, and 168 represents the global mismatch. The alignment cost is jointly associated with the mismatch in travel timing, in travel frequency and in choice of tour type. The overall cell mean of the distance matrix is 43.98, meaning on average 44 *indel*/substitution operations are required to equalize a pair of travel-activity sequence. This means that, compared to the extreme case of 168, we can expect above 25% of cells to be different between any two sequences.

We further disaggregated the distance matrix based on the socio-demographic variables (listed in Table 5-2). Figure 5-7 shows the within-group means across the population groups. A higher within-group dissimilarity mean implies larger discrepancy within this group. It also suggests the travel behavior and trip chaining pattern for those in this group are less predictable. It has important implications for travel demand modeling because, for subgroups that have higher degree of within-group discrepancy, more controlling factors and even longer travel survey period are needed in order to obtain an accurate behavioral approximation. In brief, females, young people under the age of 35, licensed drivers who own or share vehicle, low to medium income people, people with a large household, people who do not have children at home, and people living within the ½ mile proximity to Expo stations have a higher level of within-group deviations during Wave I of Expo Study.

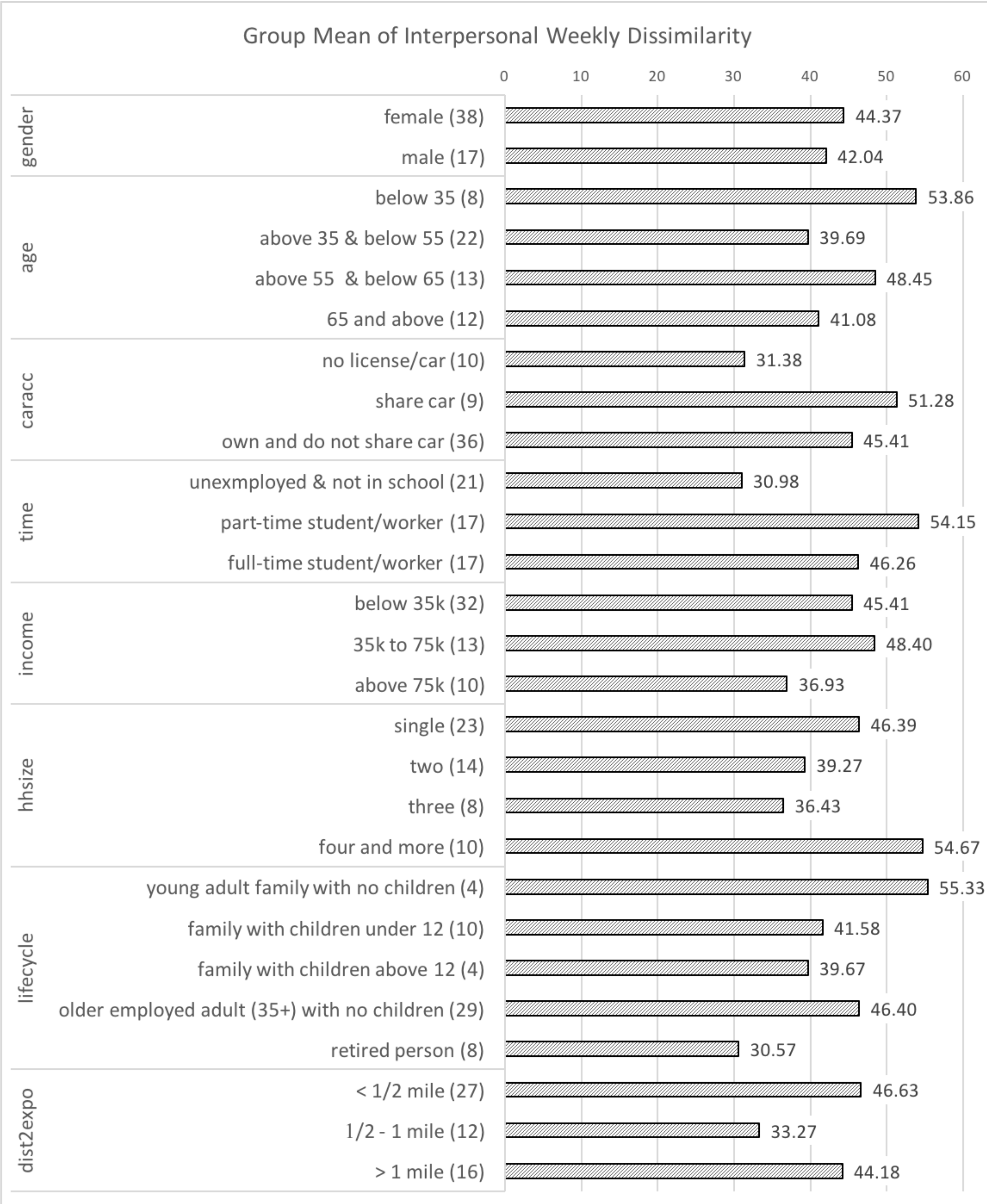


Figure 5-7 Group Mean of Weekly Interpersonal Dissimilarity

Univariate Associations

Table 5-5 summarizes the results of the discrepancy analysis of interpersonal deviation in trip chaining behavior characterized by TCI. Eight categorical socio-demographic variables are selected to perform a univariate analysis against a distance matrix with each cell representing a pairwise dissimilarity between two individuals' week-long sequences. The population grouping factor for family life cycle explains approximately 8.5% of the total discrepancy given the pseudo R^2 is 0.085. However, the variable is not statistically significant, giving a pseudo F score as 1.166 and the permutation test p-value much higher than 0.05. The most significant variable is time (schedule flexibility: unemployed, part-time employed/studying, or full-time employed/studying) which explains 6.3% of the total discrepancy. It attains a pseudo F score of 1.735 and has not been surpassed by any value in the 1,000 permutations. Another significant variable is dist2expo (distance from home to nearest Expo stations), which explains 5.6% of the total variance and is statistically significant at a 0.05 level. Admittedly, although these two factors are statistically significant, the percentage of variance they explain is fairly insubstantial. This suggests further investigation based on other personal attributes may be helpful and that interpersonal deviation appears to be naturally high even within population groups.

Table 5-5 Univariate Discrepancy Analysis

Variable	SSW	SSB	Pseudo F	Pseudo R^2	p-value[a]
gender	1157.163	30.292	1.387	0.026	0.108
agegroup	1121.836	65.618	0.994	0.055	0.485
caracc	1141.061	46.393	1.057	0.039	0.393
time	1113.162	74.292	1.735	0.063	0.001***
incomelevel	1160.366	27.089	0.607	0.023	0.946
hhsize	1139.047	48.408	0.722	0.041	0.900
lifecycle	1086.152	101.303	1.166	0.085	0.176
dist2expo	1120.461	66.994	1.555	0.056	0.016**

a. p-value is obtained via permutation test with 1000 iterations

With the important socio-demographic variables identified, it maybe of interest to test whether the discrepancy within groups differs significantly. Originally choosing the Bartlett T for within group discrepancy homogeneity test (Studer et al., 2010), Studer et al. (2011) indicated that Bartlett T is very sensitive to the case distribution across groups so is unsuitable for the randomized permutation test. Instead, the author proposed an alternative approach generalized based on a Levene test which is more powerful with randomization tests. Table 5-6 presents the results from both tests applied to our dissimilarity matrix. Again, the p-value is obtained from permutation test with 1,000 iterations. The significance level suggested by the Bartlett and Levene tests are largely consistent. The three subgroups of participants categorized by the time variable (unemployed, part-time employed/studying, and full-time employed/studying) exhibit statistically significant differences in weekly trip chaining behavior discrepancy. The difference also exists subgroups of participants categorized by different vehicle accessibility levels (no license/vehicle, own and share vehicle, own vehicle and does not share). However, the factor categorizing participants in terms of their home's distance from an Expo station (dist2expo), which is found to be to be statistically significant in explaining discrepancies in the univariate analysis in the previous section, was not associated with statistically significant differences across levels.

Table 5-6 Homogeneity Test of the Within-group Discrepancy

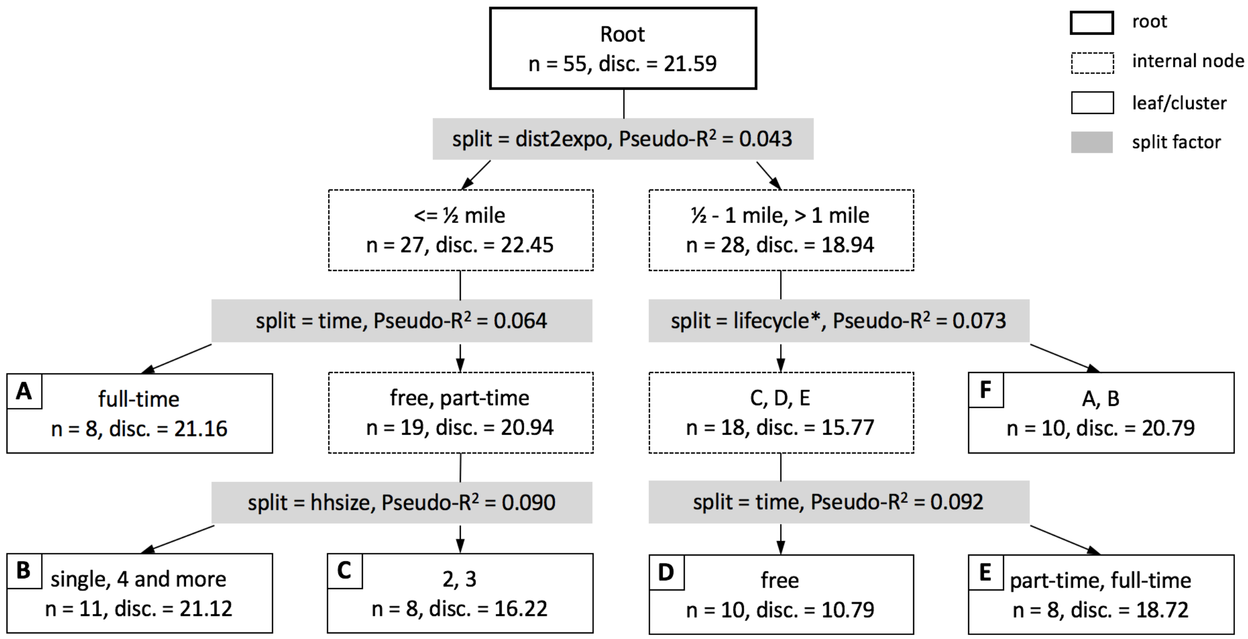
Variable (# of level)	**Bartlett**		**Levene**	
	T	**p-value**[a]	**L**	**p-value**[a]
gender (2)	0.042	0.544	0.398	0.542
agegroup (4)	0.211	0.595	0.800	0.485
caracc (3)	0.666	0.071*	3.104	0.051*
time (3)	1.370	0.003***	7.681	0.001***
incomelevel (3)	0.274	0.326	1.282	0.296
hhsize (4)	0.496	0.231	1.895	0.143
lifecycle (5)	0.765	0.228	2.000	0.117
dist2expo (3)	0.533	0.105	2.534	0.095*

a. p-value is obtained via permutation test with 1,000 iterations

Tree Structure Analysis of Representative Pattern

Using of the same set of population group factors, a regression tree is built to cluster the 55 individuals based on the pairwise dissimilarity matrix measuring trip chaining behavior (Figure 5-8). The root of the tree is on the top of the figure and the tree grows from top to bottom. The splitting factor and its pseudo-R^2 from univariate discrepancy analysis are presented in the gray box below its corresponding parent node. Node size and within node discrepancy are displayed for each node along with the splitting criterion. Note that the term of "discrepancy" is a generalized concept for "variance". So a "univariate discrepancy analysis" is similar to a "one-way ANOVA" and "within node discrepancy" is similar to "within group variance". In addition, the label of cluster index is included at the upper-left corner of each leaf node (node with no child).

The result shows that 6 population clusters (identified as leaf/cluster A-F in Figure 5-8) are identified to explain the total discrepancy in the sample's trip chaining behavior. The global pseudo-R^2 is 0.16, which means that this clustering structure explains 16% of the total discrepancy. No previous studies have utilized the same method to examine trip chaining behavior sequencing. Therefore, there are no comparable statistics to assess the quality of this regression tree. However, we find that the global pseudo-R^2 is at a similar level with that of other studies using the same method in other fields. The most relevant study, an analysis of daily activity pattern which regressed activity purpose sequencing against personal characteristics and obtained a global pseudo-R^2 of 0.189 (Kim, 2014). Another study about life trajectories had a global pseudo-R^2 of 0.187 (Studer et al., 2011). The factors that are found to be significant in the univariate discrepancy analysis in a previous section (dist2expo and time) both have been selected to grow the tree. Meanwhile, two other variables (lifecycle and hhsize) are also used to develop the tree.



**Level for lifecycle: A. young adult (below 35) family with no child, B. family with children under 12, C. family with children above 12, D. older adult (above 35) family with no child, and E. retired person*

Figure 5-8 Regression Tree of Weekly Trip Chaining Sequence

Some interaction between population grouping factors can be observed from the tree structure. For participants who live close to an Expo station (less than ½ mile away), schedule flexibility is found to be more influential on trip chaining behavior than any other factors. More specifically, people who work or study full-time appear to be different from people who are less time-constrained. For the latter, household size is a factor used to further distinguish the trip chaining patterns. Individuals who live alone and those who live with more than three people are found to have similar patterns, and together these groups are different than individuals who live with one or two other people. On the other hand, for people who live farther away from an Expo station (> ½ mile), family life cycle plays a more important role on trip chaining pattern than other factors. Within this subgroup, individuals in an earlier life stage (young single/adult family without children, and adult family with children under 12-year-old) are found to be different than individuals in a more senior life stage (family with children older than 12-year-old, older working family without children, and retired persons). According to Jones et al. (1983), these family life cycle factors are expected to be associated with quite distinguishable travel patterns due to different family structures, income levels and needs for social activity and out-of-home activities. For people in more senior life stage, schedule flexibility again shows significant impact on trip chaining patterns. In this regard, more senior participants who are free (neither employed nor student) are differentiated from the other two groups who are partially or full-time constrained due to employment or school obligations.

The time-series and distribution plot are used to visually inspect the trip chaining patterns across the population clusters identified as leaf. Note that Clusters A, B, and C are for people who live within ½ mile from an Expo station and Clusters D, E, and F are for those who live further away. Overall, Clusters A, B, and C exhibit higher daily travel demand and more complex trip chaining behavior (longer tails and darker cells), while Clusters D, E, and F show lower level of out-of-home travel demand and simpler chaining behavior (shorter tails and lighter cells). The difference described above can be observed from both time-series and distribution plots (Figure 5-9).

People in Cluster A are those who live close to an Expo station and are full-time workers or students. Their time-series plot indicates they tend to have a regular daily pattern compared to other population clusters. Moreover, their travel timing is similar given their colored cells are well aligned, especially from Monday to Thursday. This seems reasonable since Cluster A is composed of participants who face strong time-constraints. From the distribution plot, we can see most of them choose a mixture of simple and complex trip chaining.

Although Clusters B and C are both comprised of those who live close to an Expo station and are part-time worker/students or non-employed people, Cluster B includes those who live alone or with more than three other people and Cluster C includes those living in a household with two or three members. Compared to Cluster A, the travel demand of those in Cluster B is more evenly spread over the week and those in Cluster B have a mixture of simple and complex trip chaining according to the distribution plot. Participants in Cluster C seem to have much less intensive trip chaining behavior than Clusters A and B, given their colored cells are sparse in the time series plot and their colors are light in the distribution plot.

Clusters D and E are both comprised of those who live farther from an Expo station (> ½ mile) and are in a more senior life stage (family with children older than 12-year-old, older working family without children, and retired persons). Cluster D includes those with no employment and not enrolled in educational program and Cluster E includes those who employed or attending school, either full-time or part-time. Most of the colored cells in Cluster D are not aligned vertically in time (time-series plot) possibly reflecting this group faces lower time constraints. They also show limited travel demand and tend to undertake simpler trip chains based on the distribution plot. Cluster E shows more out-of-home activities and more complex trip chains than Cluster D, but the interpersonal alignment by time and weekly regularity is very limited.

Cluster F is both comprised of those who live further from an Expo station (> ½ mile) and are in an earlier life stage (young single/adult family without children, or adult family with children under 12-year-old). Based on the time-series plot, Cluster F is associate with slightly more regular travel patterns on weekdays than weekends. Participants in Cluster F seem to have higher out-of-home activity demand than Clusters D and E according to the distribution plot given the tails are on average longer compared to those for Clusters D and E. Participants in Cluster F, though, tend to make non-complex trip chains given a majority of the colored cells are light.

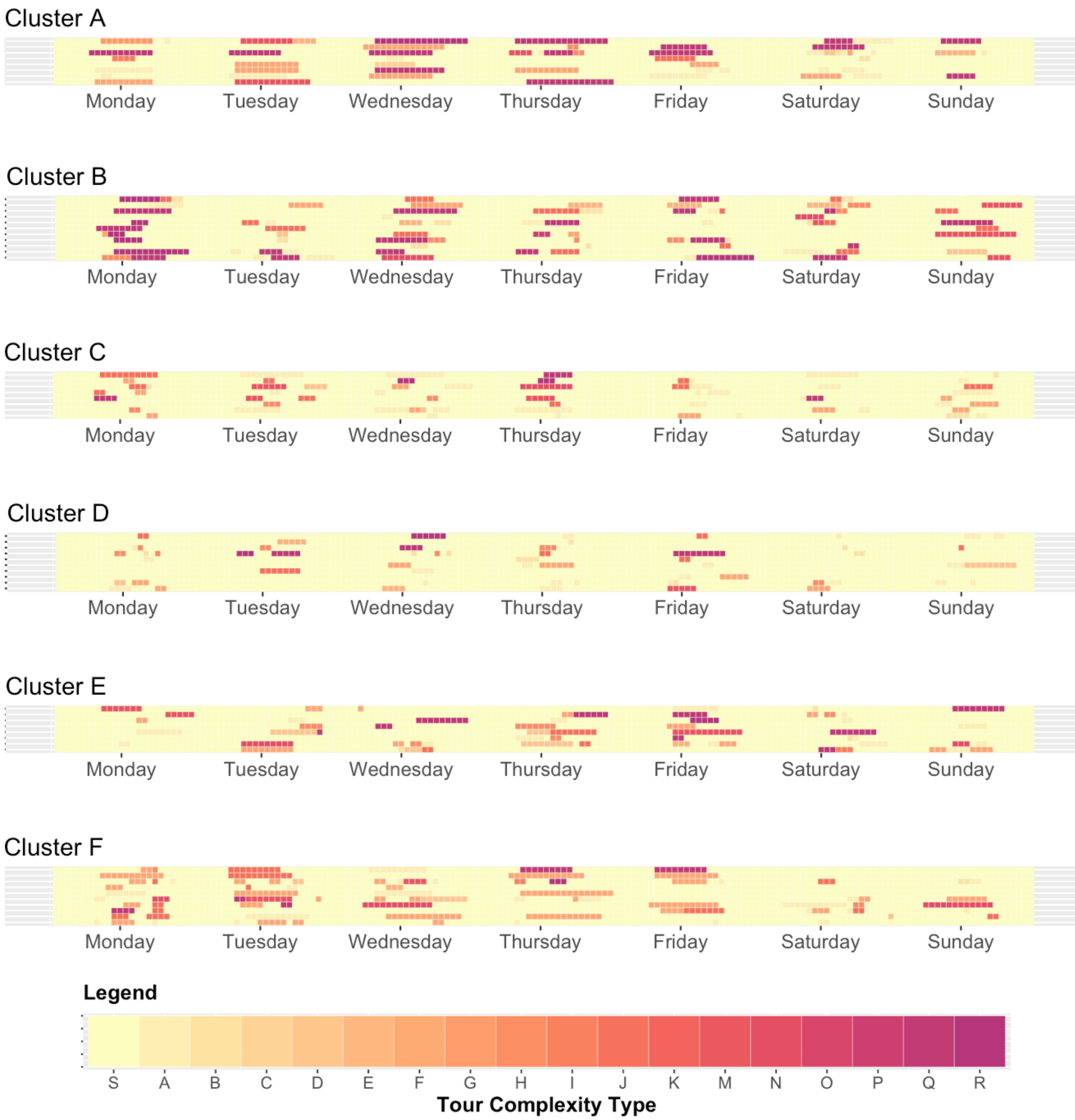


Figure 5-9 Time-series Plot by Population Cluster

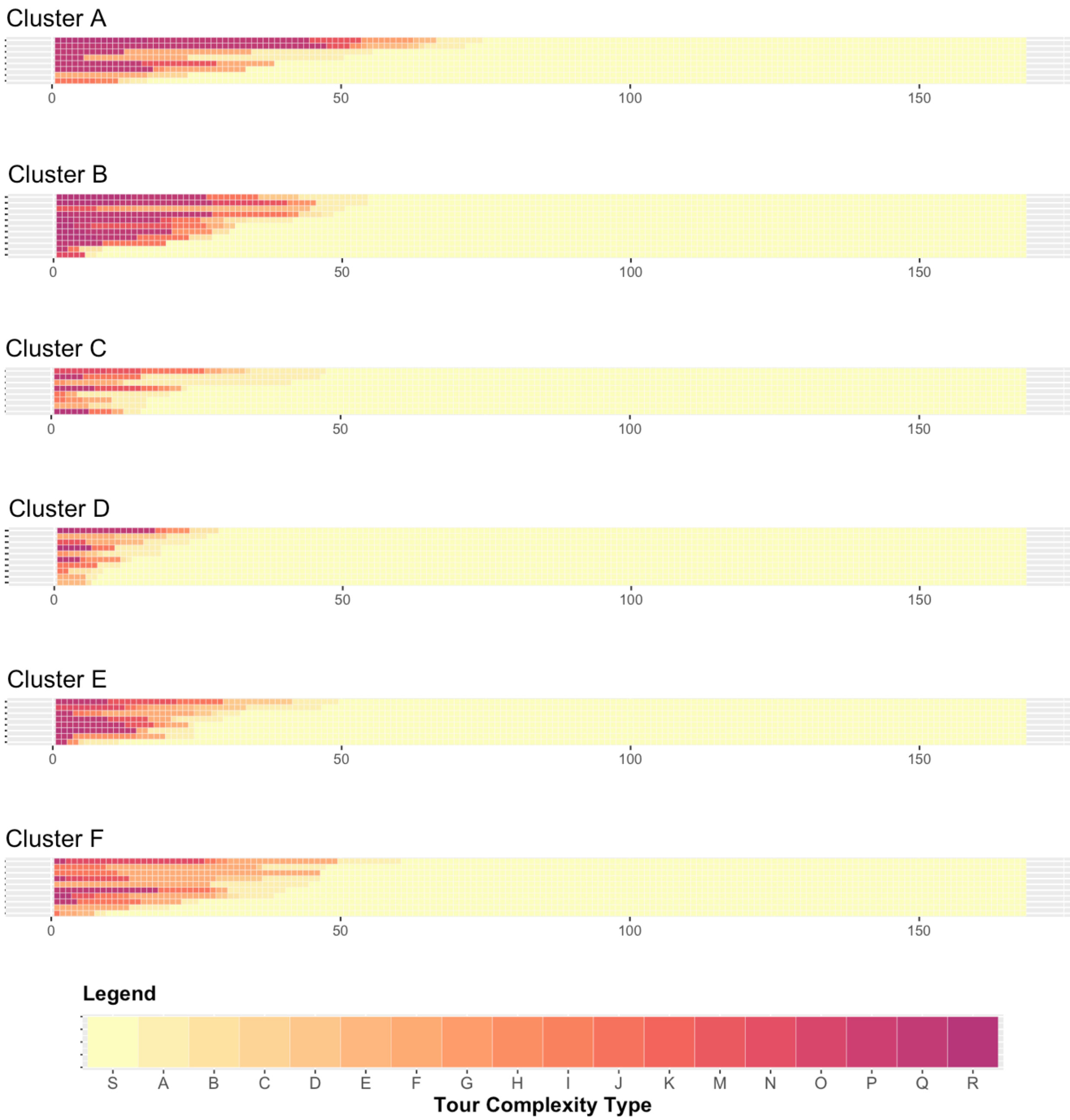


Figure 5-10 Distribution Plot by Cluster by Population Cluster

Travel and Trip Chaining Pattern by Cluster

Given the population clusters have been identified and the difference among clusters have been visually inspected, the next step is to investigate the travel-activity characteristics of each cluster. Three aspects that are directly involved in our SAM are analyzed reviewed by cluster: 1) timing and frequency of daily travel, and 2) choice in TCI tour type.

Table 5-7 Daily Activity/travel Hour Duration by Cluster

Cluster	Mon.	Tue.	Wed.	Thur.	Fri.	Sat.	Sun.	Overall
A	7.17	10.17	9.57	10.40	6.83	8.50	6.00	8.38
B	6.89	5.14	6.10	6.00	4.57	5.25	6.00	5.71
C	3.71	6.29	5.57	4.00	5.00	5.75	3.75	4.87
D	4.00	5.25	3.29	2.75	3.86	1.60	5.75	3.78
E	1.00	5.17	4.20	7.33	6.00	3.43	6.75	4.84
F	7.20	10.00	8.63	9.86	7.83	1.17	4.20	6.98

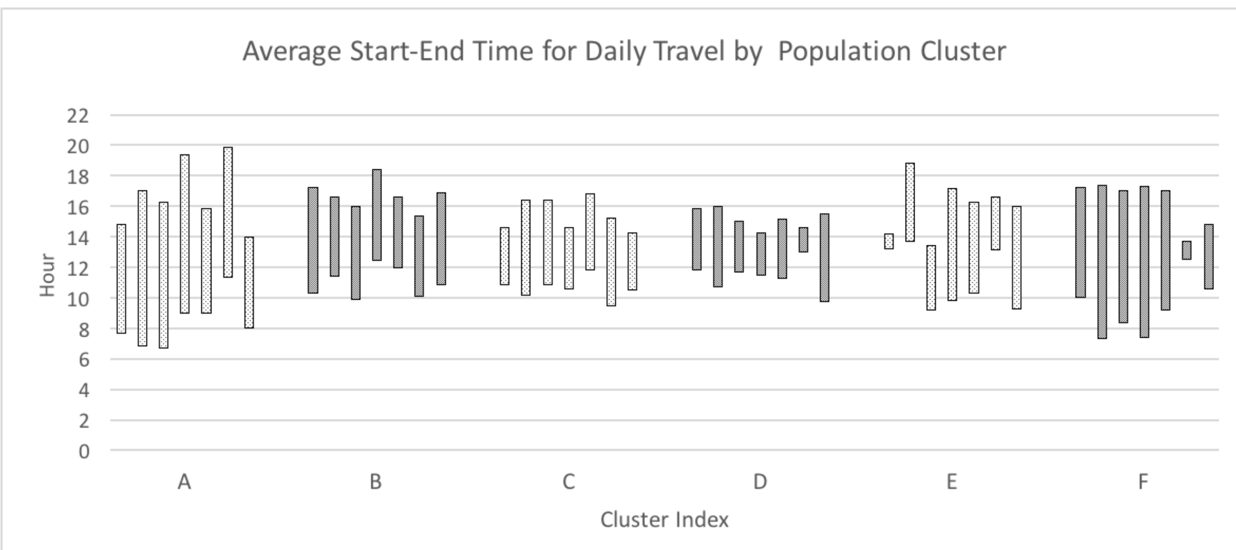


Figure 5-11 Timing of Daily Travel by Cluster

Table 5-7 lists the average activity/travel hour duration for each cluster by day of week. Figure 5-11 visualizes the start and end time of the average period of activity/travel. For each cluster, the 7 bars represent the 7 days of a week from Monday to Sunday. Cluster A (<= ½ mile to Expo stations, full-time worker/student) and Cluster F (> ½ mile to Expo, early stage in life cycle) have longer activity/travel durations than the other clusters, but Cluster F shows a decreased travel demand on weekends. It also has more of an aligned home-return hour during weekdays while Cluster A show much less regularity in departure or return hour. Clusters B (<= ½ mile to Expo stations, free or part-time worker/student, single or large family), C (<= ½ mile to Expo stations, free or part-time worker/student, two or three-person family), D (> ½ mile to Expo stations, senior life stage, not employed/enrolled in education program) and E (> ½ mile to Expo stations, senior life stage, part-time/full-time worker/student) have shorter activity/travel hours than Clusters A and F. People in these clusters on average start their daily travel after 10:00, much later than those in Clusters A and F who start their activity/travel between 7:00 to 9:00. People in Clusters B and E tend to end the day between 16:00 and 18:00, while people in Clusters C and D tend to end the day between 14:00 and 16:00. Overall, Cluster D has the shortest out-of-home duration. Moreover, the daily variation between days of the week is very high in Cluster E. Cluster E includes people above 35-year-old with/without older school-age children and people above 65 who are still working. The high daily variability of this group supports the claim that families with older school-age children have more intensive travel demands as children start to have greater social needs but are still dependent on parents to drive them (Jones et al., 1983). These patterns are also consistent with the findings by Kim (2014) that suggest older workers spend less time on work and have less clear peak hour points on weekdays compared to younger workers.

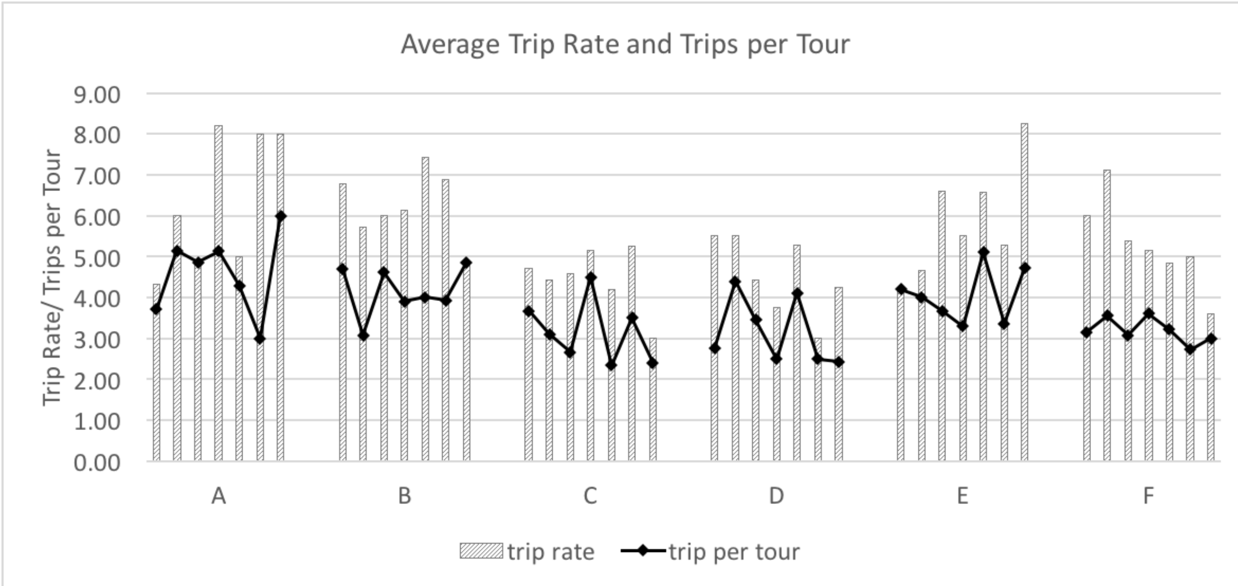


Figure 5-12 Daily Trip Rate and Trips per Tour by Cluster

Figure 5-12 presents the daily trip rate and average number of trips chained per tour. Similar to Figure 5-11, each population cluster contains 7 bars representing Monday through Sunday. Note that only days that contain out-of-home activities are included and days with no travel-activity are excluded from this analysis. Hence, the trip rates reported could be inflated. The trip rate is increasing from Monday to Sunday for Clusters A and E, while it tends to be decreasing more or less for other clusters. As for number of trips per tour, participants in Clusters A, B and E tend to chain more overall than people in Clusters C, D and F. Although Cluster F shows a high trip rate, participants in this cluster chain less. This could be due to the reason that Cluster F, who's defining characteristic is an early life cycle, has more out-of-home activity demand and is less obligated to family which leads to less time and monetary constraints. Cluster C and D have similar patterns in trip rate and chaining intensity, which are both lower than other clusters. The common characteristics between people in Cluster C and D is that they are not full-time employed or enrolled in school. So it seems very likely that the low trip rate of these clusters are associated with the lack of commuting trips. However, the activity/travel duration for participants in Cluster C is twice long as that of participants in Cluster D. Recall that the overall average number of trips chained per tour is highest on Friday and lowest on Saturday. Based on Figure 5-12, we can see the complexity of chaining on Friday is contributed mainly by people in Clusters D and E, while the simplicity of chaining on Saturday could be contributed mainly by people in Clusters A and F.

Table 5-8 Out-of-home Time Allocation by TCI Type

TCI Code	**Description**				**Travel-Activity Time Allocation (minute)[b]**					
	Seg.[a]	**Clst.[a]**	**Div.[a]**	**Effi.[a]**	**A**	**B**	**C**	**D**	**E**	**F**
A	single	0	0	1	62	25	93	24	47	54
B		1	0	1						
C	simple	0	0	0						
D		0	0	1					23	10
E		0	1	0		15				
F		0	1	1	78		23	16	32	116
G		1	0	0						
H		1	0	1					14	
I		1	1	0						
J		1	1	1		35	25	10	10	33
K	complex	0	0	0						
L		0	0	1						
M		0	1	0	26	19	12		34	
N		0	1	1					13	21
O		1	0	0						
P		1	0	1						
Q		1	1	0	122	71	10	13	69	20
R		1	1	1	16	31				
Daily Total (minute)					343	241	190	101	253	269

a. Seg. = Segment Index; Clst. = Cluster Index; Div. = Diversity Index; Eff. = Efficiency Index
b. statistics is only shown for TCI types with more than 10 minutes per capita per day

Finally, the distribution of time on different tour types are examined for each population cluster (Table 5-8). For simplicity, the allocated time is only shown for tour types that lasted for more than 10 minutes per traveler per day. The top 3 tour types are highlighted for each cluster as well. Consistent with what can be observed from the distribution plot, people in Clusters A, B and E spend more time on complex and simple chained tours, while Clusters C, D and F have a higher percentage on simpler tours, especially for Clusters C and D.

Conclusion and Discussion

This study explores a new approach for studying longitudinal travel-activity patterns that characterizes out-of-home activity using TCI tour types. The combination of SAM and discrepancy analysis provides an analytical tool that can embed socio-demographic factors into the clustering process, rather than analyzing their influence in a posterior way. Based on the distance matrix computed from all the individual pairs, six population groups or clusters are identified. The tree-structured clustering method sheds light on how socio-demographic factors influence weekly travel patterns and how factors interact with each other.

Results demonstrate a new methodological approach for the travel behavior literature and demonstrates that trip chaining information extracted from GPS data provide valuable information for identifying population groups with different travel habits. The study also shows that population groups have different tendencies for choosing tour complexity types including the spatial relationship between destinations, routing arrangement and land use combination, not merely how many trips to make per tour. The proposed TCI classification method provides a better approach for understanding trip chaining behavior compared to the traditional binary classification scheme (simple vs. complex) used in most previous studies.

Results also demonstrate that the characteristics of individuals, including residential location, employment status, family life cycle and household size, have an important influential on the weekly travel patterns and trip chaining complexity. Employed people or students are found to have more complex trip chaining compared to people whose out-of-home activities depend less on other household members. The tree-structured analysis indicates that interactions between socio-demographic factors across levels of clustering. For people who are unemployed or part-time employed, household size is an important discriminant of travel patterns. For older people older than 35 years old, employment status has an important influence on travel patterns.

Limitations and Future Work

Several aspects of this study could be improved. First, the alignment cost parameters could be fine tuned. Different cost settings can be applied based on the type of alignment operations (i.e., *indel* and substitution) and based on the position of alignment. Second, the current approach is uni-dimension with a single attribute due to limited sample size. A simultaneous consideration of multiple travel-activity attributes could provide more comprehensive clustering results. In particular, those attributes that are found to be associated with trip chaining behaviors, including travel mode, destinations and activity duration should be explored further. Third, there is more room for developing a scheme to incorporate geographical information into the sequence. Further, it could be worthwhile for future studies to explore alignment methods that allow for the presence of both numerical features and categorical attributes; currently SAM requires categorical variables. Finally, it is important to apply the analytical framework on a larger dataset in order to justify the methodological approach and empirical results.

Chapter 6 Conclusion

Trip chaining is a complicated, multi-dimensional phenomenon not only often involving multiple grouped trips but also involving destination choice, activity sequencing, route arrangement, and scheduling. Trip chaining complexity has been traditionally analyzed using a simplified binary classification scheme. That is, most previous studies classify a tour (or trip chain) either as “simple” or as “complex” depending on whether it includes more than one stop in the tour. This simplified classification scheme has resulted in ambiguous research findings. In an attempt to address the limited insights in the literature regarding trip chaining, this dissertation proposes a new framework to integrate multi-dimensional trip chaining characteristics into travel behavior research. Further, this dissertation demonstrates the applicability of the proposed tour complexity measurement in intrapersonal and interpersonal travel-activity pattern research.

This dissertation introduced and demonstrates applications of the Tour Complexity Index (TCI), a new approach to qualitatively assess trip chaining complexity. It characterizes the complexity of home-based tours (trip chains) from four dimensions or components: number of chained trips, the geographical relationship or spatial clustering among destinations, visiting sequence and routing, and land use diversity across destinations. These TCI components are referred to as the Segment Index, Cluster Index, Efficiency Index, and Diversity Index, respectively. Based on explorative analysis, this study first confirms that chaining behavior is common: 64% of home-based tours have at least two locations chained. Second, when tours with at least two stops chained are analyzed using the three other TCI indices, results show that tours with multiple and non-clustered stops, optimal routing, and mixed use destinations are undertaken most frequently.

This dissertation contributes to the literature in several ways. By comparing TCI classification methods with the traditional binary classification method, Component I first verifies the findings from existing trip chaining complexity studies. For number of chained trips, results agree with previous research by Currie & Delbosc (2011) who used number of trips chained instead of the binary classification method to characterize tours. Both studies have found that as the destinations are further away from a highly urbanized area (such as the central business district or downtown), tours tend to be simpler because less stops are typically chained to access such distant destinations. Component I results also show that the presence of clustered destinations (within 0.5 miles) is more associated with the use of public transit than the number of trips chained. This partially supports Ho and Mulley's research which examines the spatial relationship between activity locations and found that clustered destinations encourage the use of public transit (Ho & Mulley, 2013). The difference lies in that this previous study claims the positive association is monotonic while this dissertation study shows the association disappears when the number of chained locations reaches a certain level. This means that although trip chaining is not a barrier to transit ridership, extremely complex travel plans are still more likely to be carried out by private vehicles.

More importantly, model results of Component I demonstrate that the TCI multi-dimensional measurement offers more information than the traditional binary classification scheme regarding the behavior of chaining trips and activities. By introducing the concept of efficient routing and diverse land use across destinations, this study finds that non-motorized modes like walking has a close relationship with complex trip chains as it is the only mode that is found to be significantly associated with all four TCI indices. Besides number of chained stops and presence of a destination cluster, tours including a walking trip are highly likely to have a non-optimal route and multiple dominant land uses across destinations. The result suggests that non-motorized travel mode might be underestimated in the existing literature. The main stream of research in trip chaining focuses on debate regarding trip chaining behaviors relating to private vehicle and public transit modes. However, the non-motorized walking mode appears to have a strong association with complex chaining behavior.

Component II contributes to the literature by assessing intrapersonal daily variability characterized by jointly considering the influence of trip chaining and mode choice. The analysis first categorizes home-based tours according to the four TCI indices; then it uses a sequential alignment approach to quantify the daily deviation for tours made by the same person. Although the current study has some methodological differences with previous studies, it confirms the findings of previous variability-related studies that the amount of within-individual day-to-day variability in generic travel behavior is significant (Hanson & Huff, 1988; Pas, 1987; Pas & Sundar, 1995; Raux et al., 2016). From the perspective of trip chaining, the study reinforces the argument that the common single-day travel survey may produce biased results because the collected data cannot fully capture the variation in travel behavior that can be observed using week-long travel data.

Existing research has examined the association of traveler characteristics and variability because if the daily variability level can be associated numerically with certain social-demographical factors, it would be easier to incorporate the daily variability factors into travel demand model development and transportation policy design. Mainly constrained by the nature of datasets, previous research has attempted to identify such determinants using cross-sectional analysis. However, little evidence has been found to support the hypothesis that level of within individual variability is connected with social-demographical factors (Hanson & Huff, 1982; Kitamura & Van Der Hoorn, 1987). Component II of this dissertation takes a different approach to this problem by taking advantage of the longitudinal nature of the Expo tour dataset. A before-and-after analysis is performed to examine whether the Expo Line would affect the variability. First, the study finds a statistically significant decline in intrapersonal variability of chaining complexity for people who live within ½ mile of an Expo station. In other words, the combination of trip chaining patterns and mode choice for people who live close to an Expo stations became more predictable. Second, it cannot be concluded that any socio-demographic factors are related with this before-and-after change. This finding is consistent with previous research but from another direction.

Component III extends the research of Component II by analyzing the deviation between individuals. That is, how different people chain trips and activities. Component III also aims to explore the applicability of TCI on established travel-activity pattern recognition methods (Wilson, 1998). The major contribution of this component is that it directly classifies out-home travel behavior and embeds socio-demographic information in the process of pattern recognition. First, this study addresses the limitations of existing studies that heavily rely on activity information which is an upper-scale concept according the theory that travel is a demand induced by activity (Recker, 2001). However, accurate activity information is difficult and expensive to collect as pointed out by previous research (Kim, 2014; Saneinejad & Roorda, 2009). This dissertation study also focuses the pattern analysis on out-home travel information derived from GPS data, a data source that is more accessible than traditional travel survey data. Second, this study contributes by applying discrepancy analysis and regression tree methods which allows socio-demographic information to be integrated into the pattern recognition process at an early stage rather than as a posterior analysis as done in previous studies. Component III demonstrates that the proposed TCI method can be used in multi-day travel pattern analysis. It also investigates the individual characteristics that influence travel patterns and the tendency to choose complex trip chains. A set of individual characteristics, including residential location, schedule flexibility, household size, and family life cycle, is found to be significantly influential to weekly travel patterns. Further, this factors also influence day-to-day variability of out-home activity hour, tour/trip rates, and choice of tour type defined by tour complexity.

Reference

Anderson, M. J. (2001). A new method for non parametric multivariate analysis of variance. *Austral Ecology*, *26*(2001), 32–46. http://doi.org/10.1111/j.1442-9993.2001.01070.pp.x

Baddeley, A. (2008). *Analysing spatial point patterns in R. CSIRO, Canberra, Australia*. http://doi.org/10.1007/s00415-011-6369-2

Boarnet, M. G., Houston, D., & Spears, S. (2013). *The Exposition Light Rail Line Study: "Before-After" Opening Travel Impacts and New Resident Sample Preliminary Analysis*.

Bowman, J. L. (2000). Activity-based disaggregate travel demand model system with activity schedules, *35*.

Cervero, R., & Kockelman, K. (1997). Travel Demand and the 3Ds: Density, Diversity, and Design. *Transportation Research Part D: Transport and Environment*, *2*(3), 199–219.

Currie, G., & Delbosc, A. (2011). Exploring the trip chaining behaviour of public transport users in Melbourne. *Transport Policy*, *18*(1), 204–210. http://doi.org/10.1016/j.tranpol.2010.08.003

de Abreu e Silva, J., Sottile, E., & Cherchi, E. (2014). Effects of Land Use Patterns on Tour Type Choice. *Transportation Research Record: Journal of the Transportation Research Board*, *2453*, 100–108. http://doi.org/10.3141/2453-13

Ettema, D., & Lippe, T. (2009). Weekly rhythms in task and time allocation of households. *Transportation*, *36*(2), 113–129. http://doi.org/10.1007/s11116-009-9190-3

Gower, J. C. (1971). A General Coefficient of Similarity and Some of Its Properties. *International Biometric Society*, *27*(4), 857–871. Retrieved from http://www.jstor.org/stable/2528823

Greenwald, M. J., & Mcnally, M. G. (2008). *Land Use Influences on Trip Chaining in Portland , Oregon*.

Hagerstrand, T. (1970). What about people in regional science? *Papers and Proceedings of the Regional Science Association*, *24*, 122–145.

Hanson, S., & Huff, J. (1986). Classification issues in the analysis of complex travel behavior. *Transportation*, *13*(3), 271–293. http://doi.org/10.1007/BF00148620

Hanson, S., & Huff, J. O. (1982). Assessing Day-to-Day Variability in Complex Travel Pattern. *Transportation Research Record*, 18–24.

Hanson, S., & Huff, O. J. (1988). Systematic variability in repetitious travel. *Transportation*, *15*(1–2), 111–135. http://doi.org/10.1007/BF00167983

Harding, C., Miller J, E., Patterson, Z., & Axhausen W, K. (2015). Multiple Purpose Tours and

Efficient Trip Chaining: An Analysis of the Effects of Land Use and Transit on Travel Behavior in Switzerland, (April), 17p. Retrieved from https://trid.trb.org/view/1338693

Hensher, D. A., & Reyes, A. J. (2006). Trip chaining as a barrier to the propensity to use public transport, 341–361.

Ho, C., & Mulley, C. (2015). Intra-household Interactions in tour-based mode choice: The role of social, temporal, spatial and resource constraints. *Transport Policy*, *38*, 52–63. http://doi.org/10.1016/j.tranpol.2014.12.001

Ho, C. Q., & Mulley, C. (2013). Multiple purposes at single destination: A key to a better understanding of the relationship between tour complexity and mode choice. *Transportation Research Part A: Policy and Practice*, *49*, 206–219. http://doi.org/10.1016/j.tra.2013.01.040

Houston, D., Luong, T. T., & Boarnet, M. G. (2014). Tracking daily travel: Assessing discrepancies between GPS-derived and self-reported travel patterns. *Transportation Research Part C: Emerging Technologies*, *48*, 97–108. http://doi.org/10.1016/j.trc.2014.08.013

Joh, C.-H., Arentze, T., & Timmermans, H. (2001). Pattern Recognition in Complex Activity Travel Patterns: Comparison of Euclidean Distance, Signal-Processing Theoretical, and Multidimensional Sequence Alignment Methods. *Transportation Research Record*, *1752*(1), 16–22. http://doi.org/10.3141/1752-03

Joh, C. H., Arentze, T. A., & Timmermans, H. J. P. (2001). A position-senstive sequence-alignment method illustrated for space-time activity-diary data. *Environment and Planning A*, *33*(2), 313–338. http://doi.org/10.1068/a3323

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. http://doi.org/10.1016/S0191-2615(01)00009-1

Jones, P. M., Dix, M. C., Clarke, M. I., & Heggie, I. G. (1983). *Understanding Travel Behavior*. (G. P.B., Ed.). Gower.

Jou, R., & Mahmassani, H. S. (1997). Comparative Analysis of Day-to-Day Trip-Chaining Behavior of Urban Commuters in Two Cities. *Transportation Research Record: Journal of the Transportation Research Board*, *1607*(1), 163–170. http://doi.org/10.3141/1607-22

Kim, K. (2014). Discrepancy Analysis of Activity Sequences. *Transportation Research Record: Journal of the Transportation Research Board*, *2413*, 24–33. http://doi.org/10.3141/2413-03

Kitamura, R., & Van Der Hoorn, T. (1987). Regularity and irreversibility of weekly travel behavior. *Transportation*, *14*(3), 227–251. http://doi.org/10.1007/BF00837531

Krizek, K. J. (2003). Neighborhood services , trip purpose , and tour-based travel.

Transportation, *30*, 387–410.

Kruskal, J. B. (1983). An overview of sequence comparison: Time warps, string edits, and macromolecules. *SIAM Review*, *25*(2), 201–237.

Lee, M., & McNally, M. G. (2006). An empirical investigation on the dynamic processes of activity scheduling and trip chaining. *Transportation*, *33*(6), 553–565. http://doi.org/10.1007/s11116-006-7728-1

Levinson, D., & Kumar, A. (1995). Activity, Travel, and the Allocation of Time. *Journal of American Planning Association*, *61*(4), 458–470. http://doi.org/10.1080/01944369508975657

Ma, J., & Goulias, K. (1997). A dynamic analysis of person and household activity and travel patterns using data from the first two waves in the Puget Sound Transportation Panel. *Transportation*, *24*(3), 309–331.

Ma, J., Mitchell, G., & Heppenstall, A. (2014). Daily travel behaviour in Beijing, China: An analysis of workers' trip chains, and the role of socio-demographics and urban form. *Habitat International*, *43*, 263–273. http://doi.org/10.1016/j.habitatint.2014.04.008

Maat, K., & Timmermans, H. (2006). Influence of Land Use on Tour Complexity: A Dutch Case. *Transportation Research Record: Journal of the Transportation Research Board*, *1977*(1), 234–241. http://doi.org/10.3141/1977-29

Maat, K., & Timmermans, H. J. P. (2009). A causal model relating urban form with daily travel distance through activity/travel decisions. *Transportation Planning and Technology*, *32*(2), 115–134. http://doi.org/10.1080/03081060902861285

McGuckin, N., Zmud, J., & Nakamoto, Y. (2005). Trip-Chaining Trends in the United States Understanding Travel Behavior for Policy Making. *Transportation Research Record: Journal of the Transportation Research Board*, *1917*, 199–204.

Moore, D. S., McCabe, G., Duckworth, W., & Sclove, S. (2003). *Bootstrap Methods and Permutation Tests*. *The Practice of Business Statistics: Using Data for Decisions*. W. H. Freeman, New York. http://doi.org/10.1002/wics.182

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. http://doi.org/10.1016/0022-2836(70)90057-4

Pas, E. I. (1987). Intrapersonal variability and model goodness-of-fit. *Transportation Research Part A: General*, *21*(6), 431–438. http://doi.org/10.1016/0191-2607(87)90032-X

Pas, E. I., & Koppelman, F. S. (1986). An examination of the determinants of day-to-day variability in individuals' urban travel behavior. *Transportation*, *13*(2), 183–200. http://doi.org/10.1007/BF00172463

Pas, E. I., & Sundar, S. (1995). Intrapersonal variability in daily urban travel behavior: Some additional evidence. *Transportation*, *22*(2), 135–150. http://doi.org/10.1007/BF01099436

Perchoux, C., Kestens, Y., Thomas, F., Van Hulst, A., Thierry, B., & Chaix, B. (2014). Assessing patterns of spatial behavior in health studies : Their socio-demographic determinants and associations with transportation modes (the RECORD Cohort Study). *Social Science & Medicine*, *119*, 64–73. http://doi.org/10.1016/j.socscimed.2014.07.026

Primerano, F., Taylor, M. a. P., Pitaksringkarn, L., & Tisato, P. (2007). Defining and understanding trip chaining behaviour. *Transportation*, *35*(1), 55–72. http://doi.org/10.1007/s11116-007-9134-8

Raux, C., Ma, T. Y., & Cornelis, E. (2016). Variability in daily activity-travel patterns: the case of a one-week travel diary. *European Transport Research Review*, *8*(4). http://doi.org/10.1007/s12544-016-0213-9

Recker, W. W. (2001). A bridge between travel demand modeling and activity-based travel analysis, *35*, 481–506.

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. http://doi.org/10.1016/0191-2607(85)90064-0

Saneinejad, S., & Roorda, M. J. (2009). Application of Sequence Alignment Methods in Clustering and Analysis of Routine Weekly Activity Schedules. *Transportation Letters: The International Journal of Transportation Research*, *1*(January), 197–211. http://doi.org/10.3328/TL.2009.01.03.197-211

Santos, A., McGuckin, N., Nakamoto, H. Y., D., G., & Liss, S. (2011). *Summary of Travel Trends: 2009 National Household Travel Survey*.

SCAG. (2016). *The 2016-2040 Regional Transportation Plan/Sustainable Communities Strategy*. Retrieved from http://scagrtpscs.net/Pages/2016RTPSCS.aspx

Schlich, R., & Axhausen, K. W. (2003). Habitual travel behaviour - evidence from a six week travel diary. *Transportation*, *30*(1), 13–36.

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. http://doi.org/10.1111/j.1467-8306.2007.00536.x

Stopher, P. R., Kockelman, K., Greaves, S., & Clifford, E. (2008). Reducing burden and sample sizes in multiday household travel surveys. *Transportation Research Record: Journal of the Transportation Research Board*, *2064*, 12–18. http://doi.org/10.3141/2064-03

Stopher, P., & Zhang, Y. (2011). Repetitiveness of Daily Travel. *Transportation Research Record*, *2230*, 75–84. http://doi.org/10.3141/2230-09

Studer, M., Ritschard, G., Gabadinho, A., & Müller, N. (2011). Discrepancy Analysis of State Sequences. *Sociological Methods & Research*, *40*(3), 471–510.

Studer, M., Ritschard, G., Gabadinho, A., & Müller, N. S. (2010). Discrepancy Analysis of Complex Objects Using Dissimilarities. In *Advances in knowledge discovery and management. Springer Berlin Heidelberg* (pp. 3–19).

Wilson, C. (2001). Activity Patterns of Canadian Women: Application of ClustalG Sequence Alignment Software. *Transportation Research Record*, *1777*(1), 55–67. http://doi.org/10.3141/1777-06

Wilson, C. (2008). Activity patterns in space and time: Calculating representative Hagerstrand trajectories. *Transportation*, *35*(4), 485–499. http://doi.org/10.1007/s11116-008-9162-z

Wilson, W. (1998). Activity pattern analysis by means of sequence-alignment methods. *Environment and Planning A*, *30*(6), 1017–1038. Retrieved from http://www.envplan.com/epa/fulltext/a30/a301017.pdf

Xianyu, J. (2013). An Exploration of the Interdependencies between Trip Chaining Behavior and Travel Mode Choice. *Procedia - Social and Behavioral Sciences*, *96*(Cictp), 1967–1975. http://doi.org/10.1016/j.sbspro.2013.08.222

Ye, X., Pendyala, R. M., & Gottardi, G. (2007). An exploration of the relationship between mode choice and complexity of trip chaining patterns. *Transportation Research Part B: Methodological*, *41*(1), 96–113. http://doi.org/10.1016/j.trb.2006.03.004