SPACE-TIME MODELING OF URBAN POPULATION DAILY TRAVEL-ACTIVITY PATTERNS USING GPS TRAJECTORY DATA

by

Ruojing Wang Scholz, B.E., M.E.

A dissertation submitted to the Graduate Council of
Texas State University in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
with a Major in Geographic Information Science
May 2018

Committee Members:

Yongmei Lu, Chair

Denise Blanchard

Edwin Chow

Jean-Claude Thill

FAIR USE AND AUTHOR'S PERMISSION STATEMENT

Fair Use

Duplication Permission

ACKNOWLEDGEMENTS

Finishing up my dissertation has a great impact on my career and life. I would like to thank the people who have supported and helped me during this process, especially my advisor, the chair of my committee, Dr. Yongmei Lu. She taught me everything about academic research, teaching, and writing. I highly appreciate the tremendous time and efforts she spent to improve my work. She has shown me what a great example of a college professor should be. I would also like to thank my dissertation committee members, Dr. Denise Blanchard, Dr. Edwin Chow, and Dr. Jean-Claude Thill, who have provided extensive professional guidance and support for my dissertation. I am also very grateful to the Department of Geography, Texas State University for giving me the opportunity to study in the PhD program and offering me a graduate assistantship as financial support. I would like to thank all the faculty, staff, and colleagues at Texas State University whom I have had the pleasure to work with during this time.

Last but not least, I would like to thank my family members and friends who have always supported and encouraged me in the pursuit of my PhD degree, especially my loving husband, Michael, and my wonderful daughter, Chelsea. Without you, this would not be possible. Thank you.

TABLE OF CONTENTS

LIST OF TABLES

Table **Page**

LIST OF FIGURES

Figure **Page**

LIST OF ABBREVIATIONS

Abbreviation	Description
CRAWDAD	Community Resource for Archiving Wireless Data
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
DB-SMoT	Direction-Based Stops and Moves of Trajectories
GIS	Geographic Information Science
GPS	Global Positioning System
HMM	Hidden Markov Model
LBS	Location-Based Services
MAUP	Modifiable Areal Unit Problem
MBB	Minimal Bounding Boxes
MDSAM	Multidimensional Sequence Alignment Method
MTUP	Modifiable Temporal Unit Problem
SAM	Sequence Alignment Method
SMoT	Stops and Moves of Trajectories
T-A	Travels and activities
UDSAM	Unidimensional Sequence Alignment Method

ABSTRACT

People conduct travel to engage in a variety of activities every day. Their activities include working, dining, shopping, and so forth. Their travels include trips to these activity locations. Each individual may have the same or a different schedule of travels and activities (T-A for short) on different days. A group of individuals may have similar or completely different daily T-A routines. Human T-A behaviors are very complex. Similarities and differences in space, time, and attribute exist among different individuals and on different days. Modeling human T-A patterns at the individual and collective levels is a research challenge in the field of geography and transportation. Previous methods for modeling collective T-A patterns failed to combine the spatial and temporal dimensions. For individual T-A pattern modeling, existing methods do not combine travel and activity events in one model nor do they make a connection between them.

This research aims to develop effective methods to model urban population collective activity patterns and individual daily T-A patterns. To accomplish this, the proposed method for modeling collective activity patterns identified the locations and times of activity hot spots in a large, metropolitan city, San Francisco, and tracked the evolvement process of these hot spots over time. GPS trajectory data of 536 taxi cabs over twenty-two days in San Francisco were analyzed to demonstrate the effectiveness of the proposed method and reveal collective activity patterns across the city and over time. Taxi passengers' pick-up and drop-off locations and times were extracted from the

trajectories and treated as passengers' activity instances. Census tracts with a significantly large number of activity instances during a one-hour interval were defined as activity hot spots. The evolving process of activity hot spots included emergence, expansion, stableness, shrinkage, displacement, and decease. This process was evaluated relative to the hot spot status at the census tract that hosted the hot spot and its neighboring tracts at two consecutive time intervals. The results indicated that collective activity patterns on a weekday were substantially different from those on a weekend day, and historical average data might be used to predict up-coming collective activity patterns. The proposed method for individual daily T-A pattern modeling identified the most frequent daily T-A events and their sequential relationships. The T-A events of an individual in one day was represented with a sequence of T-A elements. Daily T-A sequences of an individual from different days were grouped based on element similarity. The representative sequences of each group revealed the individual's different daily routines. GPS trajectory data for two individuals living in the northern area of Beijing was used to demonstrate the proposed method. The results showed that an individual might have several different daily T-A patterns or no apparent pattern. The proposed methods provide researchers with tools to study complex T-A behaviors of urban people, and calls into question a fundamental assumption in transportation geography that each individual repeats their T-A routine every day (Huff and Hanson 1986; Hanson and Huff 1988; Stopher and Zhang 2011). Further testing of this assumption may change the current design of transportation surveys as well as the modeling of transportation demand, urban

planning, traffic management, the delivery of Location-Based Services (LBS), and other services.

I. INTRODUCTION

Research Problem

Every individual conducts a number of travels (e.g., a trip from home to the workplace, a trip from home to a grocery store, etc.) and activities (e.g., working, exercising, shopping, etc.) every day. The number, type, location, time, and order of these daily travels and activities (T-A for short) may be particular to each person on each day. For each individual, T-A may be similar in some days and very different in other days. A group of individuals may share the same daily T-A routines, while others may have very different ones. In short, human T-A behaviors are very complex. Repetitions and variations exist in each individual's daily T-A behaviors. Similarities and differences can be found among different individuals. Spatial, temporal, and attribute information is all involved to describe individual or collective T-A behaviors. Thus, modeling human T-A patterns at the individual or a collective level has been a research challenge for transportation and geography researchers.

Previous studies have used different methods for modeling T-A patterns at the individual level and collective levels. For collective activity pattern modeling, descriptive statistical methods summarize the characteristics (e.g., average time spent on different types of activities in a day) of activities conducted by a group of people and associate them with socioeconomic characteristics (e.g., Van Der Hoorn 1979; Ding, Lu, and Zhang 2016); however, neither activity location nor timing is incorporated in the patterns. The space-time graph method identifies representative space-time graphs to represent typical daily T-A patterns of individuals (Recker, McNally, and Root 1985; Chen et al. 2011); however, absolute activity locations are not incorporated in the model. The

Sequence Alignment Method (SAM) discovers typical daily activity patterns for different groups of people (e.g., Joh, Arentze, and Timmermans 2005; Wilson 2008). Similar to the space-time graph method, it is suitable for discovering different lifestyles, but cannot reveal the citywide activity landscape, nor the changes of this landscape over time. Activity density surfaces method reveals citywide activity patterns but fails to show how the pattern changes over time (e.g., Kwan 2000; Chen et al. 2011). A new space-time modeling technique needs to be developed to combine the spatial and temporal dimensions. This technique would not only identify collective activity patterns across the city landscape but also trace the evolvement of the patterns over time.

For individual daily T-A pattern modeling, the daily activity sequence method identifies representative daily activity patterns accounting for activity ordering, type, time and/or other attributes, but ignores the locations of activities and the travels between them (e.g., Pas 1983; Huff and Hanson 1986; Hanson and Huff 1988). The activity location modeling method generates probabilistic models for transitioning between anchor locations, but does not reveal an individual's daily T-A in a sequenced manner (Ashbrook and Starner 2003; Hariharan and Toyama 2004). The travel route modeling method reveals an individual's frequently traveled routes but ignores activities; also it does not reveal an individual's daily T-A in a sequenced manner (e.g., Liu and Karimi 2006; Qiao et al. 2010). The Stops and Moves Model identifies frequent stops and moves from trajectory data, but fails to make a connection and order among them (e.g., Alvares et al. 2007a; Grengs, Wang, and Kostyniuk 2008). In summary, existing individual daily T-A pattern modeling methods fail to combine daily travels and activities in one model and, further, do not make connections between them.

Research Purpose

The need of analyzing and understanding human T-A behaviors is the drive for developing new T-A pattern modeling techniques. Without effective modeling techniques, people's understanding of human T-A patterns may be biased. For example, a widely accepted assumption in transportation geography states that each individual follows a daily routine in conducting his T-A, and the T-A behaviors of a sample on one randomly assigned day is representative of the overall daily T-A patterns of the population (Huff and Hanson 1986; Hanson and Huff 1988; Pas and Sundar 1995; Stopher and Zhang 2011). Most travel surveys, transportation demand modeling, and T-A behavioral analysis have been carried out with the assumption. It has rarely been questioned or examined partially due to a lack of long-term T-A behavioral data from a sufficient sample and space-time modeling techniques to reveal day-to-day variations in individual and collective T-A patterns.

This research proposed effective techniques to model urban population space-time collective activity patterns and individual daily T-A patterns. For collective activity pattern modeling, previous methods lack a full treatment regarding the spatial and temporal dimensions. The space-time collective activity pattern modeling technique developed in this research identified urban areas with dense human activities as urban activity hot spots. Urban dynamic activity patterns were revealed by tracing the development processes of these activity hot spots.

Hot spot detection is one of the techniques in spatial statistics that identifies high density areas of randomly distributed spatial events (Lawson 2010). Traditional local indicators of hot spots such as Local Moran's I, Getis-Ord G, and Geary's C identifies

only the locations of hot spots during a particular time period (Anselin 1995). In another word, the time dimension is not incorporated in these methods. Although researchers can divide the study time period into multiple time intervals and apply these local indicators in each time interval, the changes in hot spot patterns among different time intervals can be seen. However, the connections between patterns at consecutive time intervals cannot easily be traced.

Space-time statistical techniques for revealing spatial-temporal patterns were developed by a few researchers. However, Knox and Bartlett's (1964) contingency table method and the extended K-function method by Diggle et al. (1995) are both global indicators for space-time clustering. They cannot provide information about where and when the hot spots are and how the hot spots evolve. The Space-Time Permutation Model in SaTScan (Kulldorff et al. 2005) extended the search window from a circle to a cylinder, where the time dimension was added. This method reports on where and when the hot spots are, but cannot address how the hot spots evolve through time. GeoDa (Anselin 2005) adapted the Bivariate Spatial Correlation method (Wartenberg 1985) to detect space-time correlations between the observations of the same variable at a center cell and its surrounding cells through two consecutive time periods. However, only two variables can be considered at a time: either the center value at time one and the neighbor value at time two, or the center value at time two and the neighbor value at time one. This method does not provide a whole picture of the hot spot development between two consecutive times. All four variables (including the center value at time one, the neighbor value at time one, the center value at time 2, and the neighbor value at time 2) are needed to assess the development process of hot spots, but the method in GeoDa only considers

two variables at a time. Thus, a new space-time statistical technique needs to be developed to reveal not only the location and time of activity hot spots but also the development process of the hot spots.

The proposed space-time collective activity pattern modeling technique identified the location and time of activity hot spots in a large metropolitan city using Poisson distribution. An activity hot spot was defined as a spatial unit with a significantly large number of activity instances during a certain time period. An activity instance was defined as an activity conducted by one person at a particular location and time. Then the hot spot status of a center unit and its neighboring unit between two consecutive hour intervals were combined to determine the development stage of a hot spot. A total of six development stages were defined for a hot spot: emergence, expansion, stableness, shrinkage, displacement, and decease. These were regarded as a hot spot six-stage life cycle. By tracing the development stages of activity hot spots in an urban area, dynamic collective activity patterns were revealed. Thus, temporal variations (such as hour-to-hour and day-to-day) in collective activity patterns can be assessed.

To demonstrate the effectiveness of the proposed method, GPS trajectory data of 536 taxi cabs over twenty-two days in the city of San Francisco were used to reveal collective activity patterns across the city and over time. Taxi passengers' pick-up and drop-off locations and times were extracted from the trajectories to represent taxi passengers' activity instances. As taxi passengers' activities mainly involve business, tourism, and entertainment types, other types of urban activities were underrepresented, such as grocery shopping, daily commuting, etc. Thus, the activity hot spots detected and the dynamic collective activity patterns discovered in this research as a demonstration of

the proposed method and analysis mainly reveal urban commercial, tourism and entertainment activity patterns at a certain level.

For individual daily T-A pattern modeling, previous methods failed to combine travels and activities in one model or observe the order of travels and activities as they occur during a day. The proposed space-time technique encoded one individual's one-day T-A events in a sequence with location, time, and transportation mode information. By analyzing these daily T-A sequences of an individual with Sequence Alignment Method (SAM), representative daily T-A sequences can be found, which reveals the individual's daily T-A routines.

SAM has been used to analyze human activity sequences since 1995. Usually, one day is divided into a number of episodes (e.g., 48 episodes with 30 minutes each). Activities of an individual conducted during these time episodes are encoded as characters and lined up in a sequence following the chronological order. Activity sequences can be unidimensional or multidimensional. Unidimensional sequences record only one attribute of the activities conducted at different time episodes (for example, activity type). Multidimensional sequences records more than one attribute of the activities (for example, activity type, location, and other information). The character or character combination recording the attribute(s) of an activity conducted at a time episode is an element of an activity sequence. SAM makes the optimal alignment between two sequences by transforming one into the other with the smallest number of character edit operations (including insertion, deletion, and substitution) (Wilson 1998; Kwan, Xiao, and Ding 2014). The similarity of two aligned sequences is the sum of the similarity scores of all aligned pairs of elements minus the cost of all the edit operations in order to

align the two sequences (ClustalTXY). SAM classifies all sequences into groups based on their similarities and generates representative sequences for each group.

Most previous applications of SAM on activity sequences analyzed sequences of multiple individuals. The sequences were classified into groups and each group represented an unique daily activity pattern. Each unique daily activity pattern was then associated with the socio-economic characteristics of the people in the group. The purpose of these studies was to identify differences in daily T-A patterns among different individuals, which was referred as interpersonal variability in T-A behaviors (Pas & Sundar, 1995). Since different individuals rarely share the same absolute activity locations (unless they are from the same household or workplace), previous studies applied different treatments to activity locations. Joh, Arentze, and Timmermans (2005, 2007) used semantic locations (such as home, work, restaurant, etc.) instead of absolute geographical locations. Vanhulsel et al. (2011) used relative movements (angles and arc lengths) instead of absolute geographical locations. Shoval and Isaacson (2007) and Kwan, Xiao, and Ding (2014) divided the study area into polygons. Activity locations were represented by polygon IDs in which the activity occurred. When measuring sequence similarity, a particular integer (e.g., 2) was assigned as similarity score if the pair of aligned activity locations (polygon IDs) were identical; otherwise, a score of 0 was assigned. Thus, activity location similarity was not measured based on the Euclidean distance between activity locations. Wilson (2008) also divided the study area into polygons and used polygon IDs to represent activity locations. Activity location similarity was measured by the Euclidean distance of the two polygon centroids instead of the absolute activity locations in order to simplify distance calculation. In summary,

previous studies using SAM intended to discover interpersonal variability in T-A behaviors. They generally ignored travels between activities and did not measure activity location similarity based on the Euclidean distance between absolute activity locations.

The proposed space-time individual daily T-A pattern modeling technique intended to identify day-to-day differences in T-A behaviors of an individual over a long period of time (e.g., a month, a year, etc.). This was referred as intrapersonal variability (Pas & Sundar, 1995). An individual's one-day T-A events were encoded as characters and lined up in a sequence in the chronological order. These T-A sequences contained information on trip origin, trip destination, trip starting time, trip ending time, and transportation mode. As trip origins and destinations were also activity locations, and the ending time of the previous trip and the starting time of the next trip were also the starting and ending time of the activity conducted in-between the two trips. Thus, both travel and activity events were included in a sequence. For any individual, locations visited at least once a week were usually considered personal anchor activity locations (such as home, work, favorite restaurants, etc.). Thus, an individual usually has a limited number of anchor locations and their geographic coordinates can be calculated as the individual's long-term GPS trajectory data cluster at these locations. When measuring trip origin or destination location similarity, if both aligned locations were anchor locations, a similarity score was calculated based on the Euclidean distance between the two anchor locations; if one of the aligned location was not an anchor location (meaning a random activity location), the lowest similarity score was assigned. Thus, location similarity was measured based on the Euclidean distance between two absolute locations in the proposed technique. Moreover, most previous methods used a nominal measurement

scale/range for sequence element similarity. The proposed technique defined different similarity measurement scales (including nominal, ordinal, and ratio) for different attributes of an element (trip origin, trip destination, trip starting time, trip ending time, and transportation mode), and combined these measurements together to be the element similarity measurement. This allowed element similarity measurement to be more sensitive and accurate. Daily T-A sequences of an individual were constructed and element similarity measurement (including location similarity) were defined following the proposed technique. Sequence alignment were implemented in the ClustalTXY software. Daily T-A sequences were classified into several groups and representative sequences for each group were identified. The representative sequences of each group revealed one daily T-A routine of the individual. Thus, day-to-day variations in individual T-A behaviors can be revealed by comparing his multiple daily routines.

To demonstrate the effectiveness of the proposed method, GPS trajectory data of two individuals living in the northern area of Beijing, China was used to derive daily T-A patterns. As some of the GPS data were not recorded by the participants due to privacy concerns or technical issues, data incompleteness and inconsistency appeared in the dataset, which might have impacted the daily T-A patterns discovered.

Research Questions

This research proposed space-time modeling techniques to reveal temporal variations in collective activity patterns and individual daily T-A patterns. Two overarching research questions were:

1. Can the proposed space-time modeling technique effectively identify temporal variations in spatial collective activity patterns?

2. Can the proposed space-time modeling technique effectively identify temporal variations in individual daily T-A patterns?

As space-time collective activity patterns were revealed by the spatial temporal distributions of urban activity hot spots, there were two sub-questions under the first overarching research question:

1) Can the proposed space-time modeling technique effectively reveal temporal variations in the spatial distribution of urban activity hot spots?

2) To what extent might the future spatial temporal distribution of urban activity hot spots be predicted based on the revealed temporal variations?

To answer these research questions, taxi passengers' activity instances were summarized into a set of spatial units at each time interval. Poisson distribution was applied to every spatial-temporal unit to evaluate whether the unit was an activity hot spot. Thus, activity hot spots were located spatially and temporally. When a hot spot was focused, the status on whether its neighboring zone was a hot spot at the focused time interval, as well as whether the focused spatial unit and its neighboring zone were hot spots at the previous and the next time interval were considered to assess the current and next developing stage of the focal hot spot. By tracing the development process of all activity hot spots in the city, the hour-to-hour variations of the spatial distribution of activity hot spots can be revealed. By comparing the hot spot development processes of the whole city between two single days, a day-to-day variation of the distribution of activity hot spots can be revealed. This process answered the first sub-question.

Based on the initial observation of the data and the revealed hour-to-hour and day-to-day variations of the distribution of urban activity hot spots, an assumption on a weekly repetitive cycle for the spatial temporal distribution of urban activity hot spots was made. Thus, the number of activity instances in a spatial unit during a future time interval was estimated as the historical average number of activity instances in the spatial unit during the same time interval of a day on the same day of a week. Thus, activity hot spots can be predicted for the future time interval. The predictions for consecutive future time intervals can be connected to assess the development stages of the future hot spots. The predictions of the development of future hot spots can be compared with the ground truth data to evaluate prediction accuracy. This process answered the second sub-question.

As an individual's daily T-A patterns were revealed by his representative daily T-A sequences, there were three sub-questions under the second overarching research question:

1) Can the proposed space-time modeling technique effectively identify a number of representative daily T-A sequences of an individual?

2) Whether the proposed space-time modeling technique is sensitive to the dimension of the daily T-A sequences?

3) Whether the proposed space-time modeling technique is sensitive to location similarity measurements?

To answer these research questions, two individuals' long-term T-A behavioral data were encoded into daily T-A sequences with different dimensions. One-dimensional T-A sequences contained only trip origin or destination location in the sequences. Two-

dimensional T-A sequences contained trip origin and destination locations in the sequences. Three-dimensional sequences contained trip origin or destination location, trip starting or ending time, and transportation mode information in the sequences. Five-dimensional sequences contained trip origin location, trip destination location, trip starting time, trip ending time, and transportation mode information in the sequences. One-, two-, three-, and five-dimensional sequences were all analyzed with the proposed method. The daily T-A patterns derived from the one-, two-, three-, and five-dimensional sequence sets were compared. This answered sub-question two. Moreover, three methods for measuring location similarity were implemented in the sequence alignment of an individual's one-, two-, and three-dimensional weekday sequences. The daily T-A patterns discovered using the three location similarity measurements were compared within the one-, two-, and three-dimensional sequence sets. This answered sub-question three. The overall daily T-A patterns discovered for the two individuals answered sub-question one.

Significance of the Research

The contributions of this research to the scholarly research are multi-fold. First, it proposes space-time modeling techniques to identify individual and collective T-A patterns from GPS trajectory data, which will assist transportation geographers analyze individual and collective level long-term T-A behaviors. Second, it defines the typology of a hot spot's life cycle and proposes a method to access the development stage of a hot spot. This adds to the research literature on hot spot detection and evolvement assessment techniques in spatial statistics and will benefit researchers in many fields. Third, this

research calls into question a fundamental assumption in transportation geography that each individual repeats the same T-A events on a daily basis (Huff and Hanson 1986; Hanson and Huff 1988; Stopher and Zhang 2011). Testing of this assumption may have profound implications for the design of transportation surveys as well as the modeling of transportation demand and urban planning. The individual daily T-A pattern modeling may potentially contribute to the shift of spatial modeling and environmental exposure analysis from a place-based to people-based framework, which may be an interest to environmental and health researchers (Kwan 2009; Fang and Lu 2012).

This research also has multiple contributions in professional applications. The space-time modeling of urban population T-A patterns may help transportation researchers and urban planners understand urban T-A behaviors and assist in redesigning transportation surveys, as well as, the modeling of travel demand and urban planning. The dynamic patterns of urban activity hot spots constructed with historical data may be used to retrieve large social, cultural, economic, and political events from the past and will assist in building city-wide profiles of collective activity patterns. This will result in more informed decisions for traffic management, public safety control, emergency response, and other services. Moreover, understanding the repetition and variability of individual's daily T-A patterns may help predict daily T-A behaviors especially towards detecting abnormalities or deviations from daily routines. This will assist law enforcement for monitoring parolees and may help prevent terrorist activities. Furthermore, individual daily T-A pattern modeling may assist in improving individual-level environmental exposure estimation, health risk monitoring, driving risk assessment, etc.

More specifically, collective and individual T-A pattern modeling has great potential in the delivery of Location-Based Services (LBS) by providing predictions of the location and time of activity hot spots in urban areas as well as each individual's near future travel routes and destinations. This will enable more accurate deliveries of Location-Based Services (LBS) to individuals, for example, traffic condition reports, locations of gas stations and restaurants nearby. Relevant and personalized services may be provided based on the understanding of each individual's T-A routines and lifestyles, such as friends and carpool recommendations.

Individual T-A pattern modeling may be very important for public safety management and criminal justice analysis (Rossmo, Lu, and Fang 2012). Since September 11, 2001, whether terrorist attacks might have been prevented and how to increase security agencies' levels of awareness and knowledge of terrorist activities became very important questions for government officials to answer (Seifert 2004). Thus, identifying and tracking individual terrorists became one of the key objectives of many homeland security initiatives (Seifert 2004). Significant efforts have been made regarding how to better collect and analyze terrorists' traveling and activity information toward providing insightful, carefully constructed, and predictive consequences (Seifert 2004). Terrorists' traveling trajectories may be obtained through their cell phone signals, thus, their T-A patterns may be modeled and used to monitor their movements. Any deviation from their daily routines may be an indicator of abnormal events, which deserves serious attention from the security agencies. These are examples of analyses that may help to prevent and fight against terrorist activities.

Individual T-A pattern modeling may also be useful for customer-centric marketing. It is becoming increasingly apparent that marketers are considering customer-level information when they generate a marketing strategy for a business. Customers are now demanding personalization and customization of products and services that help them feel unique (Bianco 2004). Customer-centric marketing considers the needs, wants, and other information of customers as the starting point of the marketing process (Cheng and Dogan 2008). Businesses that start their marketing process with the needs of customers are better adapt to various market scenarios than their competitors and are able to manage the supply side rather than the demand side of the marketing process (Sheth, Sisodia, and Sharma 2000). Thus, knowing their customers is essential to the whole marketing process. Traditional customer information contains basic demographic and economic information that may be collected from credit cards, online registration, etc. This information is helpful but limited. Daily T-A patterns of customers may provide more insight on customers' lifestyles, social roles, and their needs. Products or services may be better personalized and customized according to this information. Businesses knowing what a customer is most likely to buy has a significant advantage over the competition (Kumar and Petersen 2005).

Dissertation Organization

The remainder of this dissertation is organized as follows. Chapter Two discusses the theoretical framework including the limitations and opportunities present in the T-A data source, the previous T-A modeling techniques, and the connections between collective and individual level T-A patterns. Chapter Three reviews previous studies on

the analysis and modeling of collective and individual T-A patterns. Chapter Four introduces the proposed collective activity pattern modeling techniques and the empirical data, and reports on the findings and implications. Chapter Five thoroughly describes the proposed framework for the individual daily T-A pattern modeling and the case study data, and discusses results on the pattern discovery and sensitivity analysis. Chapter Six summarizes the findings of this research and recognizes the limitations in the data sources and the modeling procedures, as well as makes suggestions for future studies.

II. THEORETICAL FRAMEWORK

The spatial tradition of geography (Pattison 1964) makes the geography discipline different from other social sciences. Movement is considered one of the true essentials of the spatial tradition (Pattison 1964). The study of movement emerged into a separate subfield of research in geography, transportation geography, when Ullman summarized research topics in transportation geography in American Geography: Inventory & Prospect in 1954 (Ullman 1954). While traditional research topics in transportation geography remains active for the past few decades, new research topics gradually add on as the development of societies, economies, and technologies. Table 1 lists traditional and contemporary research topics in transportation geography (Ullman 1954; Knowles 1993; Taaffe and Gauthier 1994; Graham 1999; Preston 2001; Goetz et al. 2003; Black 2004; Shaw 2006).

Table 1. Traditional and contemporary research topics in transportation geography.

Geographic pattern of transportation systems
Transportation demand modeling
Spatial interaction modeling and the gravity models
Accessibility analysis
T-A behavioral analysis
Location analysis
Allocation modeling
Shortest path analysis
Flow analysis
Network analysis
Transportation policies
Transportation and economic development
Sustainable transportation

T-A behavioral analyses have received more attentions since the 1990s (Taaffe and Gauthier 1994; Graham 1999; Goetz et al. 2003). As individuals participate in daily activities at different times and different locations, their daily T-A patterns would be very complex (Hanson and Hanson 1981). Modeling and analysis of urban population daily T-A patterns at the individual and collective levels has been a research challenge for transportation geographers.

Limitations and Opportunities by the Data Source

Since the late 1970s, the primary data source for T-A pattern modeling has been household travel surveys conducted via mail, telephone, face to face interview, or internet (Stopher and Greaves 2007). These surveys usually ask the participants to report on all the trips and activities they conducted in one randomly selected day. More specifically, for each trip, the origin and destination addresses, the starting and ending time of the trip, travel distance, and travel mode information are generally recorded. For each activity, the type, location, starting, and ending time of the activity are recorded in the survey. This self-reporting data collection approach has a few disadvantages. First, precise location, time, travel speed, travel route, and other information are generally not reported. Second, it puts heavy burden on the survey participants to record their T-A details while engaging in daily activities. It also requires a lot of labor work and long processing time for implementing the survey and data input. Thus, it is not suitable for collecting data over a long period of time or among a large population. Third, the T-A data collected tends to be less accurate and reliable as participants might have misreported, over-reported, or under-reported their trips or activities. These disadvantages have made the household travel

survey data very limited in modeling T-A patterns with precise spatial and temporal information, and in revealing pattern changes over time.

Recent developments in pervasive location acquisition technologies enabled the collection of a new type of data, trajectory data, for T-A studies (Lu and Liu 2012). Trajectory data records movement histories of objects. It contains sequences of precise location information and time stamps of moving objects. Trajectory data may be represented by

$$T = \{(x_1, y_1, t_1), (x_2, y_2, t_2), \dots, (x_n, y_n, t_n)\} \quad (1)$$

$(x_i, y_i), i \in 1,2, \dots, n,$ is the coordinate of the moving object at time t_i. The difference between t_i and t_{i+1} typically ranges from one second to one minute. T-A information may be extracted from this type of data and used to model T-A patterns at the individual and collective levels.

Trajectory data can be collected continuously through the Global Positioning System (GPS) in almost all environmental conditions. Humans, vehicles, animals, and products may be equipped with a GPS device so that their trajectory data will be automatically recorded in the device. GPS devices are wearable, small, and light weight. It has substantial storage capacity of several Gigabytes of data. Its battery can be easily recharged. GPS devices do suffer a few problems of data loss, such as signal loss or degradation in urban canyons and tunnels, longer warm-up time, device failure, dead battery, etc. (Krenn et al. 2011). However, compared to household travel surveys, GPS trajectory data collection has apparent advantages. Precise location, time, travel speed, travel route, points of interest (e.g., school, work, restaurant, etc.), and other information can be easily collected for a longer period of time (e.g., a couple of weeks, months, or

even years). While it provides more accurate and reliable data, it requires much less work from the participants and survey implementation and data input technicians. Besides, the recent integration of the GPS function in mobile phones provides the possibility of trajectory data collection through mobile phones. This could greatly enlarge the pool of potential participants and reduce the cost of data collection.

Researchers have demonstrated the capacity of GPS devices to provide detailed, accurate, and reliable T-A data (Murakami and Wagner 1999; Wolf et al. 1999; Yalamanchili et al. 1998; Stopher, FitzGerald, and Zhang 2008; Wiehe et al. 2008). For example, Murakami and Wagner (1999) compared data accuracy between the self-reporting household travel survey data collection method and the GPS trajectory data collection method in a field test in Lexington, Kentucky in the fall of 1996. They found that self-reported distances and travel times were longer than GPS recorded distances and travel times. Wolf et al. (1999) evaluated GPS positioning accuracy and concluded that GPS data collection method could provide sufficient accuracy levels for movement data. Yalamanchili et al. (1998) demonstrated that GPS devices could capture short and intermediate stops that occurred within chained T-A better than the self-reporting travel surveys. Bohte and Maat (2009), Greaves et al. (2010), and Oliveira et al. (2011) further used an internet based prompted recall method to let the participants verify information extracted from GPS trajectory data and add additional information related to their travels and activities (such as activity types). Results proved that highly complete and accurate T-A data could be collected for a long period of time with reasonable burden on the participants. Bohte and Maat (2009) even claimed that the traditional household travel survey data collection method might be entirely replaced by methods using GPS, GIS,

internet, and related technologies in the near future. GPS trajectory data has provided a promising opportunity for space-time modeling of urban population T-A patterns.

Collective Activity Pattern Modeling

GPS trajectory data can provide precise activity location and time information. Citywide activity density surfaces or activity hot spots at selected time intervals may be used to reveal collective activity patterns at those time periods (e.g., Kwan 2000; Chen et al. 2011). However, the connections between the patterns at different time intervals cannot easily be traced. The failure of the existing studies to model collective level space-time activity patterns is originated from a lack of effective space-time modeling techniques. Traditional spatial statistics were commonly used to reveal spatial patterns at static snapshots, such as Ripley's K-function (Ripley 1976) and LISA (Anselin 1995).

Space-time statistical techniques for revealing spatial-temporal patterns were developed by a few researchers. Knox and Bartlett (1964) proposed a test using a 2x2 contingency table to detect space-time clustering in the distribution of epidemic events. This was the first effort to detect space-time clusters in point events. Diggle et al. (1995) combined the spatial and temporal dimensions in Ripley's K-function. However, both the contingency table method and the extended K-function method are global indicators for space-time clustering. They cannot provide information about where and when the clusters are and how the clusters evolve.

By extending the search window from a circle to a cylinder, SaTScan (Kulldorff 2010) added the time dimension to its clustering analysis. The Space-Time Permutation Model in SaTScan (Kulldorff et al. 2005) defined a cluster as a spatial unit with a higher

proportion of events than the rest of spatial units during a specific time period. This method in SaTScan reports on where and when the clusters are, but cannot address how the clusters evolve through time.

Bivariate spatial correlation (Wartenberg 1985) was originally used to detect spatial association between one variable at the center location and another variable at the surrounding locations on a lattice grid. In GeoDa (Anselin 2005), this method was adapted to detect space-time correlation between the observations of the same variable at a center cell and its surrounding cells through two consecutive time periods. The space-time correlation analysis may be conducted between time 1 at a center cell and time 2 at its surrounding neighbors, or between time 2 at a center cell and time 1 at its neighbors. This space-time correlation method from GeoDa does not consider the development process of the clustering patterns through time. A new spatial-temporal statistical technique needs to be developed, in order to reveal not only the location and time of activity hot spots but also the development process of these hot spots.

Individual Daily Activity Pattern Modeling

An individual's activities conducted during different time periods in a day may be encoded into an activity sequence following the chronological order. Usually, one day is divided into a certain number of episodes (e.g., 48 episodes with 30 minutes each). Then, activities conducted in these episodes are lined up in a sequence chronologically. Sequence Alignment Method (SAM) may be used to analyze an individual's daily sequences. SAM was originally developed for comparing DNA, RNA, or protein sequences. Through identifying regions of similarity between two sequences, functional,

structural, or evolutionary relationship may be revealed. SAM was introduced to social science by Abbott for career pattern analysis (Abbott 1995). Since then, several researchers have applied SAM to analyze human activity sequences.

When the element in a sequence contains only one attribute, for example, the type of the activity, Unidimensional Sequence Alignment Method (UDSAM) may be used to measure sequence similarity and discover representative sequences. The smallest number of operations (including insertion, deletion, and substitution) required to equalize two sequences was used to calculate their level of similarity (Wilson 1998). This method is good for identifying daily activity patterns concerning the type and order of activities. However, the locations of activities are not considered.

When more information is added to an activity sequence, such as activity type, location (e.g., home, work, etc.), and travel mode, each element in the sequence includes multiple attributes and the sequence becomes multidimensional. Thus a Multidimensional Sequence Alignment Method (MDSAM) needs to be used to analyze these sequences. Joh and colleagues (Joh, Arentze, and Timmermans 2001, 2005, and 2007; Joh et al. 2002) proposed a MDSAM to measure similarities between these sequences. The method would find the least cost alignment for each dimension/attribute first, then integrate the alignments of all dimensions, and finally calculate the cost for the integrated alignment. A hierarchical clustering algorithm was then applied to group these multidimensional sequences. Each cluster would represent a typical daily activity pattern. Joh and colleagues' MDSAM method captures multiple attributes of activities while the UDSAM does not. However, activity locations included in the attributes are semantic locations. Geographic locations are not incorporated in the method.

Wilson (2008) also proposed a MDSAM and developed the ClustalTXY software to implement it. The method analyzed two-dimensional sequences. Each element in the sequence had two attributes: activity type and location. The similarity between two corresponding activity elements was a combination of activity type similarity and activity location similarity. Activity location similarity was measured by the Euclidean distance between the centroid of the two zones in which the two activity locations were (in order to simplify the distance calculation). The similarity between two activity sequences was the sum of the similarity scores for all aligned pairs of elements minus the alignment operation cost in order to equalize the two sequences. Wilson (2008) further defined four types of representative sequences to reveal activity patterns. A "consensus sequence" is a generated sequence that has the most frequent element from the sequence set at each position. It resembles the mode of a univariate distribution. A "modal sequence" is a sequence from the sequence set that has the minimum difference to the consensus sequence. A "median sequence" is a sequence from the sequence set that has the minimum sum of differences (maximum similarity) from itself to all other individual sequences, which resembles the median of a univariate distribution. An "average sequence" is the sequence from the sequence set that has the minimum sum of squared distances from itself to all other sequences. It resembles the mean of a univariate distribution. The ClustalTXY software can identify these four types of representative sequences after calculating similarity scores between activity sequences (ClustalTXY). Wilson's method analyzes activity sequences with multiple attributes. Location similarity is part of the total similarity measurement between two activity sequences, although it is

measured by zone distance instead of the absolute distance between two activity locations.

Wilson's MDSAM is powerful for revealing daily activity patterns as it captures activity type, order, time, duration, and location information. It works for up to 10-dimensional sequences (meaning that an element can have as many as ten attributes). It also provides the opportunity for the users to define their own element similarity measurement. When activity type, location, and other attributes are included in each element, the element similarity score may be defined as a combination of similarity scores from each attribute (including how each attribute is weighed against each other). Euclidean distance between absolute activity locations may be incorporated in calculating similarity scores of the location attribute, which then may be included in the calculation of element similarity scores. Thus, location similarity between sequences may be measured. In summary, Wilson's MDSAM has many advantages for analyzing sequenced events. If an individual's daily travels and activities can all be lined up in a sequence in a way, then Wilson's method may be used to reveal individuals' daily T-A patterns.

Connections between Collective and Individual T-A Patterns

In transportation geography, collective T-A pattern modeling and analysis has been built upon the understanding of the individual T-A behaviors. That is individuals' T-A behaviors are highly repetitive from day to day (Huff and Hanson 1986; Hanson and Huff 1988; Stopher and Zhang 2011). In another word, each individual follows a daily routine in conducting his T-A. This assumption has granted that one day is adequate in capturing an individual's daily T-A behaviors. Further, the T-A behaviors of a randomly

chosen sample from a population reported on a randomly chosen day out of some longer time period constitute an unbiased sample of the population over that time period (Pas and Sundar 1995). This randomly chosen day is most often a single weekday. Thus the resulting data is representative of the T-A behaviors of the population on an average weekday. This assumption has been widely applied in the data collection and modeling procedures in transportation demand modeling, urban and regional collective T-A pattern analysis, and transportation planning. Most nationwide and regional travel surveys have been carried out with the assumption, including the National Household Travel Survey (NHTS) conducted by the U.S. Department of Transportation, The Sydney Household Travel Survey conducted by the Australian Bureau of Transport Statistics, The Mobility in Germany Survey conducted by Germany Federal Ministry of Transport and Digital Infrastructure, etc. The analysis and modeling of urban T-A patterns has been based on a single day record of each individual in the sample (e.g., Recker, McNally, and Root 1985; Wilson 1998; Collia, Sharp, and Giesbrecht 2003; Gutierrez and Garcia-Palomares 2007; Chen et al. 2011). These studies sought to reveal the differences in T-A behaviors between people, which was referred as interpersonal variability by Pas and Sundar (1995). Those differences were then related to differences in the characteristics of the individuals or households.

It has rarely been questioned whether each individual follows a daily routine in conducting their T-A, and whether the T-A behaviors on one randomly assigned day from a sample is representative of the overall daily T-A patterns of the population. Only a few studies (e.g., Huff and Hanson 1986; Kitamura and Van Der Hoorn 1987; Pas and Koppelman 1987; Hanson and Huff 1988; Pas 1988; Buliung, Roorda, and Remmel

2008; Stopher and Zhang 2011) examined and discussed the validity and reliability of this assumption, partially due to the limitation of the household travel survey data and a lack of space-time modeling techniques to reveal day-to-day variations in T-A patterns. The space-time modeling techniques of individual and collective T-A patterns proposed in this research provide means for testing the repetitiveness assumption in the future once long-term T-A behavioral data from a sufficient sample became available. Further, the relationship between individual and collective T-A patterns may also be explored. Day-to-day variations in individual T-A behaviors may result in day-to-day variations in collective T-A patterns. An examination of the nature of individual day-to-day variations will help explain how collective T-A patterns are different from day to day. As stated by Pas and Sundar (1995), individual day-to-day variations in T-A behaviors were either systematic or random. Systematic variations may be related to the day of the week or other processes, such as even or odd dates, etc. The discovery of the systematic variations in individual daily T-A behaviors may help explain the systematic variations in collective T-A behaviors. On the other hand, day-to-day variations in collective T-A patterns may somewhat reflect the individual day-to-day variations in daily T-A behaviors of the majority people. However, the nature of the day-to-day variations (systematic, random, or both) in collective T-A patterns does not automatically imply the same mechanism in individual level day-to-day variations in T-A behaviors. These are research concerns to be addressed in the future.

III. LITERATURE REVIEW

Urban population T-A patterns have been modeled at both the individual level and collective levels. This chapter reviews the existing modeling techniques for collective and individual T-A patterns.

Space-Time Modeling of Collective T-A Patterns

Five main strategies have been developed to model collective travel and/or activity patterns. Statistics on travels and activities for groups of people were calculated and associated with socioeconomic characteristics. Individuals' daily T-A were transformed into two- or three-dimensional graphs. The geometric similarity of these graphs was used for clustering and identifying representative patterns. Individuals' daily activities were also transformed into letter sequences and SAM was used to calculate sequence similarity and identify representative activity patterns. Moreover, GPS trajectories of vehicles in urban areas were clustered to reveal citywide travel patterns and traffic flows. Activity density surfaces were also constructed to reveal citywide activity patterns.

Descriptive Statistical Methods

Early T-A studies modeled urban population collective T-A patterns by simple statistics, such as daily average number of trips, travel distance, travel time, time spent on different types of activities, etc. These statistics were often summarized by socioeconomic characteristics (e.g., age, gender, employment status, etc.), in order to find the association between the generalized T-A patterns and participants' socioeconomic

characteristics. For example, Van Der Hoorn (1979) summarized the average number of trips and travel time per day by car ownership, rural-urban areas, and socio-demographic status (such as working men, working women, school children, etc.). To reveal general activity patterns, he calculated the average duration and frequency of conducting each type of activities (such as housekeeping, work, travel to/from work, shopping, etc.) by socio-demographic status during the surveyed week. Collia, Sharp, and Giesbrecht (2003) compared daily travel patterns between the elderly (age 65 and above) and the working force (age 19-64). They found that elderly people especially females took fewer trips, traveled shorter distances, and had shorter travel times daily than the working force. Moreover, elderly people were more likely to suffer from medical conditions that limit their travel. Ding, Lu, and Zhang (2016) analyzed individuals' activity time use between vehicle usage rationed days and non-rationed days, using data from a seven-day GPS survey in Beijing, China. Vehicle usage rationing is a transportation management policy which restricts the use of vehicles in one weekday of a week depending on the last digit of the vehicles' license plate (e.g., vehicles with two or seven as the last digit of their license plate are banned on Tuesdays). Results showed that the average amount of time spent on and the participation frequency in maintenance and discretionary activities per person per day had a significant decrease on rationed days. The descriptive statistical methods are very commonly used to reveal urban collective T-A patterns. However, precise location and time information are generally not incorporated in the statistics, nor does it reveal any changes of the patterns over time.

Space-Time Graph

Some researchers composed a two- or three-dimensional space-time graph to represent one individual's T-A in one day. The space-time graphs from different individuals were then classified into several groups based on their geometric similarity. Representative graphs for each group were identified to reveal typical T-A patterns of the people in each group. Recker and colleagues proposed a two-dimensional graph model to describe individuals' one-day T-A behaviors (Recker, McNally, and Root 1985). Times of activities, activity types, distances from home, and activity durations were captured in the graph model (see Figure 1 in Recker, McNally, and Root 1985). Participants' one-day T-A graphs were then classified and a representative pattern was identified for each group. These T-A patterns were associated with the socioeconomic characteristics of the participants in each group. Chen et al. (2011) used a conceptual framework from Time Geography, space-time paths, to represent individuals' one-day T-A events. Six clusters of these paths were identified based on travel distance, number of trips, and activity durations. Each cluster revealed a typical daily T-A pattern for the group of participants. The space-time graph method does not incorporate absolute activity locations in the model, nor does it reveal changes of the typical patterns over time.

Sequence Alignment Method

Previous studies mainly used the Sequence Alignment Method (SAM) to analyze daily activity sequences from a group of individuals. Typical daily activity patterns were found accounting for the order, duration, and transition of activities. Wilson (1998) analyzed daily activity sequences of women aged 65 or older using the UDSAM. Elderly

women's daily general activity routines were revealed. Shoval and Isaacson (2007) and Shoval et al. (2015) extracted activity locations from the one-day GPS trajectory data of a group of tourists. For each tourist, activity locations were lined up in a sequence following the order in which they were visited. All tourists' activity-location sequences were analyzed using the UDSAM and a number of distinctive visiting patterns were identified. The UDSAM can discover activity patterns with only one attribute. Other attributes of activities are left out.

Joh, Arentze, and Timmermans (2007) discovered frequent daily activity sub-patterns (such as "at home task - at home leisure" and "work - at home task") from a sample of participants using a MDSAM. These sub-patterns contained activity type, location (semantic), and travel mode information. Joh, Arentze, and Timmermans (2005) used a similar MDSAM to calculate similarities between 6950 three-dimensional activity sequences (activity type, location, and travel mode) obtained from a sample of individuals from some selected neighborhoods in the Amsterdam-Utrecht corridor, the Netherlands. An interactive hierarchical clustering method was applied and seven clusters were identified. For each cluster of activity sequences, the characteristics of the activity pattern (including the average relative frequency of each type of activity and the average number of episodes associated with each location and transport mode) and the socioeconomic characteristics (including gender, age, income, car ownership, with or without children, weekday or weekend) of the participants were summarized and associated. Wilson (2008) also used a MDSAM to analyze activity sequences of a random sample of participants from the 1972 Reading Urban Survey. These activity sequences had two dimensions: activity type and location. The proposed MDSAM

classified the sequences into three groups and identified four types of representative sequences (consensus sequences, modal sequences, median sequences, and average sequences) for each group. General activity patterns were described for each group based on the alignment of the sequences and the representative sequences. Vanhulsel et al. (2011) constructed multidimensional activity sequences using relative movements instead of absolute geographical locations. The proposed method transformed and normalized geographical locations of activities into angles and arc lengths to reflect the relative movements made within each sequence. Similarities of this relative movement dimension and other dimensions (such as activity type) between two sequences were calculated using Wilson's (2008) MDSAM. This method identified activity sequences which were similar in relative geographical movements rather than in absolute geographical locations. Kwan, Xiao, and Ding (2014) proposed the Multi-Objective Evolutionary Algorithm (MOEA) to assess similarities among multidimensional activity sequences and to classify them into groups with distinctive activity patterns. The experiments demonstrated that the proposed method outperformed ClustalG software.

MDSAM discovers collective activity patterns with multiple attributes. One concern associated with activity locations in collective activity pattern modeling is that the randomly selected participants do not share the same activity locations. This issue was handled differently among the above methods. Joh, Arentze, and Timmermans (2005, 2007) used semantic locations instead of geographical locations. Wilson (2008) used zone centroids to represent absolute activity locations in order to simplify distance calculation. Vanhulsel et al. (2011) used relative movements instead of absolute geographical locations. In general, SAM is more suitable for discovering daily activity

routines and associating them with socioeconomic characteristics. It does not reveal the citywide activity landscapes, nor the changes of the landscapes over time.

Trajectory Clusters

As precise travel routes are revealed by the GPS trajectory data, many studies modeled collective travel patterns by grouping similar trajectories. A group of trajectories with spatial, temporal, and/or attribute similarity were considered a trajectory cluster. Trajectory clusters showed distinct collective travel patterns or major traffic flows during particular times in the past. Trajectories may be grouped together by sharing the same origins and destinations. Andrienko and Andrienko (2008) aggregated trajectories with origins and destinations in the same city grid to reveal major traffic flows.

Trajectories may also be grouped together by passing through the same road sections or city grids (Brakatsoulas, Pfoser, and Tryfona 2004; De Almeida and Guting 2005; Mouza and Rigaux 2005; Pfoser and Jensen 2005; Li and Lin 2006; Abraham and Lal 2010; Roh et al. 2011). These studies first indexed trajectories with a sequence of road segments or city grids. Then the trajectories that shared the same sequence or subsequence of road segments or city grids were grouped as a cluster. This method reduced the dimensionality of trajectories and made similarity measurement between trajectories easy to conduct.

More often, trajectories are grouped together based on geometric similarity measurement, including distance in length, direction, speed, time, space, etc. Several studies developed algorithms to measure distances among trajectories and to identify trajectory clusters, such as Andrienko and Andrienko (2008), Lin and Su (2008), Gao et

al. (2010), Li et al. (2010), Giannotti et al. (2011), Dodge, Laube, and Weibel (2012), and Pelekis et al. (2012). A few studies partitioned trajectories into subsections and calculated distances among subsections considering location, length, direction, and speed. Similar subsections were clustered based on the calculated distance matrix. The mean center of each cluster was used to present general patterns (Nanni and Pedreschi 2006; Lee, Han, and Whang 2007; Buchin et al. 2011; Wu et al. 2013). Some studies projected trajectories onto road networks. Similarities between trajectories were measured by the spatial and temporal distances between paired road segments or paired nodes in the network (Hwang, Kang, and Li 2005; Chang et al. 2007; Tiakas et al. 2009; Roh and Hwang 2010). Elnekave, Last, and Maimon (2007) and Zhao and Xu (2011) both represented trajectories as a sequence of Minimal Bounding Boxes (MBB). Similarity between trajectories was defined as the sum of similarities between corresponding MBBs, including differences in space, time, duration, and the density of data points within the MBBs.

Trajectory clusters reveal citywide collective travel patterns and major traffic flows at selected time intervals. However, how these clusters evolve over time had never been explored. Zhou et al. (2015) developed a different method to uncover collective travel patterns. They identified critical intersections from an urban transportation network using taxis' travel trajectories. The spatial and temporal variation patterns of these critical intersections were also revealed. The network of critical intersections uncovered urban population space-time traveling patterns.

Activity Density Surfaces

With the advancements in GIS analysis techniques (e.g., Geocoding), activity locations of travel survey participants may be displayed on a map. Collective activity patterns may be revealed by activity density surfaces at selected time intervals, or by the time series of activity events of the spatial units in the study area. For example, Kwan (2000) used an activity density surface to reveal urban population activity patterns across the urban area at a selected time interval. Chen et al. (2011) constructed activity density surfaces at different times of a day for selected neighborhoods. Jiang, Ferreira, and Gonzalez (2012) displayed the distribution of different categories of activities at different times of a day over the Chicago metropolitan area. Shoval (2008) mapped the overall density of tourists' activities on a regular grid (10m*10m) across Akko's Old City. He also visualized the total amount of time the tourists spent at different parts of the city. Yue et al. (2009) identified urban areas with dense human activities at selected time periods of a day by analyzing taxi pick-up and drop-off locations. Pan et al. (2013) established temporal profiles of activity events for selected city blocks. They first identified city blocks with high density of taxi pick-up and drop-off events. Then, for each block, they calculated the average number of pick-up and drop-off events during each hour of a day for weekdays and holidays. These time series were used later for land use classification and land use change detection.

Activity density surfaces method incorporates precise location information in the collective activity pattern modeling. However, those activity density surfaces are static snapshots of collective activity patterns. Whether and how the pattern changes over time is not in the consideration.

Space-Time Modeling of Individual T-A Patterns

Four main approaches have been used for modeling individual travel and/or activity patterns. Some studies represented an individual's one-day activities by a letter sequence and used the SAM to analyze the individual's daily sequences to identify his daily activity patterns. Some studies extracted anchor locations of an individual and constructed a probabilistic model using his historical visiting and transitioning records at those locations. Some studies revealed an individual's frequently traveled routes and established a network model to predict future routes and destinations. Other studies identified frequent stops and moves from GPS trajectories and inferred activity types and travel modes for those stops and moves.

Daily Activity Sequence

One individual's one-day activities may be lined up in a sequence following the chronological order. The type, time, and location of those activities are often included in these sequences as attributes. Similarities between an individual's daily activity sequences may be measured and representative sequences may be found to reveal the individual's daily typical activity patterns. This method has been used to model individuals' daily activity patterns since the 1980s. Before Sequence Alignment Method (SAM) was introduced to study human T-A behaviors, researchers defined their own sequences and developed their own algorithms to calculate sequence similarity.

Pas (1983) represented one individual's one-day activities as a sequence of activity stops. Each stop contained information about the type of the activity and the time when it was performed. Similarity measurement between two corresponding activity

stops and two activity sequences were defined. An agglomerative hierarchical clustering method was used to group individual's daily activity sequences and a small number of clusters were identified. The activity sequence closest to the cluster centroid was defined as the representative sequence. The individual's daily activity patterns were revealed by the representative sequences containing information on the number, type, time, and order of activities. Location of activities and travels between activities were not considered.

Hanson and Huff (Huff and Hanson 1986; Hanson and Huff 1988) also represented one individual's one-day activities as a sequence of activity stops with some attribute information, including activity type, travel mode, time of arrival, and the location zone of the activity. For each individual, they defined "the most representative day" as the single day during which the activity sequence was the most similar to the activity sequences of the other days. In the empirical study, they identified the five most representative days for each participant over a five-week period. For the rest of the 30 days, each day was grouped with one of the five representative days based on similarity. The results of the empirical study showed that: 1) individuals did not simply repeat the same T-A pattern every day, nor did they conduct completely random T-A; 2) each individual had more than one typical daily pattern, and they were fundamentally different from each other; 3) the most representative daily pattern was not adequate to describe the individual's daily T-A behaviors over the five-week period; 4) even the five most representative daily patterns could not fully describe the individual's daily T-A behaviors over the survey period, as considerable variability was not accounted for; 5) no one weekday was more representative than other days; 6) weekend days were less likely to be an individual's most representative day than weekdays, but they appeared to be the

second to the fifth most representative days frequently; 7) for many individuals, none-T-A day was a typical daily pattern. Thus the authors concluded that there were both repetition and variability in individuals' daily T-A behaviors. They suggested that the data collection period for T-A behavioral studies should be long enough to capture at least three most representative days. This was one of the fundamental studies for individual daily T-A behaviors. However, absolute activity locations were not incorporated in the modeling, nor were the travels between activities a focus.

Pas (1988) described a daily pattern by the number of stops outside home, the type of activity at each stop, the time when each stop occurred, and the distance of each stop from home. He further described a weekly pattern by the frequencies of the five most frequent daily patterns, which reflected the lifestyles of individuals. A dataset that contained five-day T-A data of 112 employed people in Reading, England in 1973 was used for the empirical study. The five most typical daily patterns and weekly patterns were identified. The results showed that the five most typical daily patterns were independent of the day of the week, meaning that the author did not find any systematic variability across the days of the week. Furthermore, all five typical weekly patterns contained at least two different typical daily patterns, meaning that there was day-to-day variability in individuals' T-A behaviors within the five-day period.

Stopher and Zhang (2011) defined several typical daily T-A patterns, such as home-work-home, home-shopping-home, home-work-shopping-home, etc. after examining a GPS trajectory dataset which contained volunteers' seven- to fifteen-day trajectory data. The authors found that each volunteer's T-A behaviors over the survey period comprised a number of typical daily patterns and little repetition was present in the

volunteers' T-A behaviors over the study period. For each volunteer in the dataset, even for the daily pattern with several repetitions, very few were similar on all four attributes: total travel distance, total travel time, the start and end time of activities, and total activity duration.

Lv, Chen, and Chen (2013) created a matrix/sequence to describe an individual's one-day activities. One day was split into 24 time intervals (one hour each) as the columns in the matrix; in each column, the time spent (in minutes) at each anchor location is recorded (see Figure 2(a) in Lv, Chen, and Chen 2013). If the sum of the staying time at all anchor locations in each column is less than 60 minutes, then the remaining time is allocated to travel and assigned to "on the way". They then defined a similarity measure to compare two one-day activity matrices. For the corresponding columns in two matrices, the cosine coefficient was used to calculate their similarity. The overall similarity between two one-day activity matrices is the average value of the similarities between corresponding columns. They then used the bottom-up agglomerative clustering algorithm to group the one-day activity matrices. The algorithm begins with treating each one-day activity matrix as a cluster. In each iteration, the similarity between each pair of clusters are calculated. The two clusters with the maximum similarity are merged into a new cluster. The iteration stops when there is maximum similarity within clusters and minimum similarity between different clusters. For each cluster, a representative one-day activity matrix was calculated (see Figure 2(b) in Lv, Chen, and Chen 2013). Each entry in the matrix represents the probability of staying at that location during the specific time interval (see Figure 2(b) in Lv, Chen, and Chen 2013). Of all the above studies, activity type, the time, and order of activities are

incorporated in the description of daily T-A patterns, but the location of activities and the travel between activities are not in the consideration.

Activity Location Modeling

Some studies model an individual's visiting pattern of anchor locations to reveal the individual's T-A patterns. Historical visiting records of anchor locations, or historical transitioning records between anchor locations were used to predict the next activity location where the individual is most likely to be.

Ashbrook and Starner (2003) derived all trips from one individual's GPS trajectory data and represented each trip with its origin and destination location ID. Frequencies of conducting these trips were used to generate a Markov Model describing the transition probabilities between two locations. This model can be used to predict the individual's most possible destination given his current location. Hariharan and Toyama (2004) extracted the top five most frequently visited locations and frequently traveled trips from individuals' trajectory data. Hidden Markov Model (HMM) was used to describe transition probabilities between locations. Different from the above study, temporal information was incorporated in the HMM, meaning that the transition probabilities between two locations are conditioned on time intervals. Activity location transition probability models reveal an individual' general T-A patterns in a unique way. However, it cannot reveal the sequence of T-A occurred in one day. Thus, it is not suitable for modeling daily T-A patterns.

Scellato et al. (2011) established activity profiles at anchor locations for an individual by recording the arrival and staying time of each previous activity at each

anchor location from the individual's GPS trajectory data. The arrival and staying time of the next activity at a anchor location is calculated based on the activity profile at this location. Based on this method, all future activities at all anchor locations can be predicted for the individual. This activity profile method focuses on single activities at specific times of a day. Information on travels between activities and the order of conducting activities on a daily basis is not revealed.

Travel Route Modeling

Frequently traveled routes can be identified from an individual's historical trajectories. Traveling frequencies of these routes are used to generate a network model to reveal the individual's travel patterns and to predict his future travel path and destinations.

Liu and Karimi (2006) and Qiao et al. (2010) each summarized an individual's historical trajectory data and calculated the probabilities of turning into each road segment at each intersection using the continuous Time Bayesian Networks. The predicted future route of the individual is the route with the highest probability. Jeung et al. (2010) also developed a network mobility model that captures the turning probabilities at road intersections and the average travel speed on road segments based on mobility statistics from an individual's historical trajectories. The maximum likelihood travel route and destination of the individual can be predicted.

Kim et al. (2007) used a similar method to predict the most possible travel route of an individual given his current travelled trajectory, proposed destination, and his historical trajectory database. The method first searches for candidate trajectories in the

database whose sub-trajectory matches the current travelled trajectory and shares the same destination. Then it groups these candidate trajectories based on their similarities and count the frequencies. The most possible travel route of the individual between his current location and the proposed destination follows the route of the trajectories with the highest frequency (see Figure 2 in Kim et al. 2007).

Alvarez-Garcia et al. (2010) generated a HMM for path and destination prediction. They first extracted support points for all trajectory crossings in the database. When two trips cross at an intersection, two support points were created for each trip: one before and one after the crossing along each of the two trajectories. After support points were extracted for all trajectory crossings, each trip was represented and simplified by a sequence of support points and two trip end points. A HMM was generated on these support points to describe the probability of reaching each destination at each support point using an individual's historical GPS data.

Vu, Ryu, and Park (2009) indexed trajectories with sequences of grids that they pass through. One individual's historical trajectories passing the same sequence of grids at the same time period were identified and considered to represent a frequent movement pattern. Future travel and destination can be predicted if the individual's current trajectory matches the sub-trajectories of a frequent movement pattern.

Sadahiro, Lay, and Kobayashi (2013) indexed trajectories with directed and ordered road segments in a road network. Primary routes are defined as frequently visited connected sets of road segments. Primary routes were extracted from an individual's trajectory data collected over two years (see Figure 7 in Sadahiro, Lay, and Kobayashi 2013). Thus the individual's daily travel patterns were revealed.

This travel route modeling approach creates an individual's own local map and reveals his general travel patterns. However, the sequence of activities and travels conducted on a daily basis cannot be revealed. Thus, this modeling strategy is not suitable for individuals' daily T-A patterns.

Stops and Moves Model

The availability of long term (e.g., one month) GPS trajectory data has stimulated much research on individuals' T-A pattern modeling. Alvares et al. (2007a) and Spaccapietra et al. (2008) developed the "stops-and-moves" model for deriving T-A patterns from individuals' trajectory data. A "stop" is a part of a trajectory where the individual has stayed for a certain amount of time, indicating that an activity is performed at the location. A "move" is a part of a trajectory between two consecutive stops, representing a trip between the two stops. Many studies have identified stops and moves from trajectory data, clustered stops into anchor locations, inferred activity types at anchor locations and travel modes for the moves, and modeled daily T-A patterns.

1) Identify Stops

To be considered as a stop, an individual has to stay at a place for a certain amount of time. This time threshold is specified depending on how long the researchers consider as significant. For example, if the researcher does not wish to include waiting at traffic lights or traffic jams as stops, then the time threshold needs to be larger than the maximum waiting time at traffic lights and traffic jams. The size or range of a space to be considered as a stop is often specified too.

Several studies used the time threshold as the only condition to identify stops. Ashbrook and Starner (2003) identified paired consecutive GPS points whose time gap was at least 10 minutes. They considered these points as stops and the trajectory between two consecutive stops were derived as moves/trips. Alvarez-Garcia et al. (2010) and Chen et al. (2010) used the same method to derive trips from GPS trajectory data, except that five minutes was used in the former study and two minutes was used in the latter study as the minimum time gap between two consecutive GPS points. This simple method works the best when the GPS device loses its signal as the individual walked inside a building and resumes as he walked outside. However, it will not work when the stops are made outside and plenty of GPS points are captured.

Some studies have used both the range of a space and the time threshold as the conditions to identify stops. Hariharan and Toyama (2004) detected stops from single trajectories by identifying a subsection of a trajectory within which all GPS points are within a circle with a 30-meter radius and met the 10 minutes minimum duration. Ye et al. (2009) and Gong et al. (2012) used similar methods to identify stops. Ye et al. (2009) extracted stops by identifying sub-trajectories whose time duration is at least 30 minutes and spatial range is within 200 meters. In Gong and colleagues' study, if the points within a subsection of a trajectory are within 50 meters of each other and the time duration of the subsection is more than 200s, then the subsection of the trajectory is identified as a stop (Gong et al. 2012). Montoliu, Blom, and Gatica-Perez (2013) clustered GPS trajectory points to extract stops from a single trajectory. For a subsection of the trajectory, if the distance between the first and the last point is smaller than a threshold, the time difference is greater than a threshold, and the time difference between each pair

of consecutive points is smaller than a threshold, then this subsection of trajectory points forms a cluster/stop. The authors found that there could be a small distance but a long time gap between two consecutive points in the trajectory due to GPS signal loss. The participant could have visited many other locations in between but the GPS device failed to capture any point; or he might went into a building that caused the GPS signal loss. In the former case, in order to prevent mistakenly identifying the two consecutive points as a stop, the authors added the third condition to the clustering algorithm. However, this condition further prevented identifying the two points in the latter case as a stop. Alvares et al. (2007b) developed an algorithm called Stops and Moves of Trajectories (SMoT) to find stops and moves in trajectories. The authors defined a set of geographical places with geometries and minimum time durations according to the participants' knowledge. When a trajectory intersects the geometry of a place and the duration of the intersection is more than the minimum time duration of the place, the intersection part of the trajectory is considered as a stop. This method is easy to implement. However, creating the list of known geographical places becomes a challenge when the sample size is large. This type of method identifies stops with or without GPS signal loss.

Other studies have used a density based approach to identify stops. Thierry, Chaix, and Kestens (2013) calculated a kernel density surface based on the distribution of GPS points in a single trajectory. The density peaks whose time duration was more than 5 minutes were considered as stops. GPS points in the trajectory were then allocated either to a stop or a move. Tang and Meng (2006) modified the DBSCAN (Density-Based Spatial Clustering of Applications with Noise) method to identify stops from trajectory data. DBSCAN is a density based point clustering algorithm (Ester et al. 1996). For each

point p_i in the dataset, it first draws a circle around the point with a predetermined radius, then it counts the number of points within the circle. If the number is greater than a threshold value, then point p_i is considered as a core point and all the other points in the circle is considered as its neighbors. If core point p_i does not belong to any existing cluster, then p_i and its neighbors form a new cluster; if p_i belongs to an existing cluster, then p_i and its neighbors join the existing cluster. Tang and Meng (2006) added a time window to the algorithm to distinguish stops made at the same location but at different times in one trajectory. For each point p_i in the trajectory, it first retrieves all the points in the trajectory that are within a time difference of point p_i; then it draws a circle around point p_i with a predetermined radius and counts the number of retrieved points within the circle. If the number is greater than a threshold, then point p_i is considered as a core point and all the retrieved points in the circle are considered as its neighbors. The rest of the algorithm is the same as the DBSCAN method. These density based methods cannot identify stops where there are GPS signal loss, for example, inside a building. Palma et al. (2008) also modified the DBSCAN method to extract stops from single trajectories. For each point p_i in the trajectory, it finds a sequence of points in the trajectory that is within a threshold distance of p_i. These distances are measured along the trajectory. If the time difference between the last and the first point in this sequence is greater than a threshold value, then point p_i is considered as a core point and all the other points in the sequence are considered as its neighbors. The rest of the algorithm is the same as DBSCAN. This method can find clusters/stops where some of the GPS points are lost. The authors further computed the threshold distance based on the distribution of distances between two consecutive points in the trajectory. Zhao and Xu (2009) later improved the

calculation of the threshold distance. They divided the trajectory into two parts: the fast-speed part and the slow-speed part. They argued that the distribution of distances between two consecutive points in these two parts is not homogeneous. They calculated the threshold distance based on the distances from the slow part. Their experiment proved that this threshold distance can significantly improve the quality of clustering. The modified density based method can not only identify stops with or without signal loss, but also discover stops where the GPS points form clusters of irregular shapes. Moreover, it is much more efficient and effective than other methods.

Very few studies used the circuity property of stops in trajectory data to separate stops from moves. For a subsection of trajectory points, if the trajectory distance between the first and the last point divided by the Euclidian distance between them is greater than seven, then this subsection of trajectory is considered as a stop (Wolf et al. 2004). Manso et al. (2010) developed the Direction-Based Stops and Moves of Trajectories (DB-SMoT) algorithm using direction change to find stops in a single trajectory. For each point in the trajectory, if the direction change at the point is bigger than a threshold then this point is considered as a candidate cluster point. For a sequence of connected candidate cluster points, if the total number of these points is more than a threshold and the time duration is greater than the minimum time duration, then this sequence of candidate cluster points forms a cluster/stop. This method is effective in identifying interesting places in trajectories with apparent direction changes, such as fishing spots in fishing vessel trajectories (with around 90 percent accuracy). However, this type of method does not work for stops with GPS signal loss.

2) Cluster Stops into Anchor Locations

When analyzing one individual's trajectory data collected over weeks or months, a significant number of stops could be identified. Many of these stops cluster at a few locations. This is because human daily activities are repetitive. They may visit the same location conducting the same activity multiple times a week/month, thus multiple stops at the same location could be extracted from the individual's trajectories and these stops represent the same anchor location and the same activity. Researchers usually cluster these stops and mark the ones in the same cluster with the same location ID. Thus we know the individual repeatedly visited a few locations and conducted the same activities, which can be used to model the individual's daily T-A patterns. Grengs, Wang, and Kostyniuk (2008) derived trip end points to represent stops and clustered them into a few anchor locations if the distance between these stops is within 100 feet. Stopher, FitzGerald, and Zhang (2008) clustered all trip end points within 200 meter buffer zones into a few anchor locations.

There are three common types of point clustering algorithms that can be used to cluster stops into anchor locations: partitioning (K-Means), hierarchical, and density based algorithms. Partitioning algorithms run iteratively to minimize the sum of the squared distances of each point to its cluster center. Ashbrook and Starner (2003) and Alvarez-Garcia et al. (2010) derived trip end points to represent stops and clustered them using a variant of K-Means clustering algorithm. There are a few problems of using partitioning algorithms to cluster stops. First, the number of clusters must be specified before running the algorithm, which can be difficult for clustering stops since the exact number of places an individual have visited is usually unknown. Second, noise points

cannot be excluded from a cluster. Not all stops are made at a anchor location. Stops made at non-anchor locations should be excluded from a cluster. Third, partitioning algorithms are not deterministic, meaning that the final clustering result depends on the initial random assignment of points into clusters. Hierarchical algorithms establish a hierarchical structure of all points. The algorithm runs either from top down that iteratively splits all points into smaller clusters or from bottom up that iteratively combines the closest points into a cluster. A termination condition needs to be specified indicating when the iteration should be stopped. The clusters are organized as a hierarchical tree and each branch of the tree represents a cluster. Hariharan and Toyama (2004) clustered trip end points using an agglomerative hierarchical clustering algorithm. The algorithm starts with treating each trip end point as a cluster. During each iteration of the algorithm, if the distance between two closest clusters is smaller than a specified distance (e.g., 250 meters), then they are merged; otherwise, the algorithm stops and outputs all remaining clusters as locations. Chen et al. (2010) also clustered trip end points into anchor locations using a hierarchical clustering algorithm. Hierarchical algorithms allow researchers to specify the spatial scale of clusters, rather than the number of clusters (partitioning algorithms) or the number of points contained in a cluster (density based algorithms). However, it is difficult to define the proper termination condition for a specific application. Density based algorithms such as DBSCAN (introduced in the last section) identifies clusters of different shapes, do not require any prior knowledge about the number of clusters, can effectively exclude noise points, and work much more efficiently than the above two methods. Lv, Chen, and Chen (2013) identified stops from trajectories and clustered them using the DBSCAN method.

3) Infer Activity Types at Anchor Locations

After identifying anchor locations, researchers often use other datasets to help infer the types of activities (e.g., work, school, shopping, recreation, etc.) conducted at these locations. Wolf et al. (2004) and Grengs, Wang, and Kostyniuk (2008) both inferred activity types using land-use data, business listings, time of the activity, activity duration, and visiting frequency. Chen et al. (2010) overlay a business listing point file, a participant's self-reported activity locations file, and a land-use parcel file onto the extracted anchor locations. For low density areas, the type of activity at a anchor location can be directly inferred from intersection with known places. For dense areas, a probabilistic model was applied to infer activity types considering the visiting history of each anchor location.

4) Infer Travel Mode for the Moves

For each derived move, travel mode was often inferred using various methods, such as rule-based models (Chung and Shalaby 2005; Chen et al. 2010; Wu et al. 2011; Gong et al. 2012), decision tree models (Reddy et al. 2010; Wu et al. 2011), Hidden Markov Models (Reddy et al. 2010), supervised learning methods (Zheng et al. 2010), neural networks (Gonzalez et al. 2010), Support Vector Machine (SVM) algorithms (Dodge, Weibel, and Forootan 2009; Bolbol et al. 2012), and fuzzy membership classification (Biljecki, Ledoux, and Van Oosterom 2013). Urban canyon effect and complicated urban transportation networks often contribute to errors in mode detection.

5) Identify Daily T-A Patterns

Based on the "stops-and-moves" model, Bogorny and colleagues (Bogorny, Kuijpers, and Alvares 2009; Bogorny, Heuser, and Alvares 2010) incorporated repetitiveness and consecutiveness into daily T-A pattern analysis. A frequent pattern is defined as a set of stops or moves that occur in a minimum number of daily trajectories during the study period. A sequential pattern is a set of stops or moves in a particular chronological order that occur in a minimum number of daily trajectories during the study period. Grengs, Wang, and Kostyniuk (2008) detected and mapped the frequently visited locations and traveled routes from an individual's GPS trajectory data collected over four weeks, which could be considered as the individual's frequent T-A patterns (see Figure 10 in Grengs, Wang, and Kostyniuk 2008). However, temporal information and connections between travels and activities (the sequential pattern) were not revealed in the study.

Of the above four methods for modeling individual daily T-A patterns, activity location modeling and travel route modeling each focuses on only one part of the T-A behaviors, and neither of them reveal the connections between travels and activities nor the order of conducting these travels and activities on the daily basis. The daily activity sequence method captures the order of conducting activities in a day, but ignores the locations of these activities and the travels between them. The stops and moves model works fine for identifying single stops and moves, but fails to make a connection and order among them.

IV. COLLECTIVE ACTIVITY PATTERNS MODELING AND ANALYSIS

This chapter demonstrates the proposed space-time modeling techniques for collective activity patterns. The empirical data contains the GPS trajectory data of a sample of taxi cabs in San Francisco, California from May to June in 2008. The analysis results and discussions are presented in the third section of the chapter.

Site Description and Data

San Francisco, California is one of the highest populated urban areas in North America. Surrounded by water on three sides, the total land area of the city is approximately 120 square kilometers with a population of about 805,235, according to the 2010 US Census. There is a total of 194 census tracts in the land area of the city (Figure 1). The census tract 2010 boundary data of San Francisco were downloaded from the U.S. Census Bureau website for analysis.

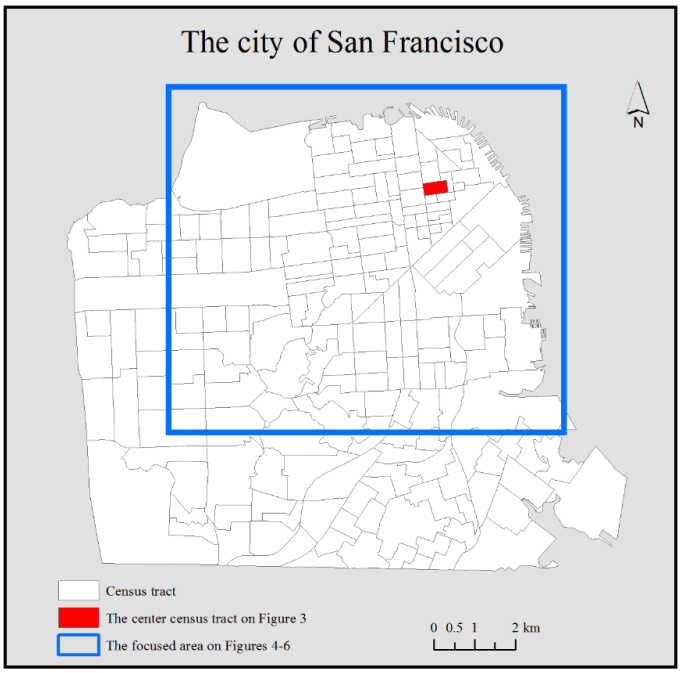


Figure 1. The city of San Francisco.

The San Francisco Dataset was downloaded from CRAWDAD (Community Resource for Archiving Wireless Data At Dartmouth) website (crawdad.cs.dartmouth.edu). It contains the GPS trajectory data of 536 taxi cabs in San Francisco over a period of twenty-two days, from May 18 to June 8 in 2008. The location-updates for each taxi contain the latitude/longitude coordinates, time stamp, and the taxi occupancy status. Data were recorded approximately every sixty seconds. There are about 1,500 authorized cabs doing business in the city of San Francisco. This trajectory dataset was collected for the taxi cabs (about one third of all the authorized cabs in the city) that belong to the Yellow Cab of San Francisco, the largest cab company in the city.

One concern of using this dataset is the representativeness of the taxi data. First, the dataset contains GPS trajectory data of about one third of all cabs in San Francisco. The taxi passengers' activities extracted from this dataset form a reasonable sample of all taxi passengers' activities in San Francisco. Second, people are more likely to use taxi services for business, tourism, or entertainment types of activities, and less likely for daily life activities, such as grocery shopping, daily trip to work and home, picking up and dropping off children to schools, etc. Thus, activity patterns extracted from this trajectory dataset do not represent urban residents' daily life activity patterns, but rather urban commercial, tourism, and entertainment activity patterns. Third, people may use other transportation modes to reach commercial, tourism, and entertainment activity destinations, such as driving a car, subway, bus, bicycle, walking, etc. These people's activities are underrepresented. Thus, activity patterns extracted from this taxi trajectory dataset represent urban commercial, tourism, and entertainment activity patterns at a

certain degree. If we consider the fact that San Francisco is one of the leading centers of commercial, tourism, and cultural activities in the United States, then analyzing taxi passenger's activity patterns may become more interesting. To reveal city residents' daily life activity patterns, other datasets need to be considered.

Due to the high spatial and temporal accuracy and continuous coverage over large samples, GPS trajectory datasets can play a unique role for transportation and activity pattern studies. A number of studies that were based on analyses of this taxi cab trajectory dataset of San Francisco have been reported. Examples include real-time traffic modeling and estimation (Herring 2010), optimal route recommendation (Hu et al. 2012), future user location prediction (Scellato et al. 2011), and trajectory data privacy protection (Gambs, Killijian, and Del Prado Cortez 2010; Hwang, Hsueh, and Chung 2012). Noteworthy is that none of these studies have explored the dynamic patterns of urban activities at the collective level.

Taxi cabs' GPS trajectory data was projected first during the initial data processing. A 3D point feature (x, y, t) was used to represent an activity instance from the trajectory dataset. The x and y coordinates recorded the location, and the t coordinate recorded time. The location of an activity instance was defined as a passenger pick-up or drop-off location. Such a location was inferred when there was a change in taxi occupancy status. The average of the x and y coordinates between the location immediately before a taxi status change and the one after the change were taken to represent a pick-up or drop-off location. The time of each activity was estimated by averaging the time stamps of these two consecutive locations. A total of 808,375 passenger activity points was identified from this dataset. These points were mapped to

the corresponding census tracts for each hour interval. The number of activities that fall within each census tract during each hour interval was recorded.

Methodology

Collective activity patterns can be revealed by the locations and times of activity hot spots in the city and the evolvement of these hot spots over time. This section defines an activity hot spot and the dynamic stages of hot spots in a life cycle and describes methods for identifying an activity hot spot and its development stage. A prediction method for future dynamics of activity hot spots is also presented in this section.

Detection of Activity Hot Spots

An activity hot spot was defined in this research as a census tract with a significantly large number of activity instances during a one-hour period. Poisson distribution was used to identify activity hot spots in this research. In probability theory and statistics, Poisson distribution calculates the probability of a given number of events occurring at a fixed time interval and/or space (Haight 1967). Two conditions are required for using Poisson distribution (Haight 1967). First, the average rate of the occurrence of the events is known. Second, the occurrence of the events is independent. The theoretical distribution of the taxi passengers' activities across the study area and through the study time was assumed to be completely random. This means that the occurrence of each activity at a specific census tract and during a specific hour interval was random and independent of other activities. Thus, it was a reasonable assumption that the number of activity instances occurring within a census tract during a one-hour

interval obeys the Poisson distribution. Since the average rate of the occurrence of these activities could be easily calculated using the San Francisco Dataset, Poisson distribution was used to estimate the probability of a certain number of activity instances occurring at a certain tract during a certain hour.

Let λ_i denote the expected number of activities that occur at census tract *i* during an hour *j*, the probability of observing *k* instances in tract *i* during hour *j* is:

$$P(k_{i,j}) = \frac{e^{-\lambda_i * \lambda_i}{}^{k_{i,j}}}{k_{i,j}!}, \ k_{i,j} = 0, 1, 2, ... \quad (2)$$

$$\lambda_i = \frac{1}{T} * \frac{a_i}{A} * N \quad (3)$$

Where a_i is the areal size of tract *i*, *A* is the total areal size of the 194 census tracts on land, *T* is the total number of hours in the data period, and *N* is the total number of activity instances. Equation (2) generates small probability for large $k_{i,j}$. When *P* is smaller than a threshold (for example, 0.01), it means it is very unlikely to observe *k* instances in tract *i* during hour *j*, indicating that census tract *i* has a significantly large number of activity instances during hour *j* thus it forms an activity hot spot. A computer program was developed to screen whether a census tract forms a hot spot during any one-hour interval.

The potential impact of Modifiable Areal Unit Problem (MAUP) (Fotheringham and Wong 1991) must be addressed. Census tract was selected as the spatial unit for the analysis due to the following considerations. First, census tracts are a stable set of geographic units for the United State Census Bureau to present statistical data. Each census tract usually covers a physically contiguous area with a population size between 1200 and 8000. Census tract boundaries generally follow physical features or administrative boundaries. Using census tracts as the spatial units provides possibility of

linking the social-economic statistical data with the discovered activity patterns. Although an equal-sized grid that is draped over the city may serve as a framework for summarizing the taxi passengers' activity data, such a spatial partition is incapable to reflect local socioeconomic characteristics. Moreover, smaller units will lead to significant increase in processing time, creating a problem for the analysis as the algorithm runs on a single desktop computer. However, it should be noted that the application of census tracts as the spatial units for the empirical analyses reported in this research does not automatically exclude other spatial partition schemes. Similar to MAUP, there is a Modifiable Temporal Unit Problem (MTUP), meaning that the hot spot patterns might be different as the time unit varies. One-hour interval was chosen because it is a common unit for daily pattern cycle. A study aiming at comparing patterns at different spatial and temporal scales should consider using other spatial and temporal units.

Dynamics of Activity Hot Spots in a Life Cycle

The dynamics of activity hot spots can be described using a six-stage spectrum of life cycle. The hot spot status of one specific census tract (the center) and that of its surrounding tracts (the periphery) at two consecutive time periods (time 1 and time 2) were considered to define the hot spot's development stage. The periphery zone for a center tract was defined as the combined area of all the census tracts which shared a boundary with the center tract. The expected and observed number of activities for the periphery zone was respectively the sum of the expected and observed number of activities in each of its member tracts. The hot spot status of the periphery zone can be

examined using equation (2) and by substituting a_i by the areal size of the periphery zone. It needs to be recognized that this definition of the periphery zone and treatment of the examination of the hot spot status of the periphery zone has an impact on the periphery zone hot spot status, the development stages of hot spots, the overall collective activity patterns across the urban space, and the evolvement of these patterns over time. Combining all the adjacent census tracts together as the periphery zone simplifies the examination of the periphery zone hot spot status and the evaluation of hot spot development stage by reducing the number of variables. However, it is unable to count the variations among the member tracts of a periphery zone. Other definitions and treatments of the periphery zone may result in the discovery of different collective activity patterns and evolvements. This aspect may be explored in future studies.

A hot spot's life cycle includes six development stages:

1) Emergence: a hot spot emerges at a center zone, its periphery zone, or both, if neither the center nor the periphery zone is a hot spot at time 1 and at least one of them becomes a hot spot at time 2.
2) Expansion: a hot spot expands if either a center or its periphery zone is a hot spot at time 1 and both of them are hot spots at time 2. An outward expansion presents if a hot spot expands from center to periphery zone; an inward expansion exists if a hot spot expands from periphery zone to center zone.
3) Stableness: a hot spot is stable if the status of both center and periphery zones remain the same from time 1 to time 2.
4) Shrinkage: a hot spot shrinks if both center and periphery zones are hot spots at time 1 but only one of them remains as a hot spot at time 2. An outward

shrink presents if the center zone loses its hot spot status; an inward shrink exists if the periphery zone loses its hot spot status.

5) Displacement: a hot spot is displaced if either the center or the periphery zone is a hot spot at time 1, and their hot spot statuses switch at time 2. An outward displacement refers to the moving of a hot spot from center to periphery zone; an inward displacement is the moving of a hot spot from periphery zone to center.
6) Decease: a hot spot deceases at center, periphery zone, or both if the center, the periphery zone, or both are hot spots at time 1 but neither is a hot spot at time 2.

Table 2 summarizes the typology of the six stages. A computer program was developed to assess the hot spot status following the six-stage spectrum typology. The dynamics of all activity hot spots were assessed by considering both central and periphery zones during two consecutive hour intervals.

Table 2. Typology of activity hot spot dynamics throughout a life cycle.

Life Cycle Stage	Dynamic Pattern	Zones in a Neighborhood	Time 1 Hot Spot Status	Time 2 Hot Spot Status
Emergence	Center Emergence	Center	No	Yes
		Periphery	No	No
	Periphery Emergence	Center	No	No
		Periphery	No	Yes
	Overall Emergence	Center	No	Yes
		Periphery	No	Yes
Expansion	Outward Expansion	Center	Yes	Yes
		Periphery	No	Yes
	Inward Expansion	Center	No	Yes
		Periphery	Yes	Yes
Stableness	Center Stableness	Center	Yes	Yes
		Periphery	No	No
	Periphery Stableness	Center	No	No
		Periphery	Yes	Yes
	Overall Stableness	Center	Yes	Yes
		Periphery	Yes	Yes
Shrinkage	Outward Shrinkage	Center	Yes	No
		Periphery	Yes	Yes
	Inward Shrinkage	Center	Yes	Yes
		Periphery	Yes	No
Displacement	Outward Displacement	Center	Yes	No
		Periphery	No	Yes
	Inward Displacement	Center	No	Yes
		Periphery	Yes	No
Decease	Center Decease	Center	Yes	No
		Periphery	No	No
	Periphery Decease	Center	No	No
		Periphery	Yes	No
	Overall Decease	Center	Yes	No
		Periphery	Yes	No

Prediction of Hot Spots Dynamics

Being able to predict hot spot dynamics has a great potential for better traffic management and service delivery. An initial observation of the case data showed a clear weekly periodic pattern on the daily total of activity instances in the study area during the twenty-two-day period (Figure 2). Based on this observation, it was reasonable to make an assumption that collective activity patterns also followed a weekly repetitive cycle. As collective activity patterns were revealed by the spatial temporal distribution of activity hot spots and their development processes, it was reasonable to assume that activity hot spot distribution and dynamics also followed a weekly repetitive cycle.

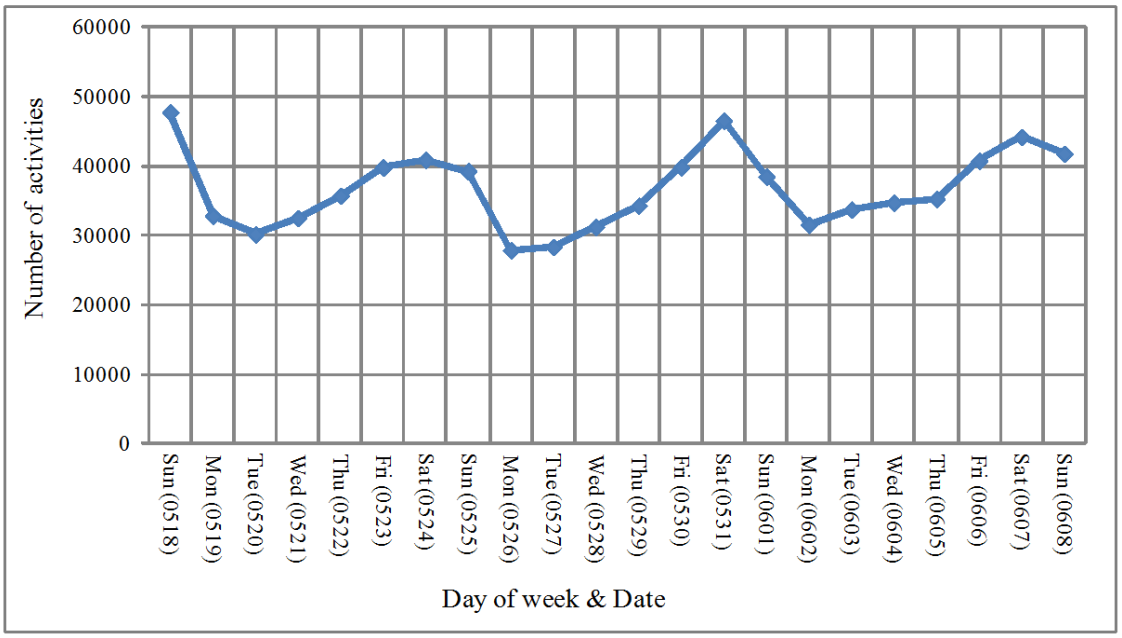


Figure 2. The daily total of activity instances in the study area during the twenty-two-day period.

With this assumption, the number of activity instances in a spatial unit during a future time interval was estimated as the historical average number of activity instances in the spatial unit during the same time interval of a day on the same day of a week. Let *d* represents the *d*th day of a week. The values of *d* $\{0, 1, 2, 3, 4, 5, 6\}$ correspond to the days of a week $\{Sunday, Monday, Tuesday, Wednesday, Thursday, Friday, Saturday\}$. j_d represents hour *j* on the *d*th day of a future week. K_{ij_d} is the estimated number of activity instances in census tract *i* during hour *j* on the *d*th day of a future week. *w* represents the *w*th week in the data collection period. *m* is the total number of weeks in the data collection period. j_{wd} represents hour *j* on the *d*th day of the *w*th week. $k_{ij_{wd}}$ is the number of activity instances in census tract *i* during hour *j* on the *d*th day of the *w*th week. Equation (4) calculates the estimated number of activity instances in a census tract during a future hour.

$$K_{ij_d} = \frac{\sum_{w=1}^{m} k_{ij_{wd}}}{m},\ w = 1, 2, 3, \dots, m,\ d = 0, 1, 2, 3, 4, 5, 6, \quad (4)$$

Thus, whether census tract *i* would host an activity hot spot during a future hour *j* can be assessed using Poisson distribution (Equation 2 and 3). The hot spot status of all census tracts in the study area during a future hour can be assessed using the above method.

A future hot spot's development stage can be evaluated using the estimated hot spot status of the center census tract and that of the periphery zone at two consecutive hours in the future. As the periphery zone for a center census tract was defined as the combined area of all the census tracts sharing a boundary with the center tract, the estimated number of activity instances in the periphery zone during a future hour was respectively the sum of the estimated number of activity instances in each of its member

tracts during the future hour. Thus, the hot spot status of the periphery zone during a future hour can be assessed using Poisson distribution (Equation 2 and a_i is substituted by the areal size of the periphery zone). When the estimated hot spot status of the center tract and its periphery zone during a future hour *j* was connected with that of the future hour *(j+1)*, the development stage of this future hot spot can be evaluated using the typology defined in Table 2. Table 3 shows the prediction method for the development stage of a future hot spot during two consecutive hours.

Table 3. The prediction method for the development stage of a future hot spot during two consecutive hours.

Zones in a Neighborhood	Future Hour *j* Hot Spot Status	Future Hour *(j+1)* Hot Spot Status
Center census tract *i*	Estimated with K_{ij_d}	Estimated with $K_{i(j+1)_d}$*
The periphery zone i_p*	Estimated with $K_{i_p j_d}$*	Estimated with $K_{i_p(j+1)_d}$*

*Note: $K_{i(j+1)_d}$ is the estimated number of activity instances in census tract *i* during hour *(j+1)* on the *d*th day of a future week. i_p is the periphery zone of census tract *i*. $K_{i_p j_d}$ is the estimated number of activity instances in periphery zone $i_p$during hour *j* on the *d*th day of a future week. $K_{i_p(j+1)_d}$ is the estimated number of activity instances in periphery zone i_p during hour *(j+1)* on the *d*th day of a future week.

When *j* is the current hour, *(j+1)* is the upcoming hour. Activity instance data during hour *j* may be collected in real time by a central server and summarized at the end of hour *j*. Assume that the central server maintains a database of activity instances for the past *m* weeks. Activity hot spot status for each census tract and its periphery zone during

hour *j* can be evaluated using Equation (2) and (3). For the upcoming hour *(j+1)*, the number of activity instances in each census tract $K_{i(j+1)_d}$ and its periphery zone $K_{i_p(j+1)_d}$ can be estimated using Equation (4). The hot spot status for each census tract and its periphery zone during the upcoming hour *(j+1)* can be assessed using Equation (2) and (3). Thus, the hot spot status at each census tract and its periphery zone during the current hour *j* and the upcoming hour *(j+1)* is calculated and estimated. They are combined to determine the upcoming development stage of all activity hot spots in the study area.

To evaluate prediction accuracy, the predicted activity hot spots and their development stages can be compared with the calculated activity hot spots and their development stages using the observation data, once a future hour became a past hour and activity data was collected. The accuracy of the prediction is closely related to the assumption on a weekly repetitive cycle for collective activity patterns. Other assumptions on a temporal repetitive cycle (such as daily, seasonal, etc.) may be explored in future studies. The prediction accuracy may also be related to the collection period (number of weeks) of the historical data. Moreover, the prediction reflects historical average weekly collective activity patterns, as it was made based on a weekly repetitive assumption. Any significant deviation of the observed hot spot dynamics from the prediction may indicate the presence of abnormal or special events in the study area (for example, a large ethnic festival, a large sports event, etc.).

Findings and Discussion

Hot spot analyses in this research were conducted for each census tract at the one-hour interval. Each one-hour interval is referred to using the starting time during a

twenty-four-hour period. For example, "hour 0 on Monday May 19" refers to the time period of 12:00 am – 1:00 am on May 19. For each of the census tracts that were identified as activity hot spots at a certain hour (i.e. time 1), the dynamics of that hot spot was assessed by connecting with the hot spot patterns centered at the same census tract at the hour immediately before (i.e. time 0) and the hour immediately after (i.e. time 2) that hour. Hence, the dynamics of a hot spot was determined by the development of the activity patterns during two consecutive one-hour intervals.

Life Cycle of an Activity Hot Spot

In order to illustrate the life cycle of an activity hot spot, the hot spot status for a focal census tract and its surrounding tracts were analyzed for Saturday May 31, 2008. The focal census tract was in downtown San Francisco (Figure 3). The maps in Figure 3 illustrate the hot spot life cycle of the center tract through the stages of hot spot emergence, expansion, stableness, shrinkage, and decease. The center tract and its surrounding tracts formed a cluster of stable hot spots between hour 0 and hour 4 (12 – 5 am) except for the two surrounding tracts that showed some changes. An outward shrinking was detected at the center tract during hour 4 to hour 5 (i.e. 4 – 6 am); it spread outward until the center tract deceased as a hot spot during hour 6 (i.e. 6 – 7 am) and remained so until hour 9 (i.e. 9 – 10 am). A hot spot emerged from the surrounding tracts during hour 10 and expanded to the center tract during hour 11. Both the center and the surrounding tracts remained as hot spots throughout the rest of the day.

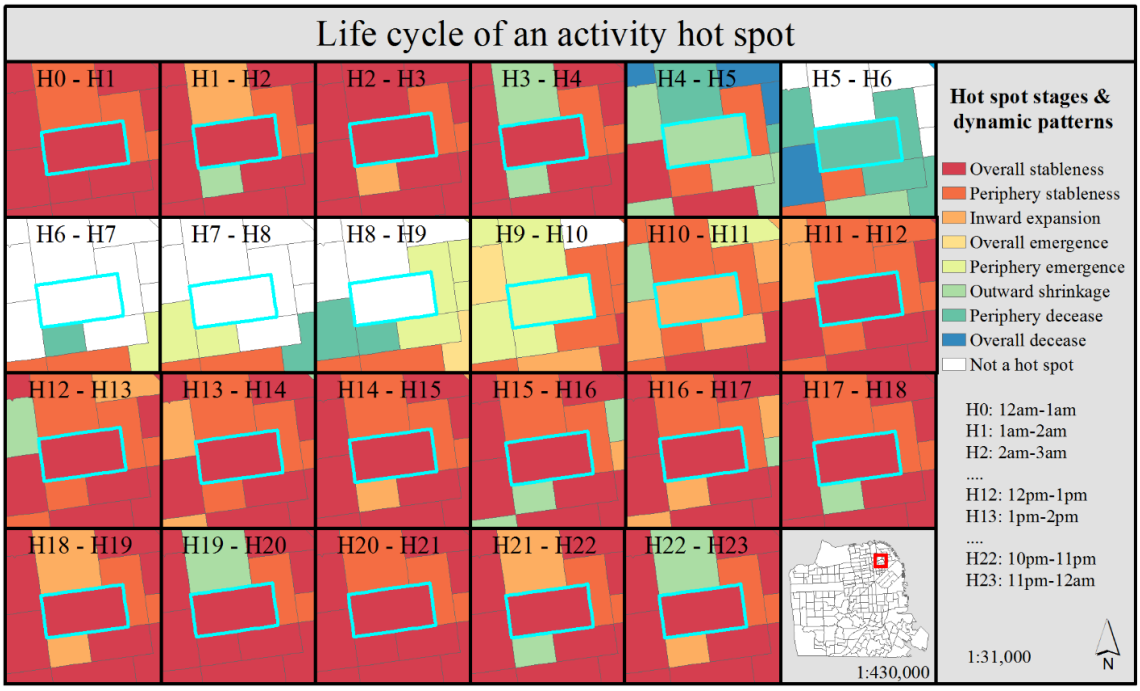


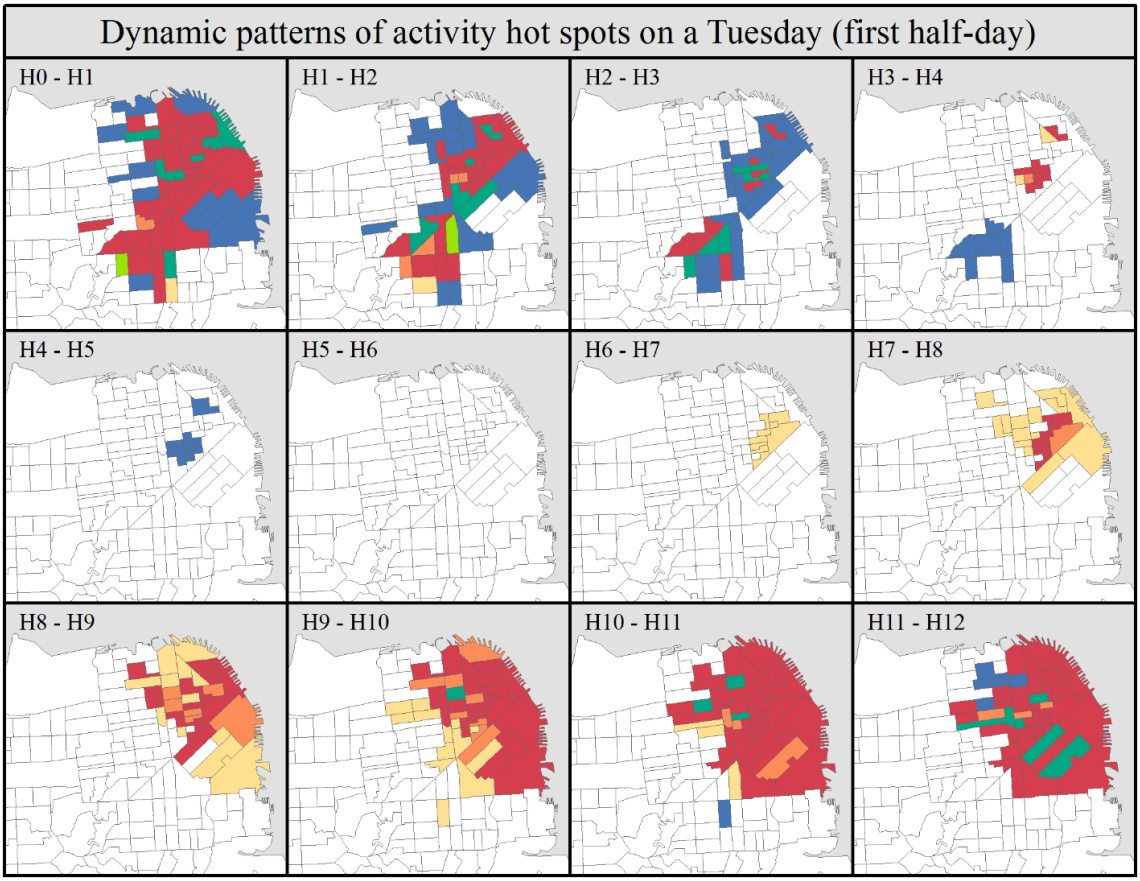
Figure 3. The life cycle of an activity hot spot.

It is important to understand that the timing and the sequences of the different stages of hot spot development vary on different days of a week and for different parts of a city. The variation reflects the spatial patterns and rhymes of urban life. After building a city-wide profile of space-time hot spots, better and more informed decisions can be made for traffic management, public safety control, emergency response and other services.

Dynamic Patterns of Activity Hot Spots during a Day

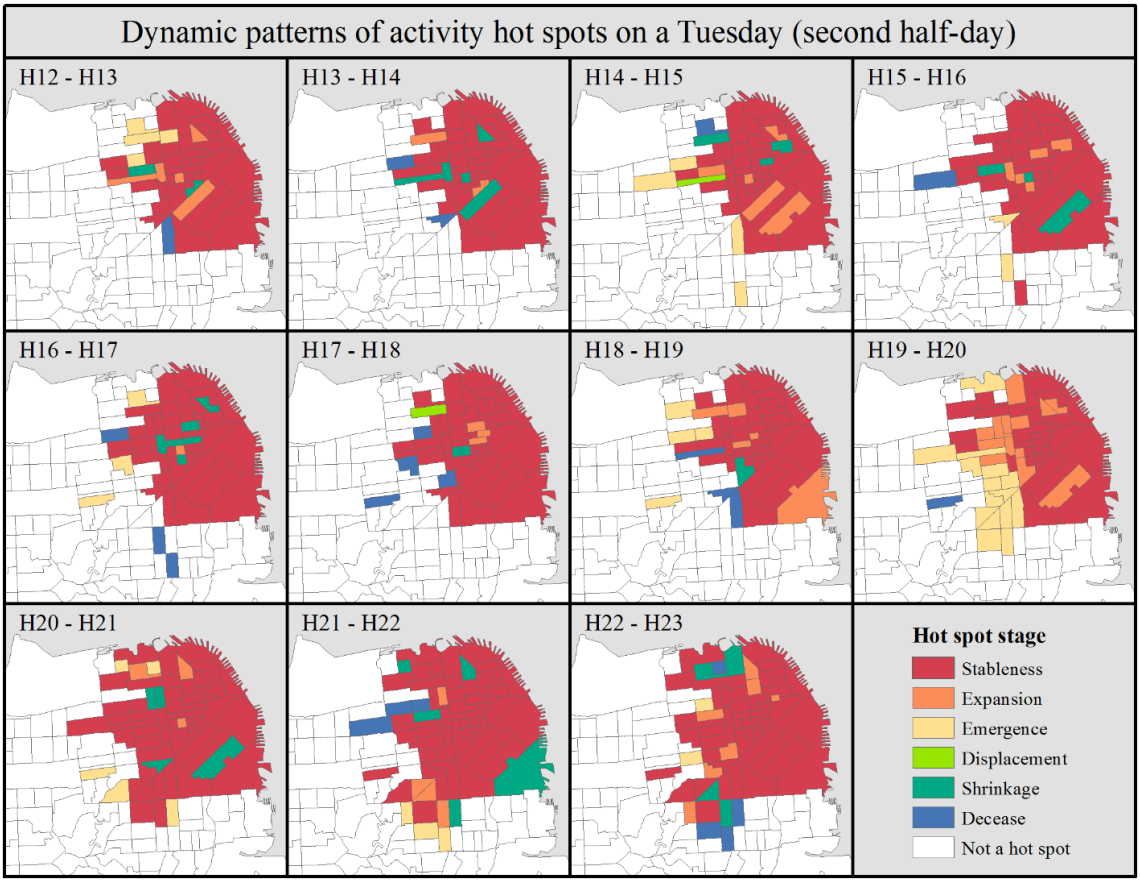
As Figure 2 showed a clear weekly periodic pattern on the daily total of activity instances in the study area, one can see that the largest number of activities existed on Saturdays, and the smallest numbers appeared to be on Mondays and Tuesdays. Thus,

one Tuesday (May 27, 2008) and one Saturday (May 31, 2008) were selected for a close examination of the variations in the spatial temporal distribution of activity hot spots and their development patterns. Figures 4 and 5 illustrate the hot spots' dynamic patterns for every two consecutive hours on these two days. For visualization purpose, only the six stages of hot spot life cycle were reported on the maps. The details of the hot spot dynamics, as described by the fifteen categories in the second column of Table 2, can be assessed following the second subsection of the methodology section of this chapter. The details of the comparison on the dynamic hot spot patterns between a Tuesday and a Saturday were discussed below.



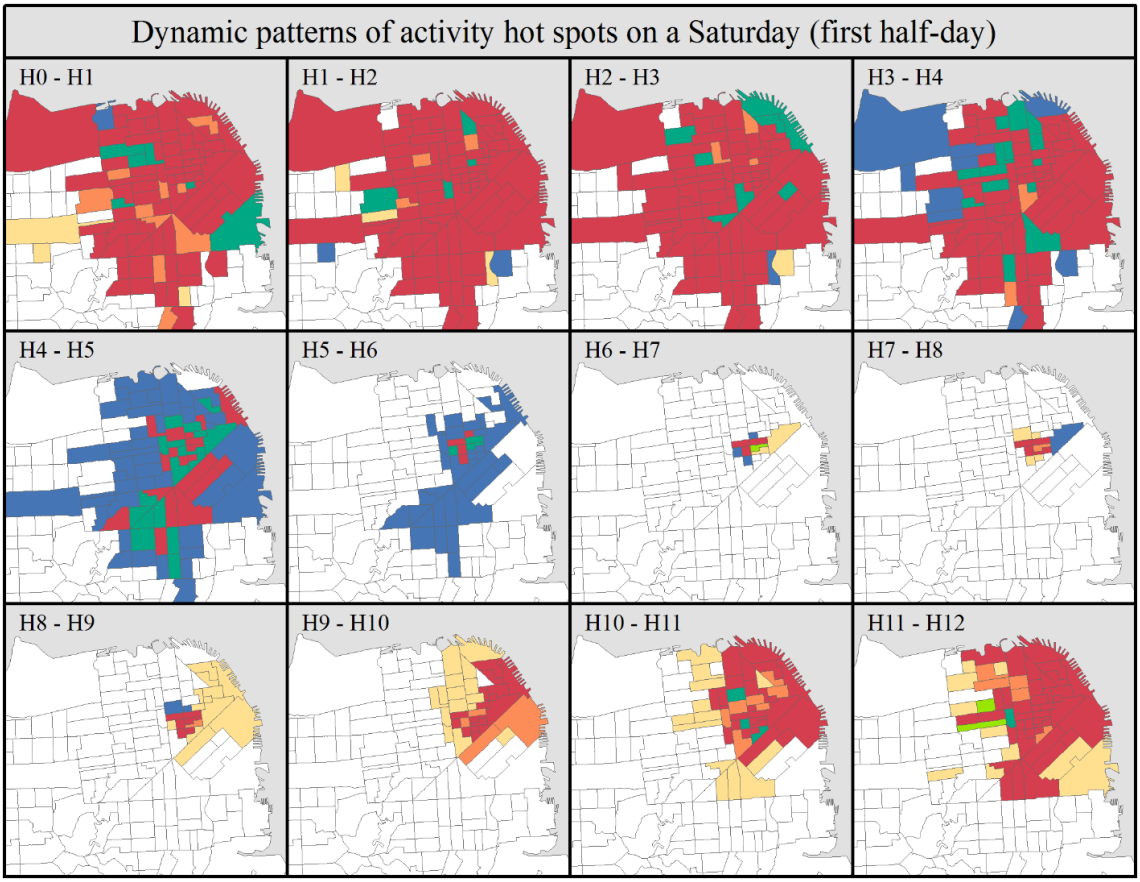
A. Tuesday: first half-day.

Figure 4. Dynamic patterns of activity hot spots on a Tuesday.



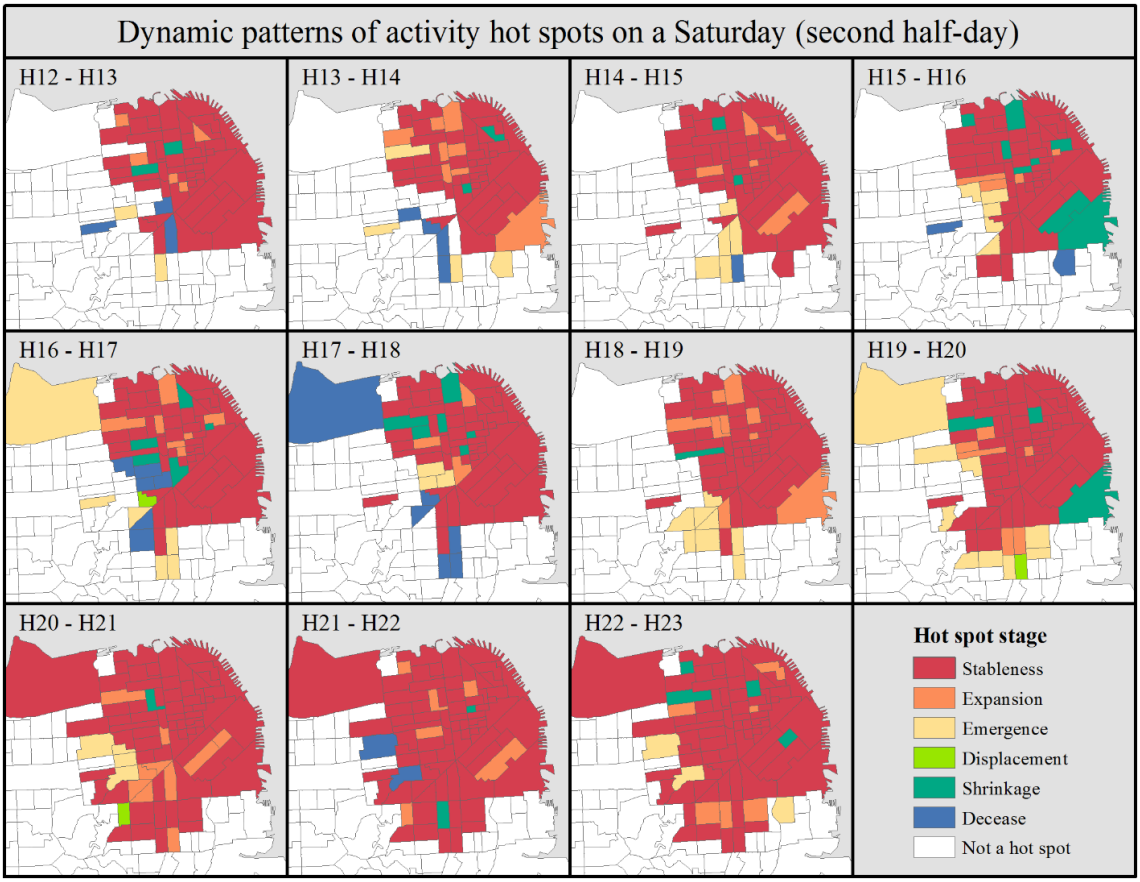
B. Tuesday: second half-day.

Figure 4. (Continued) Dynamic patterns of activity hot spots on a Tuesday.



A. Saturday: first half-day.

Figure 5. Dynamic patterns of activity hot spots on a Saturday.



B. Saturday: second half-day.

Figure 5. (Continued) Dynamic patterns of activity hot spots on a Saturday.

The pattern differences in the dynamics of activity hot spots between the selected Tuesday and Saturday showed clearly when cross-referencing Figures 4 and 5. Compared to the post-midnight hours on the Tuesday, many more hot spots existed during the same time on the Saturday and they existed in a larger geographic area. A number of hot spots started shrinking and dying during hour 0 and hour 1 (i.e. 12 – 2 am) on the Tuesday, but they did not do so on the Saturday until about hour 3 to hour 4 (i.e. 3 – 5 am), suggesting a three-hour extended active period on the Saturday. Similarly, most hot spots disappeared / died during hour 3 to hour 4 (i.e. 3 – 5 am) on the Tuesday. However, this

pattern did not show until about hour 6 and hour 7 (i.e. 6 – 8 am) on the Saturday, which is another three-hour delay. These pattern differences during mid-night and early morning hours reflected the activity rhythms of most urban dwellers on the different days of a week. People are more active in the midnight and post-midnight hours on a Saturday than a Tuesday. Many people work during the day on Tuesday, and they are likely to be resting in bed during Tuesday early morning hours, resulting in few hot spots city-wide. This observation was further confirmed by the statistics reported in Table 4 for the hours 2-3 (i.e. 2 – 4 am) and the hours 4-5 (i.e. 4 – 6 am). There were more active hot spots during these hours on the Saturday, while most hot spots were deceased during the same time on the Tuesday.

Hot spots started emerging at hour 7 (7 – 8 am) on Tuesday morning. The emerging and expanding mode lasted until hour 10 (10 – 11 am), resulting in a total of about fifty hot spots. On Saturday, hot spots did not start the emerging mode until hour 9 (9 – 10 am). The emerging and expanding mode lasted also around four hours until hour 12 (12 – 1 pm), totaling about sixty new hot spots. There seemed to be a two-hour delay for the morning activity hot spots on the Saturday compared to the Tuesday. There were more activity hot spots on the Tuesday morning than the Saturday morning. However, after the delayed peak of emerging hot spots, the total number of activity hot spots on the Saturday surpassed that on the Tuesday around noontime (Table 4, column "H11-H12"). Thus, Saturday saw more activity hot spots after noontime (see the last maps in Figures 4A & 5A). On both days, a surge of emerging activity hot spots appeared during the evening time at hours 19-20 (Figures 4B & 5B). Work and business-related activities were likely responsible for the active early morning patterns on the Tuesday, while social

and tourism activities may be reflected more by the Saturday pattern. Most people get up early for their busy weekday routines on Tuesday, but many may choose to follow a late schedule on Saturday. However, people tend to attend more social and entertainment events and tourism activities on Saturday, resulting in an overall larger number of activity hot spots on the Saturday.

Table 4. Comparing the dynamics of activity hot spots during the selected hours on a Tuesday (May 27, 2008) and a Saturday (May 31, 2008).

Hot Spot Stage	H2 – H3 (2am – 4am)		H4 – H5 (4am – 6am)		H8 – H9 (8am – 10am)		H11 – H12 (11am – 1pm)		H20 – H21 (8pm – 10pm)	
	Tue.	Sat.	Tue.	Sat.	Tue.	Sat.	Tue.	Sat.	Tue.	Sat.
Emergence	0	1	0	0	12	12	0	10	5	4
Expansion	0	3	0	0	6	1	3	5	3	10
Stableness	9	76	0	19	17	6	37	39	55	65
Shrinkage	7	8	0	18	0	0	6	1	4	1
Displacement	0	0	0	0	0	0	0	2	0	1
Decease	22	1	11	42	0	2	4	0	0	0
Total active hot spots	16	88	0	37	35	19	46	57	67	81

Note: The focused study area consists of 194 census tracts, each of which was evaluated as a potential center for an activity hot spot.

Overall, the Tuesday patterns described above may represent a typical weekday (work day) scenario: many activity hot spots emerge during the morning rush hours, and most hot spots decease around the midnight hours. The Saturday patterns reveal a typical

weekend-day situation: many hot spots start emerging in the late morning hours and remain active for longer hours, and more hot spots are generated throughout the day overall. These general patterns of activity hot spots and their dynamics reflect the overall urban activity tides and ebbs. Business related activities start getting active early in the morning on a weekday. Weekend activities tend to be related to tourism and entertainment. They are likely to start in mid or late morning hours and may last until midnight or even the next morning.

Predicting the Dynamics of Activity Hot Spots

As GPS trajectory data in the San Francisco dataset was collected over twenty-two days from Sunday May 18, 2008 to Sunday June 8, 2008, it was split into two parts. The first twenty-one days (three weeks) were used as historical data for hot spot prediction. The last day, Sunday June 8, 2008 was used as the ground truth for a comparison with the prediction and calculating prediction accuracy.

Assume the current time was at the end of hour 4 (4 - 5am) and beginning of hour 5 (5 - 6am) on Sunday June 8, 2008. Activity instance data during hour 4 had been collected and summarized into census tract units. Hot spot status during hour 4 for each census tract and its periphery zone were evaluated using Equation (2) and (3). The number of activity instances in each census tract (K_{i5_0}) and its periphery zone ($K_{i_p5_0}$) during hour 5 was estimated using Equation (4) and the three-week historical data. Thus, hot spot status during hour 5 was estimated for each census tract and its periphery zone. Compared with the calculated hot spot status during hour 5 with the ground truth data on Sunday June 8, 2008, the number of census tracts and periphery zones with the correct

prediction and prediction accuracy was reported in Table 5. The calculated hot spot status during hour 4 with the ground truth data and the estimated hot spot status during hour 5 with the historical data were combined to estimate hot spot development stages during the two hours (hour 4 - 5, 4 - 6am). The observed hot spot development stages during hour 4 and 5 were calculated with the ground truth data at hour 4 and 5 on Sunday June 8, 2008. The number of census tracts with the correct prediction on the hot spot development stage and the prediction accuracy was reported in Table 5. The upper two maps in Figure 6 show the predicted hot spot dynamics and the observed hot spot dynamics during hour 4 and 5 (4 - 6am) on Sunday June 8, 2008.

Table 5. Accuracy of status predication for activity hot spots and their developments.

Prediction Item	Census Tracts (out of 194) with Correct Prediction	Prediction Accuracy
Center zone status at hour 5 (5 – 6am)	186	95.9%
Periphery zone status at hour 5 (5 – 6am)	184	94.8%
Hot spot dynamics during hour 4 – hour 5 (4 – 6am)	178	91.8%
Center zone status at hour 13 (1 – 2pm)	180	92.8%
Periphery zone status at hour 13 (1 – 2pm)	169	87.1%
Hot spot dynamics during hour 12 – hour 13 (12pm – 2pm)	171	88.1%

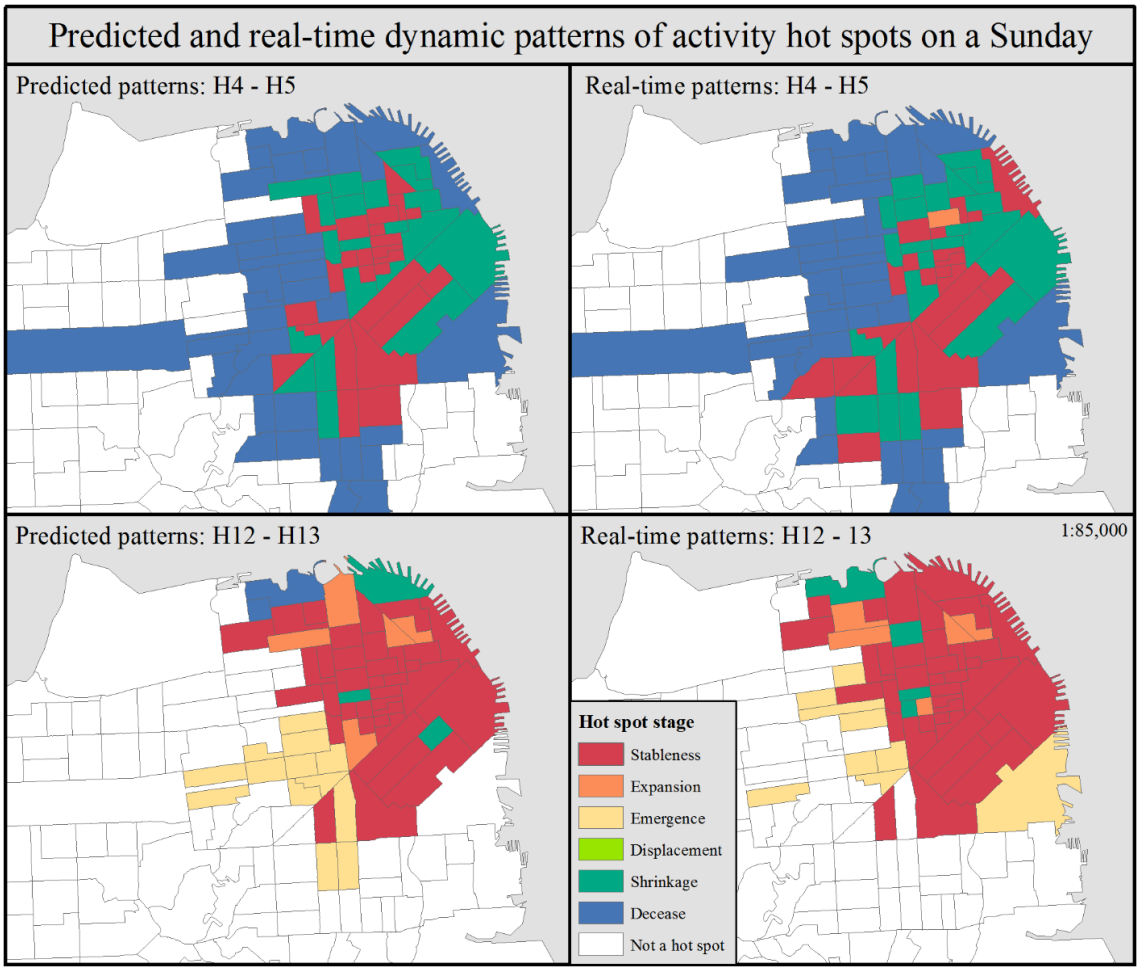


Figure 6. Predicted and real-time dynamic patterns of activity hot spots on a Sunday.

The same analysis was repeated for hour 12 and 13 on Sunday June 8, 2008. The number of census tracts and periphery zones with the correct prediction on hot spot status, the number of census tracts with the correct prediction on the hot spot development stage, and prediction accuracy were reported in Table 5. The predicted and observed hot spot dynamics during hour 12 and 13 were shown in the lower two maps in Figure 6.

The predicted hot spot status during hour 5 was 95.9 percent accurate for center census tracts and 94.8 percent accurate for periphery zones. The prediction accuracy for

hot spot dynamics during hour 4 and 5 was 91.8 percent. The predictions for Sunday early afternoon (hour 12 – hour 13) were reasonably accurate with the rates as 92.8 percent, 87.1 percent, and 88.1 percent respectively. Prediction accuracy was further analyzed with confusion matrix, commission and omission errors in Table 6 – 9. Commission and omission errors are two types of possible errors in predictive or classification models (Anderson, Lew, and Peterson 2003). Commission errors represent the proportion of items that were predicted to be in a category but actually belong to other categories (Anderson, Lew, and Peterson 2003). They are also called false positives, or overprediction. Omission errors represent the proportion of items that belong to a category but were predicted to be in other categories (Anderson, Lew, and Peterson 2003). They are also called false negatives, or underprediction. The relative proportions of these errors are usually presented in a matrix called confusion matrix, or error matrix (Anderson, Lew, and Peterson 2003). Commission and omission errors of the prediction of being a hot spot for both center census tracts and periphery zones were somewhat big (Table 9). This is because the number of hot spots were small during the predicted hour interval (5 – 6 am, Table 6 and 7). Any false prediction would lead to bigger commission and omission errors. For the hot spot development stage prediction, commission and omission errors were big for hot spots in the stage of "inward shrinkage" and "inward expansion" (Table 9). Respectively, there were only two and one observed hot spots in these two stages, one or two false prediction would lead to huge errors (Table 8).

By cross-referencing the predicted patterns and observed patterns in Figure 6, it was found that errors mainly occurred at census tracts located on the edge of the main hot spot cluster. These edge tracts had more complex life cycles than the census tracts located

in the cluster center – they tend to change through different stages of life cycle more frequently (refer to the maps in Figures 4 & 5 for examples). This indicates that their hot spot status may be more sensitive to the specific activities that are going on during a particular time on a particular day, which makes it harder for a highly accurate prediction.

Table 6. Confusion matrix for center census tracts' hot spot status predication accuracy at hour 5.

		Predicted hot spot status		Total
		Hot spot	Not a hot spot	
Observed hot spot status	Hot spot	9	3	12
	Not a hot spot	5	177	182
Total		14	180	194

Table 7. Confusion matrix for periphery zones' hot spot status predication accuracy at hour 5.

		Predicted hot spot status		Total
		Hot spot	Not a hot spot	
Observed hot spot status	Hot spot	38	6	44
	Not a hot spot	4	146	150
Total		42	152	194

Table 8. Confusion matrix for the predication accuracy of hot spot development stages between hour 4 and hour 5.

		Predicted hot spot development stage									Total
		C_D	P_D	O_D	I_S	O_S	O_Stb	P_S	I_E	None	
Observed hot spot development stage	C_D	3	0	0	0	0	0	0	0	0	3
	P_D	0	21	0	0	0	0	2	0	0	23
	O_D	0	0	9	0	1	0	0	0	0	10
	I_S	0	0	0	1	1	0	0	0	0	2
	O_S	0	0	1	1	14	4	0	0	0	20
	O_Stb	0	0	0	1	1	7	0	0	0	9
	P_S	0	3	0	0	0	0	11	0	0	14
	I_E	0	0	0	0	0	0	1	0	0	1
	None	0	0	0	0	0	0	0	0	112	112
Total		3	24	10	3	17	11	14	0	112	194

Note: “C_D” donates “Center Decease”; “P_D” donates “Periphery Decease”; “O_D” donates “Overall Decease”; “I_S” donates “Inward Shrinkage”; “O_S” donates “Outward Shrinkage”; “O_Stb” donates “Overall Stableness”; “P_S” donates “Periphery Stableness”; “I_E” donates “Inward Expansion”; “None” donates “Not a hot spot”.

Table 9. Commission and omission errors of the predication for hot spot status at hour 5 and hot spot development stages during hour 4 and 5.

Prediction item		Commission errors	Omission errors
Center census tract hot spot status at hour 5	Hot spot	0.357	0.25
	Not a hot spot	0.017	0.0275
Periphery zone hot spot status at hour 5	Hot spot	0.095	0.136
	Not a hot spot	0.0395	0.0267
Hot spot development stage during hour 4 - 5	C_D	0	0
	P_D	0.125	0.0870
	O_D	0.1	0.1
	I_S	0.667	0.5
	O_S	0.176	0.3
	O_Stb	0.364	0.222
	P_S	0.214	0.214
	I_E	0	1
	None	0	0

Note: “C_D” donates “Center Decease”; “P_D” donates “Periphery Decease”; “O_D” donates “Overall Decease”; “I_S” donates “Inward Shrinkage”; “O_S” donates “Outward Shrinkage”; “O_Stb” donates “Overall Stableness”; “P_S” donates “Periphery Stableness”; “I_E” donates “Inward Expansion”; “None” donates “Not a hot spot”.

V. INDIVIDUAL ACTIVITY PATTERNS MODELING AND ANALYSIS

This chapter demonstrates the proposed space-time modeling techniques for identifying individual daily T-A patterns. The empirical data contains the GPS trajectory data of two participants in the Microsoft Research Asia GeoLife Project (Zheng et al. 2008, 2009). The pattern discovery and the sensitivity analysis are reported in the third section of the chapter.

Site Description and Data

Beijing is located in northern China. It is the capital city and the second largest city (by urban population) of China. It is the nation's political, cultural, economic, and educational center. Beijing has been the political center of China for about eight centuries. It is known for its palaces, temples, parks, gardens, tombs, walls, and gates. These historical treasures and many universities together made Beijing a center of culture and education. It is also the home of the headquarters of many China's largest state-owned companies and it is a major hub for the country's transportation network.

The Beijing Dataset was collected mostly in Beijing through the Microsoft Research Asia GeoLife Project from April 2007 to September 2009 (Zheng et al. 2008, 2009). It contains the GPS trajectory data of thirty-two volunteers in various periods, from one week to over two years. These volunteers might be drawn from the employees who worked at a research institution in Beijing. This dataset recorded many of the outdoor travels conducted between their daily activities, such as home, work, shopping, dining, sightseeing, hiking, cycling, etc. There were two types of data files for each participant. The GPS trajectory files contained latitude, longitude, elevation, date, and

time. Location information was recorded approximately every two to five seconds. The trip label file contained information about the date, starting time, ending time, and transportation mode of each trip taken during the survey period.

There are some concerns with this dataset. First, the sample size was not adequate to generate prototypes of individual daily T-A patterns. Second, the volunteers may come from the same workplace thus share similar work schedules. The daily T-A patterns discovered from these people may only be representative for this particular workplace. Third, data incompleteness and inconsistency exist in this dataset. Many participants did not track their traveling continuously during the survey period. It was common to find that some trips during certain time periods (i.e. a few hours of a day or a few days of a month) were missing. Various reasons may contribute to this type of data incompleteness. The survey participants might choose not to record some of the trips due to privacy concerns. They might forget to turn on the GPS device timely, or encounter some technical issues with the device, such as running out of battery, etc. Data inconsistency was also found between the trip label files and the GPS trajectory files. For example, some trips were recorded in the trip label file but could not be found in the GPS trajectory files or vice versa. Sometimes, the starting and/or ending time of the same trip does not match between the two files. Fourth, other information about the participants (such as demographic and economic status) and their activities (such as activity types) were absent. Thus, the analysis will not be able to link the discovered daily T-A patterns with personal characteristics. These limitations associated with the Beijing Dataset must be recognized before proceeding to the data analysis.

Data was collected over thirty days for Eighteen out of the thirty-two participants. From these eighteen participants, two (ID: 022 and 031) were selected for the empirical analysis. These two participants lived in the northern area of Beijing when their data was collected. For each participant, the locations of trip origins and destinations were identified by cross-referencing the trajectory files and the trip label file. The trip origins and destinations were then used to create a trip end points file containing information on location, date, time, and transportation mode. Each participant's trip end points were projected and shown on a map. A few clusters of these points could be observed. These clusters represented anchor locations of the participant. Density based clustering algorithm (Ester et al. 1996) was implemented on each participant's trip end points. For each trip end point p_i in a participant's trip end point set, a circle was drawn around the point with a predetermined radius (200 meters in this case study). Then the number of points within the circle was counted. If the number was greater than a threshold value (20 in this case study), then point p_i was considered a core point and all the other points in the circle was considered its neighbors. If core point p_i did not belong to any existing cluster, then p_i and its neighbors formed a new cluster. If p_i belonged to an existing cluster, then p_i and its neighbors joined the existing cluster. Density based clustering algorithm determines the number of clusters automatically, identifies clusters of different shapes, and excludes noise points effectively. During the initial data processing, density based clustering algorithm was run with different search radii and threshold values on each participant's trip end point dataset. The search radius and threshold value that produced the best clustering result was used in the case study. Note that, other values for the search radius and threshold may work better for different point datasets.

A few clusters of trip end points were identified for each participant (e.g., Figure 7). For each cluster, the mean center was used to represent an anchor location for a participant, and a single letter (such as "B") was assigned as the location ID. All trip end points that belong to a cluster were marked with the same location ID. The trip end points that do not belong to any cluster/anchor location were all assigned the letter "A". Thus, letter "A" was not a fixed location as all the other letters. It refered to all non-anchor locations of the participant. Figure 7 shows an example of one participant's trip end points clustering at a few anchor locations. Next, each trip was represented by two letters, such as "BD" (indicating a trip from anchor location "B" to "D") and "BA" (indicating a trip from anchor location "B" to a non-anchor location). All the trips of one participant were then sorted by time. It needs to be noted that many trips start and/or end at anchor locations (e.g., home, workplace, etc.) for each participant, and there is no pre-knowledge about any anchor locations from any participant. All the anchor locations of a participant were found through data processing, including trip end points identification and clustering, as illustrated above.

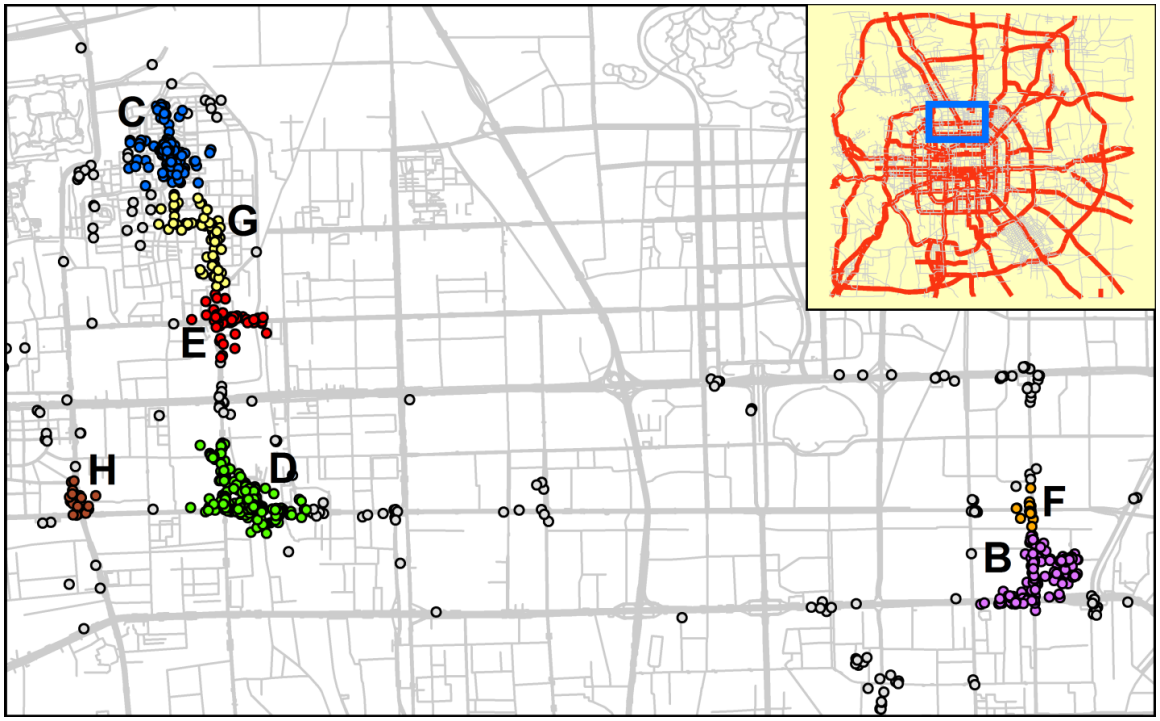


Figure 7. Participant 031's trip end points cluster at several anchor locations.

Methodology

An individual's one-day T-A events was represented as a sequence of trip origins and destinations in the chronological order. These origins and destinations were also the individual's activity locations. By analyzing the similarities between these daily T-A sequences from one individual, one or more representative daily T-A patterns may be found to describe the individual's daily routines.

Constructing Individuals' Daily T-A Sequences

Most individuals take more than one trip a day (e.g., one trip is from home to work and another is from work to home). The origins and destinations of the trips taken by one individual in one day can be lined up chronologically in a sequence. This one-

dimensional sequence records the individual's one-day T-A. Each element in this sequence has only one attribute: the location information for an origin or a destination. For example, assume that John's T-As on one particular day include staying at location "B" from 12:00 am to 7:40 am, traveling from location "B" to location "C" by car and arriving at location "C" at 8:20 am, staying at "C" until 4:30 pm, traveling from "C" back to "B" by car and arriving at location "B" at 5:10 pm, and staying at "B" until the end of the day (11:59 pm). John's T-As during the day can be represented by a one-dimensional sequence as "BCCB". Element "B" indicates both the activity location from 12:00 am to 7:40 am and the origin location of the trip from "B" to "C".

When the origin and destination locations of each trip are paired up and treated as single elements in a T-A sequence, a two-dimensional T-A sequence is formed. This two-dimensional sequence records the trips one individual had taken in one day in the chronological order. Each element in the sequence has two attributes: one for trip origin location and the other for trip destination location. Using John's example, his T-A in that day can be represented by a two-dimensional sequence as "BcCb". Each element (e.g., "Bc") indicates a trip John had taken in the day ("B" is the trip origin location and "c" is the destination location).

When information on time and transportation mode is considered, higher dimensional sequences can be constructed. A three-dimensional sequence records the location, time, and transportation mode of T-A taken during a day in the chronological order. Each element in the sequence has three attributes: trip origin/destination location, trip starting/ending time, and the transportation mode. Note that the second attribute always matches the first attribute. In this case, if the first attribute records the origin of a

trip, the second attribute records the starting time of the trip; if the first attribute records the destination of a trip, the second attribute records the ending time of the trip. Using John's example again, his T-A in that day can be represented by a three-dimensional sequence as "BheCieCqeBre". In element "Bhe", "B" is the trip origin location; "h" is the time leaving "B", "e" is the transportation mode of the trip. Time and transportation modes are coded as lower case letters separately for sequence representation (see Table 10 & 11).

Table 10. Twenty-four lower case letters to represent twenty-four hours of a day.

Letter	Hour of a Day
a	12:00:00 am - 12:59:59 am
b	1:00:00 am - 1:59:59 am
c	2:00:00 am - 2:59:59 am
.....	.....
w	10:00:00 pm - 10:59:59 pm
x	11:00:00 pm - 11:59:59 pm

Table 11. Nine lower case letters to represent twelve different transportation modes.

Letter	Transportation Mode
a	Walk or run
b	Bike
c	Bus
d	Motorcycle
e	Car
f	Taxi
g	Plane
h	Subway, railway, or train
i	Boat

When more information about one trip is included in one element, a five- or higher dimensional sequence can be constructed. The five attributes for one element are: trip origin location, trip destination location, trip starting time, trip ending time, and transportation mode. John's one-day T-A in the example can be represented as a five-dimensional sequence "BchieCbqre". In element "Bchie", "B" records the trip origin location, "c" is the destination location, "h" is the time leaving "B", "i" is the time arriving at "c", and "e" is the transportation mode. There are certainly other ways to form sequences to incorporate different attributes into a sequence. This research uses the above four types of sequences.

Similarity between Corresponding Elements

In order to find an individual's daily routines, similarity between his daily T-A sequences needs to be measured. For this purpose, the similarity between each pair of corresponding elements from two sequences needs to be quantified first. Element similarity can be calculated using similarity matrices consisting of similarity scores between the corresponding attributes of the corresponding elements.

As described in the last section, elements in one-dimensional sequences has only one attribute: the origin/destination location of a trip. The similarity of two corresponding elements from two one-dimensional sequences is the similarity of the two corresponding locations. For each individual, many of these daily trip origin or destination locations are his personal anchor locations (or anchor locations, as illustrated in the site description and data section of this chapter), and the specific geographic coordinates of these locations were calculated in the early processing of his trajectory data (through identifying "stops"

and clustering "stops" into anchor locations, as illustrated in the site description and data section of this chapter). The similarity between each pair of corresponding locations can be measured based on their geographic proximity. Two locations that are closer to each other are considered more similar and their similarity score is higher. This principle ensures a higher similarity score for two daily T-A sequences with higher spatial similarity. For example, a student spent most of Day One in the Liberal Arts Building taking classes and most of Day Two in the Library studying. The Liberal Arts Building and the Library are very close to each other. The T-A sequences of these two days receive a higher similarity score because they have higher spatial similarity. The similarity of two corresponding locations can also be measured based on location types (such as restaurants, grocery stores, etc.) or other characteristics of the locations. However, these types of information are absent from the Beijing Dataset. Thus, location similarity measurement based on distance only was adopted in this research.

Based on the above distance principle, three methods were used to calculate similarity scores between locations and construct location similarity matrices. These methods include inverse distance decay, linear distance decay, and ordinal ranking of distances. In the location similarity matrix based on inverse distance decay, similarity scores are inversely related to the distances between two locations. They are calculated as:

$$S_i = \frac{1}{D_i} \times A_r, \quad i = 1, 2, 3, \dots, n \tag{5}$$

Where S_i is the similarity score for the i th pair of locations, D_i is the Euclidean distance between the two locations in the i th pair, A_r is a constant, and n is the number of paired locations. A_r is set based on properly considering the spatial distribution of a

participant's anchor locations, so that all similarity scores are within the range of 0-10 in this dissertation. This similarity range is pre-set by the software package ClustalTXY, which was used in this research to calculate similarities between T-A sequences. It allows users to define its own personalized measurement for element similarity. However, it requires all element similarity scores to be an integer between 0 and 10. This rule was followed throughout this research. Other rules for element similarity scores may apply when other software packages or methods are used.

As distance between any two anchor locations can be calculated, the similarity score for any two of those locations can be calculated with the above equation. It must be noted that location "A" could be anywhere as it represents all non-anchor locations, thus distance between "A" and any anchor locations cannot be calculated. A score of 0 (the lowest score) is assigned to represent the similarity between location "A" and any anchor locations. A score of 10 (the highest score) is assigned to identical locations. Table 12 shows an example of a location similarity matrix created using inverse distance decay on Participant 031's trajectory data. Locations "B" to "H" were his anchor locations. Location "A" represented all non-anchor locations of his. Figure 8 is a map showing the spatial distribution of his anchor locations.

Table 12. Participant 031's location similarity matrix using inverse distance decay.

Location	A	B	C	D	E	F	G	H
A	10	0	0	0	0	0	0	0
B	---	10	0	0	0	5	0	0
C	---	---	10	1	2	0	3	1
D	---	---	---	10	2	0	1	2
E	---	---	---	---	10	0	4	1
F	---	---	---	---	---	10	0	0
G	---	---	---	---	---	---	10	1
H	---	---	---	---	---	---	---	10

Figure 8. Participant 031's anchor locations.

Another way for constructing location similarity matrix is to use a linear decay function to calculate similarity scores, where the similarity scores are linearly and negatively related to the distances, as illustrated in Equation (6).

$$S_i = 10 - \frac{D_i}{A_l}, i = 0, 1, 2, \dots, n \quad (6)$$

where A_l is a constant set according to the spatial distribution of the anchor locations of a participant, so that all similarity scores are less than 10. Any negative S_i from Equation (6) is reset to 0 to ensure the integrity of similarity scores. Table 13 shows an example of a location similarity matrix created using linear distance decay on Participant 031's trajectory data. Similarity score for each pair of anchor locations was calculated using Equation (6). Similarity scores for non-anchor locations and identical locations were defined following the same rule above.

Table 13. Participant 031's location similarity matrix using linear distance decay.

Location	A	B	C	D	E	F	G	H
A	10	0	0	0	0	0	0	0
B	---	10	0	0	0	5	0	0
C	---	---	10	0	0	0	2	0
D	---	---	---	10	0	0	0	0
E	---	---	---	---	10	0	4	0
F	---	---	---	---	---	10	0	0
G	---	---	---	---	---	---	10	0
H	---	---	---	---	---	---	---	10

Ordinal ranking of distances between locations can be used to construct location similarity matrix as well. Distances between pairs of anchor locations are ordered from the smallest to the largest. The smallest distance receives the highest similarity score (an

integer less than 10, since 10 is for identical locations). The largest distance receives the lowest similarity score (usually a 0). Similarity score decreases as the rank of the distances increases. Table 14 shows an example of a location similarity matrix created using ordinal ranking of distances on Participant 031's trajectory data. Similarity scores for anchor locations were calculated following the above rules. Note that, in this case, the smallest distance (between “B” and “F”) which ranked the first was assigned a similarity score of 5; the anchor location pairs whose distance ranked from the second smallest to the fifth was assigned a score from 4 to 1; the rest of them all received a score of 0. This score assignment was somewhat conservative. As participants did not provide any information about their anchor locations, location types were unknown in this case study. Even two locations seemed to be very close (as “B” and “F”), they could be completely different types and irrelevant (such as girlfriend’s house and a restaurant). Thus, a conservative score of 5 was assigned to the location pair with the smallest distance. Other score ranges may be more appropriate in other case studies. Similarity scores for non-anchor locations and identical locations were assigned in the same way as the last two methods.

Table 14. Participant 031's location similarity matrix using ordinal ranking of distances.

Location	A	B	C	D	E	F	G	H
A	10	0	0	0	0	0	0	0
B	---	10	0	0	0	5	0	0
C	---	---	10	0	0	0	3	0
D	---	---	---	10	1	0	0	2
E	---	---	---	---	10	0	4	0
F	---	---	---	---	---	10	0	0
G	---	---	---	---	---	---	10	0
H	---	---	---	---	---	---	---	10

A few concerns need to be addressed here. First, there are other ways to generate similarity scores based on distances. This research used the above three methods and tested whether they had an impact on the individual daily T-A pattern identification. Second, since location similarity was measured based on the physical distance between two locations (areal units were not used), MAUP is not a concern in this measurement. Third, each individual had a unique set of anchor locations, which were used to generate his location similarity matrix. Thus, each individual had its own location similarity matrices.

Assume Participant 031's two daily one-dimensional T-A sequences are: "CDDC" (as S1) and "CEEC" (as S2). Figure 9 shows the alignment of the two sequences. The two circled locations (location "D" from S1 and Location "E" from S2)

are a pair of corresponding elements. The similarity of the two corresponding elements is the similarity of location "D" and "E", which can be found from Participant 031's location similarity matrices (Table 12-14). When location similarity matrix based on inverse distance decay (Table 12) is used, the similarity score for Location "D" and "E" is 2. When location similarity matrix based on linear distance decay (Table 13) is used, the similarity score for the two locations is 0. When location similarity matrix based on ordinal ranking of distances (Table 14) is used, the similarity score is 1. Considering 0 and 10 were set to be the lowest and highest similarity scores, location "D" and "E" appear to have very limited spatial similarity no matter which location similarity matrix is used.

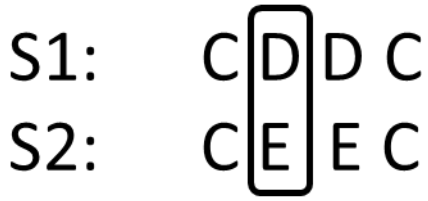


Figure 9. The alignment of Participant 031's two daily one-dimensional T-A sequences.

Elements in two-dimensional sequences has two attributes: trip origin location and trip destination location. The similarity score between two corresponding two-dimensional elements can be calculated in two steps. First, calculate the similarity score for the corresponding attributes separately. Second, multiply the two attribute similarity scores together and standardize the product by dividing it by ten. Since both attributes contain locations only, location similarity matrices defined above can be used to calculate similarity scores for each attribute.

Assume Participant 031's two daily two-dimensional T-A sequences are: "CdDc" (as S1) and "CeEc" (as S2). Figure 10 shows the alignment between the two sequences. The two circled elements ("Cd" from S1 and "Ce" from S2) are a pair of corresponding elements. The similarity of them can be calculated as shown in Figure 11.

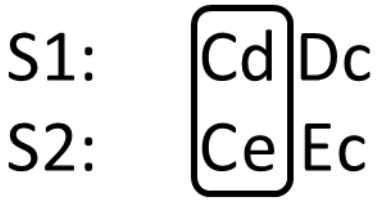


Figure 10. The alignment of Participant 031's two daily two-dimensional T-A sequences.

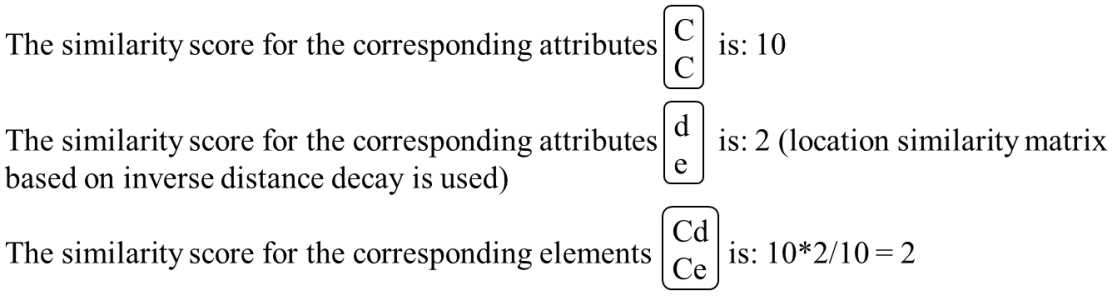


Figure 11. The calculation of similarity score for two corresponding two-dimensional elements.

Elements in three-dimensional sequences has three attributes: trip origin/destination location, trip starting/ending time, and the transportation mode. The similarity score between two corresponding three-dimensional elements can be calculated following these steps. First, calculate the similarity score for each corresponding attribute separately. Second, multiply the three attribute similarity scores together and standardize

the product through dividing it by one hundred. Location similarity matrices can be used to calculate similarity scores for the first attribute, since it contains origin/destination locations. Similarity score for the second attribute (contains trip starting/ending time) can be calculated accordingly on a scale of 0-10. A score of 10 is assigned when the two times share the same value. Otherwise, a score of 5 is assigned. For the third attribute (contains transportation mode), a similarity score of 10 is assigned if the two transportation modes are the same. Otherwise, a score of 8 is assigned. A score of 5 and 8 for discrepancies in time and transportation mode attributes ensures that time difference was considered more significant for T-A patterns than transportation mode in this research. Note that other score matrices can be used to quantify similarities regarding time and transportation mode. For example, two transportation modes may be considered similar if the travel speeds tend to be similar or if they result in similar travel times for a certain trip.

Assume Participant 031's two daily three-dimensional T-A sequences are: "CibDibDrbCrb" (as S1) and "CicEicErcCrc" (as S2). Figure 12 shows the alignment between the two sequences. The two circled elements ("Dib" from S1 and "Eic" from S2) are a pair of corresponding elements. The similarity of them can be calculated following the steps in Figure 13.

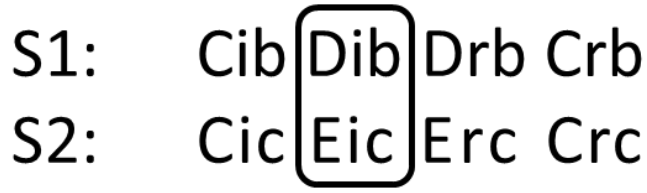


Figure 12. The alignment of Participant 031's two daily three-dimensional T-A sequences.

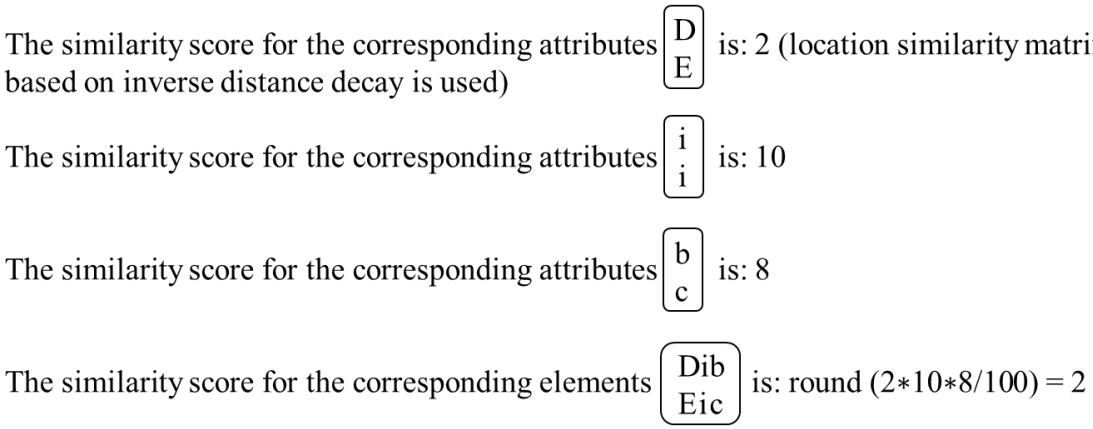


Figure 13. The calculation of similarity score for two corresponding three-dimensional elements.

Elements in five-dimensional sequences has five attributes: trip origin location, trip destination location, trip starting time, trip ending time, and transportation mode. The similarity score between two corresponding five-dimensional elements can be calculated following very similar steps as three-dimensional elements. First, calculate the similarity score for each corresponding attribute separately. Second, multiply the five attribute similarity scores together and standardize the product through dividing it by 10,000. Similarity scores for the first and second attribute can be calculated using location

similarity matrices. Similarity scores for the third and fourth attribute can be calculated following the same rules defined above for the time attribute. Similarity scores for the fifth attribute can be calculated in the same way as defined above for the transportation mode attribute.

Assume Participant 031's two daily five-dimensional T-A sequences are: "CdiibDcrrb" (as S1) and "CeiicEcrrc" (as S2). Figure 14 shows the alignment of the two sequences. The two circled elements ("Cdiib" from S1 and "Ceiic" from S2) are two corresponding elements. Their similarity can be calculated as shown in Figure 15.

Figure 14. The alignment of Participant 031's two daily five-dimensional T-A sequences.

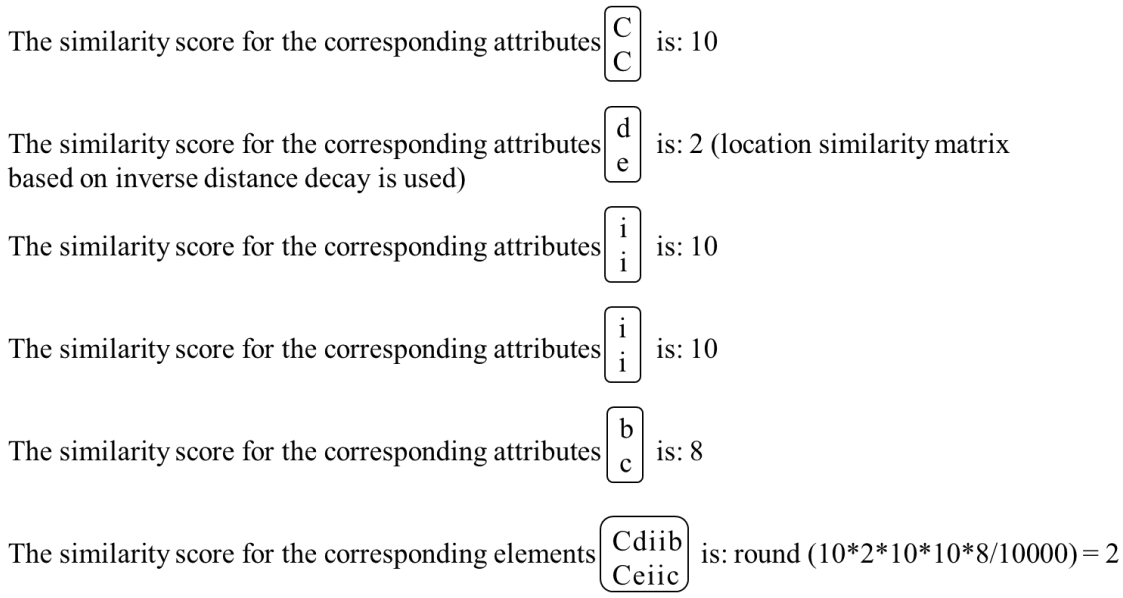


Figure 15. The calculation of similarity score for two corresponding five-dimensional elements.

Similarity between T-A Sequences, Sequence Grouping, and the Representative Sequences

To calculate similarity, two daily T-A sequences need to be aligned first. The optimal alignment between two sequences can be achieved by transforming one into the other with the smallest number of character edit operations (including insertion, deletion, and substitution) (Wilson 1998; Kwan, Xiao, and Ding 2014). This smallest number of character edit operations required to transform one sequence into the other is called Levenshtein Distance (LD), or edit distance (Navarro 2001). The similarity between two daily T-A sequences can be measured as a combined similarity score, which is the sum of the similarity scores for all aligned pairs of elements minus the penalties for gap opening and extension operations in order to equal the two sequences (ClustalTXY). Gap Opening Penalty (GOP) is the cost of adding a new gap of any length to a sequence in order to

align two sequences (ClustalTXY). Gap Extension Penalty (GEP) is the cost of each item in a gap (ClustalTXY). The suggested value for GOP is eight and GEP is three by the ClustalTXY software, so that the final alignment would resemble the general appearance of the input sequences to some extent (ClustalTXY). The similarity score for a pair of aligned elements (corresponding elements) can be calculated following the rules defined in the previous section. A high similarity score indicates a high level of similarity between the two sequences.

In the case of Participant 031, assume his two daily one-dimensional T-A sequences are "CEEDDC" (as S1) and "CDDC" (as S2). To obtain the maximum similarity (optimal alignment) between the two sequences, a gap with the length of two was introduced in Sequence S2 (Figure 16). Figure 16 shows the alignment between the two sequences. The calculation of the similarity score for Sequence S1 and S2 is shown in Figure 17.

Figure 16. The alignment of Sequence S1 and S2.

The similarity score S for each pair of aligned elements: $S\left(\begin{matrix}C\\C\end{matrix}\right) = 10$,

$S\left(\begin{matrix}D\\D\end{matrix}\right) = 10$, $S\left(\begin{matrix}D\\D\end{matrix}\right) = 10$, $S\left(\begin{matrix}C\\C\end{matrix}\right) = 10$,

GOP is 8 and GEP is 3 (these are the default setting in ClustalTXY software),

The similarity score for Sequence S1 and S2 is: 10+10+10+10-(8+2*3) = 26

Figure 17. Similarity score calculation for Sequence S1 and S2.

Based on the similarity scores calculated for any pair of sequences, all sequences can be clustered into several groups using the Neighbor-Joining Method (Saitou and Nei 1987). Sequences within a group are more similar to each other than to sequences across groups. Representative sequences, such as the consensus sequence and median sequence, can be identified for each group. A consensus sequence is comprised by the most frequent element at each position from the group of sequences (Wilson 2008). A median sequence is a sequence from the group which has the minimum sum of differences to all other sequences in the group (Wilson 2008). These representative sequences reveal the general pattern of the sequences in a group. It is important to note that a median sequence can always be identified for a group of sequences while a consensus sequence presents only when repetitive patterns exist across the sequences within a group.

A sequence alignment software package called ClustalTXY was used in this research for calculating similarity scores between individual daily T-A sequences, classifying them into groups, and generating representative sequences. The similarity

measurement between corresponding elements defined in the previous section was used as an input file for personalized alignment.

Findings and Discussion

For the two selected participants (ID "022" and "031"), each of their trajectory data was prepared following the site description and data section of this chapter. Each participant's anchor locations were identified and labeled. Each of their trips was represented with trip origin and destination location IDs. For each participant, one-, two-, three-, and five-dimensional daily T-A sequences were generated following the first subsection of the methodology section of this chapter. The daily T-A sequences for each participant were further divided into two groups: weekday sequences and weekend sequences. Thus, each participant had eight sets of sequences: a weekday set and a weekend set for each of the four different dimensional sequences.

Following the second subsection of the methodology section of this chapter, three types of location similarity matrices were generated for each participant. Similarity measurement for corresponding elements in the one-, two-, three-, and five-dimensional sequences (as defined in the second subsection of the methodology section of this chapter) was calculated and uploaded to ClustalTXY software. Sequence alignment was performed for each set of sequences for each participant using the ClustalTXY software. Representative sequences were identified and reported. The analysis results of the two participants' daily T-A patterns are reported in the rest of this section, as well as sensitivity analysis of pattern identification based on the different location similarity matrices and sequence dimensions.

Individual Daily T-A Pattern Discovery and Sequence Dimension

Participant 022's anchor locations are shown in Figure 18. Sequence alignment results for his one-, two-, three-, and five-dimensional weekday and weekend T-A sequences are reported in Table 15. Within each sequence set, T-A sequences were clustered into several groups.

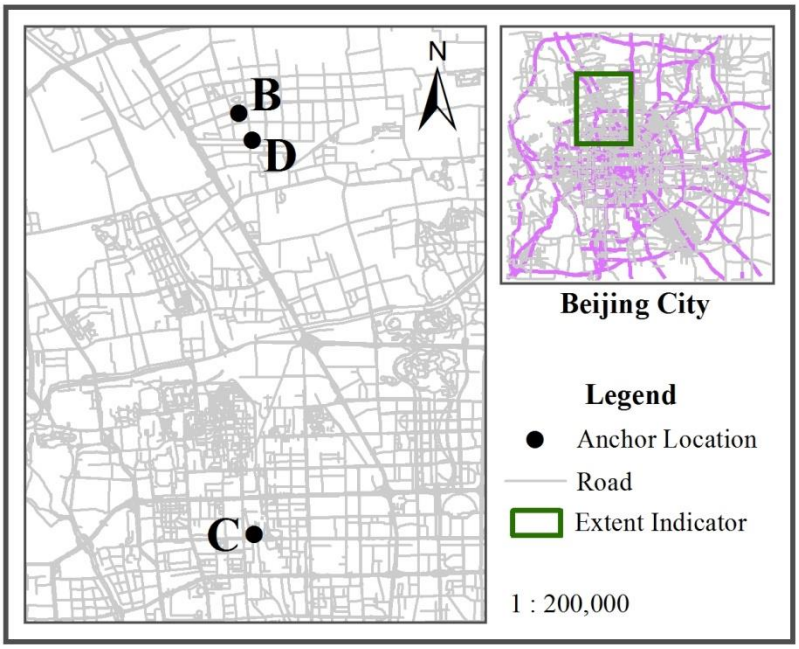


Figure 18. Participant 022's anchor locations.

Table 15. Grouping patterns and representative sequences of Participant 022's weekday and weekend sequence sets.

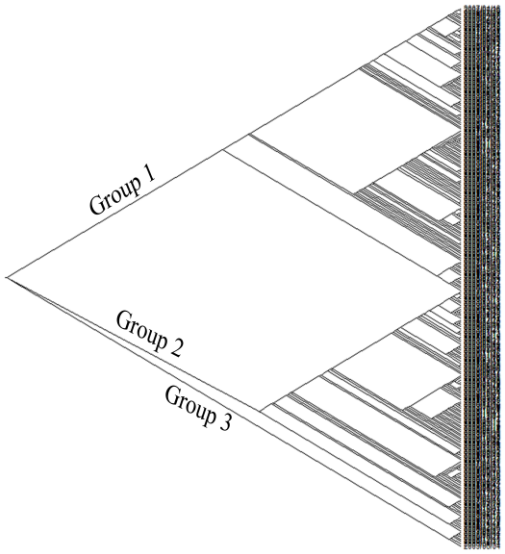
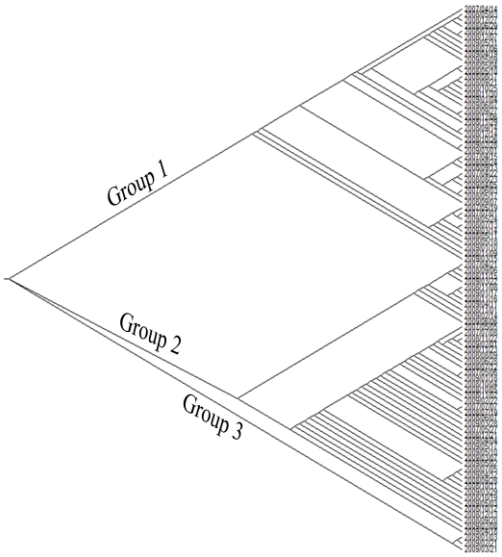
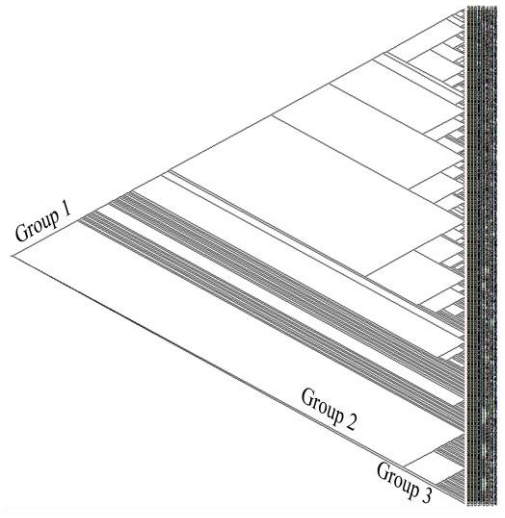
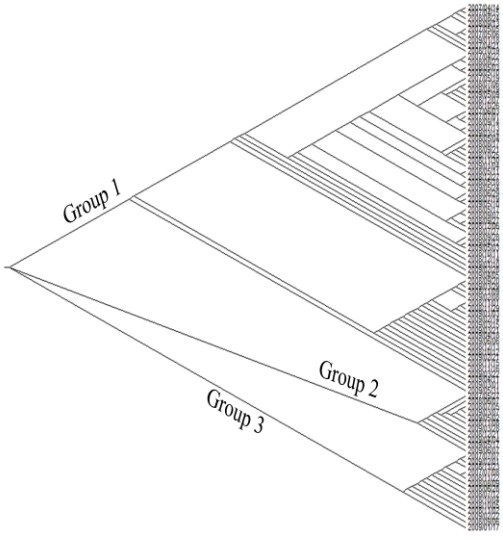
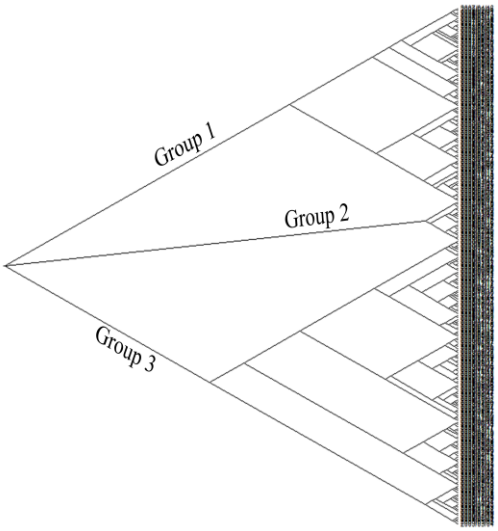
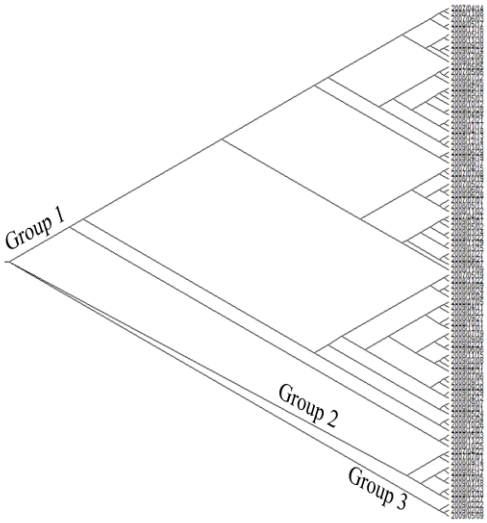
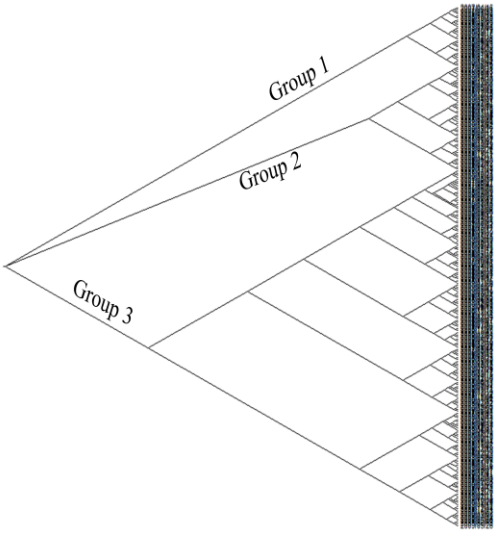
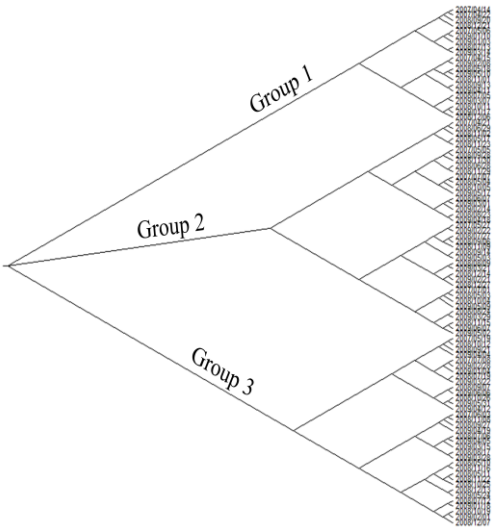
Sequence Dimension	Weekday			Weekend		
	Group	Consensus Sequence	Median Sequence	Group	Consensus Sequence	Median Sequence
One	1 (n=132)	BCCA	BCCAAC	1 (n=45)	BCCB	BCCB
	2 (n=112)	BCCB	BCCB	2 (n=48)	AAAAAA	BAAAAAAAAA
	3 (n=7)	BCCB	BCCB	3 (n=3)	BAAB	BAAB
						
Two	1 (n=215)	BcCb	BcCb	1 (n=71)	BaAa	BaAaAaAb
	2 (n=35)	BcCb	BcCb	2 (n=11)	AaAaAb	AaAaAaAb
	3 (n=1)	BcCb	BcCb	3 (n=14)	AaAaAa	AaAaAaAaAa
						

Table 15. (Continued) Grouping patterns and representative sequences of Participant 022's weekday and weekend sequence sets.

Three	1 (n=94)	DjhCjh	BiaDiaDihCjhCshBth	1 (n=83)	----------	BneAoeAoaAoaAoeBqe
	2 (n=19)	BjeCjeAte	BjeCjeAwfBwf	2 (n=10)	----------	AjeAjeAjaAmaAmeBne
	3 (n=138)	BjeCke	BjhCkhCseAseAueBue	3 (n=3)	AifAifAmgAng	AifAifAmgAng
						
Five	1 (n=29)	----------	BcjkeAbwxf	1 (n=21)	----------	BanoeAbqre
	2 (n=50)	----------	BcjkeCbuve	2 (n=40)	----------	AajoeAapue
	3 (n=172)	----------	BcjkhCbvve	3 (n=35)	----------	AaileAalnaAannneAbope
						

On weekdays, Participant 022's one-dimensional sequences were clustered into three groups (Table 15). Since Group 3 was very small and had the same representative sequences with Group 2, Group 2 and 3 were combined into a new Group 2 with a total of 119 sequences (47 percent of all sequences in this set). The consensus sequence of the new Group 2 showed there was a general pattern for all the sequences in this group as "BCCB". This suggested that a typical daily T-A pattern for Participant 022 was from anchor location "B" to "C" and then from "C" back to "B". It seems like a simple work day T-A pattern: from home to workplace and then back home. The consensus sequence of Group 1 showed a little different daily T-A pattern of this participant: from anchor location "B" to "C", and then from "C" to a non-anchor location "A". This indicated that Participant 022 went to one or more non-anchor locations after work before going back home. It could be picking up kids from school, a grocery store trip, or other life maintenance activities. A lack of a general pattern for after work activities might be the reason that the consensus sequence did not contain the trips leading the participant back home. For example, the participant may have only one after work activity on some days and more on other days. Further, some of these activities may be conducted at anchor locations, while others are not. Group 1 sequences represented a complex work day T-A pattern: from home to workplace, then some after work activities before going back home. In a summary, two daily T-A patterns were discovered from Participant 022's one-dimensional weekday sequences: one was simply home-work-home, the other contained after work activities and appeared to be complex.

Participant 022's two-dimensional weekday sequences were also clustered into three groups (Table 15). All three groups had the same representative sequences, and

only one daily T-A pattern was discovered: from anchor location "B" to "C" and then from "C" back to "B". This is the same with one of the daily T-A patterns discovered from his one-dimensional weekday sequences.

For three-dimensional sequences, as time and transportation mode information were added into the sequence, it became challenging to identify a general pattern. This echoed Hanson and Huff's study in 1988, in which they found that the less complex the daily T-A model was (such as a one-dimensional sequence), the more repetition might be identified; the more complex the daily T-A model was (such as a three-dimensional sequence), the less repetition could be found in the sequences (Hanson and Huff 1988). The three-dimensional weekday sequences were also clustered into three groups. Group 1 contained 37 percent of all sequences. The repetitive pattern showed from the consensus sequence contained only one trip, which was from anchor location "D" to "C" between 9 and 10am in the morning by subway. This should be a part of the home-work trip of the participant. An explanation for the absence of a work-home trip in the repetitive pattern may be that the time and/or transportation mode of this trip varied from day to day, or there were other activities conducted after work at different locations during different times. Thus, there was no consent on this part. Using the transportation mode information, we can infer that anchor location "D" may be a subway station. Note that a short trip from the participant's home "B" (inferred from the one- and two-dimensional sequences) to "D" was missing from the repetitive pattern. This could be caused by errors /offsets in GPS data or a generalization error when trip end points were assigned to anchor locations. Group 2 contained only 8 percent of all the sequences. Its consensus sequence showed a trip from anchor location "B" to "C" between 9 and 10 am by car.

Again, the trip from work back home was not in the repetitive pattern and the same explanation from above applies here too. Group 3 contained 55 percent of all the sequences and the general pattern revealed from the consensus sequence contained only the trip from anchor location "B" to "C" between 9 and 11 am by car. The results of three-dimensional weekday sequences suggested that the morning trip from home to work was the only identifiable representative trip when location, time, and transportation mode were all considered, although on some days the participant took subway and on other days he drove a car. The work-home trip varied significantly in time and transportation mode from day to day. The activities after work varied greatly in number, time, and location, making it impossible to identify a general pattern.

For five-dimensional weekday sequences, no general T-A patterns were found (Table 15). The median sequences seemed to suggest a common daily pattern from anchor location "B" to "C" and then from "C" back to "B" with slightly different times and transportation modes. However, this pattern was not repetitive enough to appear as the consensus sequence. This result once again conformed Hanson and Huff's finding in1988 (Hanson and Huff 1988).

Comparing the daily T-A patterns found from Participant 022's one-, two-, three-, and five-dimensional weekday sequences, we can conclude that two daily T-A patterns existed for this participant: one was from home to work and then from work back home, the other was from home to work and then to some after-work activities then back home. The home-work trips were conducted during consistent hours every morning by either car or subway, while the work-home trips were conducted at different times by different transportation modes from day to day, with after-work activities varying in number, time,

and location. Participant 022's weekday T-A behaviors appeared to be complex. Using the different dimensional sequences for T-A pattern analysis helped gain a comprehensive profile of his daily T-A.

On weekends, one-dimensional sequences were clustered into three groups (Table 15). The consensus sequence of Group 1 (contained forty-five sequences, 47 percent of all sequences in this set) indicated the same home-work-home T-A pattern as found for the one-dimensional weekday sequences. It means this participant spent nearly half of his weekends working. The consensus sequence of Group 2 indicated that the participant visited several non-anchor locations on a weekend day. These locations could be a retail store, a market, a friend's home, a park, a restaurant, etc. These places were not regularly visited by this participant. Group 3 of this sequence set was too small, thus its consensus sequence does not indicate a general T-A pattern of the participant.

Although this participant's two-dimensional weekend sequences were clustered into three groups, the consensus sequences of all these groups suggested the same T-A pattern: visiting several non-anchor locations on a weekend day (Table 15). For three- and five-dimensional weekend sequences, no general pattern was found. The only exception was Group 3 sequences of the three-dimensional weekend sequence set. However, this group was too small. Its consensus sequence does not indicate a general pattern. The reason for a lack of general T-A patterns for weekend may be that weekend activities tend to vary significantly from one weekend day to another in number, location, time, and transportation mode.

There seemed to be an inconsistency in the weekend T-A patterns discovered from different dimensional sequences. A working pattern was found from one-

dimensional sequences but not from the higher dimensional sequences. The reasons may be that time and transportation mode varied significantly on weekends for the home-work and work-home trips. Thus, including these information into analyses only added more "noise" and reduced the chances for identifying commonality. Furthermore, the computation complexity increased rapidly as the sequence dimension increased, which might have challenged the effectiveness of the software used in this research.

Individual Daily T-A Pattern Discovery and Location Similarity Matrix

Participant 031's anchor locations are shown in Figure 8. Sequence alignment results for his one-, two-, and three-dimensional weekday sequences using different location similarity matrices were reported in Table 16. For all three sequence dimensions, no matter which location similarity matrix was used, the sequence grouping patterns were the same: the same number of groups and the same number of sequences in each group. Further, consensus and median sequences were mostly the same with minimum differences across all three location similarity matrices and all three sequence dimensions (Table 16).

Table 16. Grouping patterns and representative sequences of Participant 031's weekday sequence sets using different location similarity matrices.

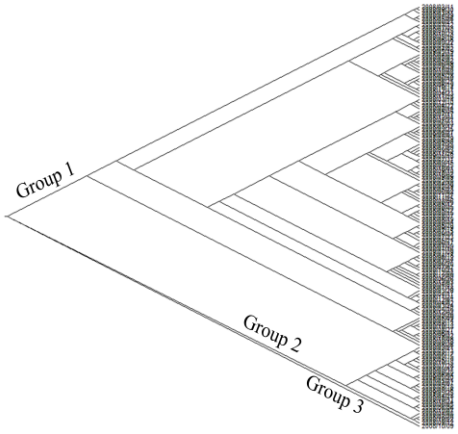
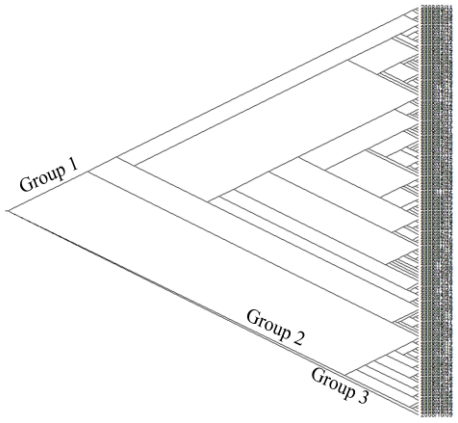
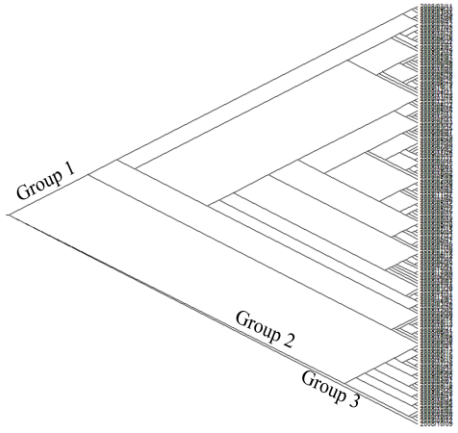
Sequence Dimension	Location Similarity Matrix	Group	Consensus Sequence	Median Sequence	Grouping Structure
One	Inverse Distance Decay	1 (n=114)	CDD	DDDAA DDDDD	
		2 (n=26)	DCCDDD	DDDCC DDD	
		3 (n=1)	DCCD	DCCD	
	Linear Distance Decay	1 (n=114)	CDD	DDDAA DDDDD	
		2 (n=26)	DCCDDD	DDDCC DDD	
		3 (n=1)	DCCD	DCCD	
	Ordinal Ranking of Distances	1 (n=114)	DD	DDDAA DDDDD	
		2 (n=26)	DCCDDD	DDDCC DDD	
		3 (n=1)	DCCD	DCCD	

Table 16. (Continued) Grouping patterns and representative sequences of Participant 031's weekday sequence sets using different location similarity matrices.

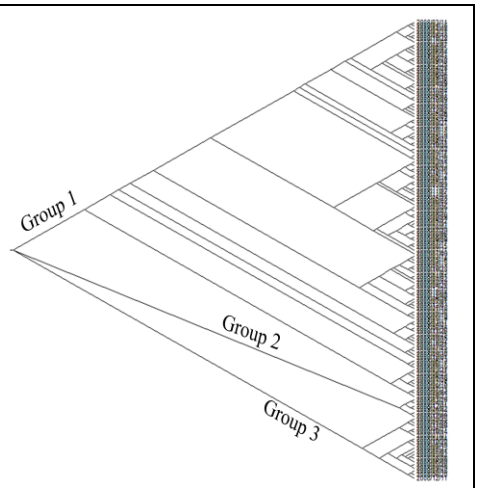
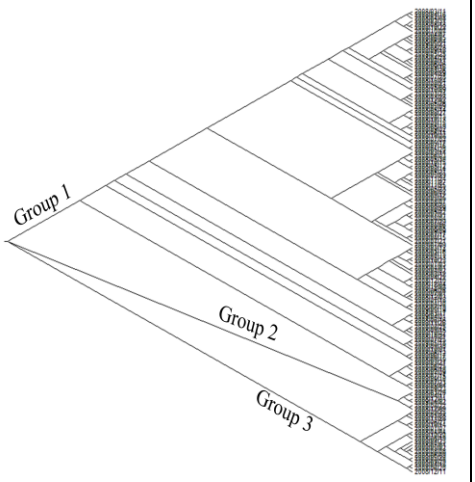
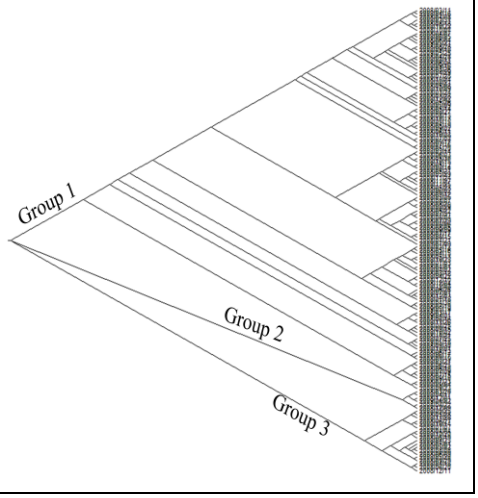
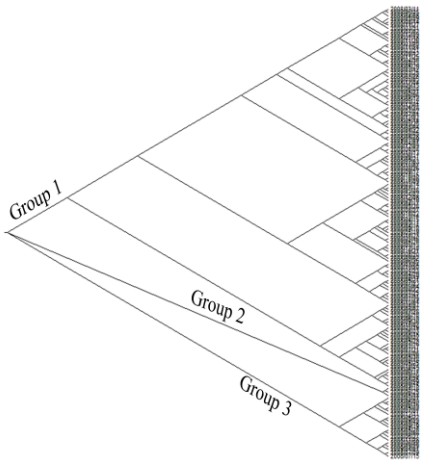
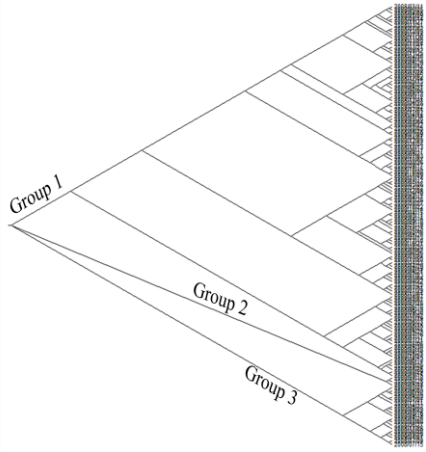
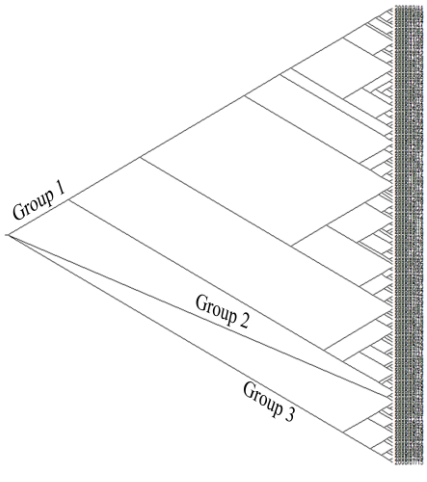
Two	Inverse Distance Decay	1 (n=116)	CdDd	DcCdDd	
		2 (n=6)	EdDd	EdDd	
		3 (n=19)	Aa	AaAaAa AaAaAa BbBa	
	Linear Distance Decay	1 (n=116)	CdDd	DcCdDd Dd	
		2 (n=6)	EdDd	EdDd	
		3 (n=19)	Aa	AaAaAa AaAaAa BbBa	
	Ordinal Ranking of Distances	1 (n=116)	DcCdDd	DcCdDd Dd	
		2 (n=6)	EdDd	EdDd	
		3 (n=19)	Aa	AaAaAa AaAaAa BbBa	

Table 16. (Continued) Grouping patterns and representative sequences of Participant 031's weekday sequence sets using different location similarity matrices.

Three	Inverse Distance Decay	1 (n=119)	Dkb	DabCabCjb DkbDkaDka	
		2 (n=3)	AifGifGiaGiaGvbDvb	AofBofBoaBoaBwfDxf	
		3 (n=19)	----------	AjaAnaAqaAqaAqcAqcAqaAqaAshAwhAwaAxaBxaBxaBxfAxf	
	Linear Distance Decay	1 (n=119)	Dkb	DabCabCjb DkbDkaDka	
		2 (n=3)	GiaGiaGmaEmaGvbDvb	AofBofBoaBoaBwfDxf	
		3 (n=19)	----------	AjaAnaAqaAqaAqcAqcAqaAqaAshAwhAwaAxaBxaBxaBxfAxf	
	Ordinal Ranking of Distances	1 (n=119)	Dkb	DabCabCjb DkbDkaDka	
		2 (n=3)	GiaGiaGmaEmaGvbDvb	AofBofBoaBoaBwfDxf	
		3 (n=19)	----------	AjaAnaAqaAqaAqcAqcAqaAqaAshAwhAwaAxaBxaBxaBxfAxf	

The results from Participant 031's weekday one-dimensional sequences showed that, no matter which location similarity matrix was used, Group 3 contained only one sequence and it corresponded well with the consensus sequence of Group 2. Thus Group 3 was combined with Group 2 (Table 16). Group 1 contained about 20 percent of all sequences with a consensus sequence of "DCCDDD". This indicated that, on about one in five weekdays, the participant took a round trip between anchor location "D" and "C". These two locations might be the participant's home and workplace. The trip "DD" after the round trip might be resulted from multiple trip end locations being assigned to the same anchor location "D". For example, the participant travels from anchor location "C" to "D" by bus, and then walk from the bus stop "D" to home in a short trip. In this case, the bus stop and home may both be marked as anchor location "D". This is related to the algorithms of clustering trip end points and identifying anchor locations. The combined Group 2 sequences shared a consensus sequence of "CDD" or "DD". It means that, on the other weekdays, the only T-A in common was a trip from anchor location "C" to "D" or a very short trip around anchor location "D". Overall, there was only one general daily T-A pattern for Participant 031, and this pattern only accounted for one fifth of all his weekdays, no matter which location similarity matrix was used. On the rest of the weekdays, his T-As were highly irregular, thus no complete daily pattern was identified. Location similarity matrices appeared to have limited impact on the daily T-A patterns identified from Participant 031’s one-dimensional weekday sequences.

For two-dimensional weekday sequences, no matter which location similarity matrix was used, some consistent results showed. First, all sequences were clustered into three groups and each group had the same number of sequences (Table 16). Second,

Group 2 contained only six sequences (4 percent of all sequences in this set) and the common T-A among them was a trip from anchor location "E" to "D". Last, Group 3 contained nineteen sequences (13 percent of all) and its consensus sequence was a trip between two non-anchor locations. However, different location similarity matrices led to the identification of different daily T-A patterns. When inverse distance decay and linear distance decay matrices were used, the general T-A pattern revealed from Group 1's consensus sequence contained only one trip from anchor location "C" to "D" ("Dd" was a very short trip around anchor location "D" thus could be ignored as explained above). No complete daily round-trip T-A pattern was discovered, indicating that the participant's daily T-As were highly irregular. However, when ordinal ranking of distances matrix was used, a complete daily pattern was discovered, which was from anchor location "D" to "C" then from "C" back to "D" (again "Dd" can be ignored). This seemed like a home-work round trip and it accounted for over 80 percent of his work days. Therefore, the selection of location similarity matrices had a certain impact on the patterns identified.

For three-dimensional weekday sequences, the sequence grouping patterns and consensus sequences were mostly the same except for minor differences among consensus sequences of Group 2 (Table 16). However, Group 2 contained only three sequences (2 percent of all), thus patterns revealed from the consensus sequence could not be considered as a general pattern for all three-dimensional weekday sequences. Thus, any difference across Group 2 consensus sequences could be ignored. There was no general daily T-A pattern detected from three-dimensional weekday sequences regardless of location similarity matrix used.

Overall, Participant 031's daily T-As were found highly irregular. Location similarity matrix was found to have a certain level of impact on the T-A patterns discovered. Moreover, sequence dimension might also have led to differences in the T-A patterns discovered. As sequence dimension increased, it became more challenging and complex to identify general patterns. This corresponded well with the findings by Hanson and Huff (1988).

VI. CONCLUSIONS AND FUTURE STUDIES

Conclusions

This research proposed new space-time modeling techniques for discovering urban population daily collective activity patterns and individual T-A patterns using GPS trajectory data. Collective activity patterns were revealed by the locations and dynamics of urban activity hot spots. GPS trajectory data of taxi cabs from San Francisco were used to extract urban population activity instances. By identifying and tracing the development processes of activity hot spots on a Tuesday and a Saturday, collective activity patterns on a weekday were compared with that on a weekend day. The results showed that weekday activities started early in the morning and ended around midnight while weekend activities started in the late morning and lasted until the next morning. Activity hot spots constructed based on historical average data were used as a prediction for up-coming collective activity patterns. Error analysis using confusion matrix, commission and omission errors showed that the prediction was reasonably accurate. These results indicated that the proposed modeling technique could effectively identify temporal variations in urban population collective activity patterns and the future collective activity patterns could be predicted at a reasonable accuracy rate. Individual daily T-A were modeled as daily T-A sequences. By grouping similar sequences and extracting representative sequences, individual daily T-A patterns were revealed. GPS trajectory data for two participants from the Microsoft Research Asia GeoLife Project was used for case studies. The results showed that one participant had two equally important daily T-A patterns and the other participant had no apparent patterns. The sensitivity analysis suggested that sequence dimension had an impact on the general daily T-A patterns

discovered: it became harder to find general patterns as sequence dimension increased. However, sequences of different dimensions together helped to construct an individual's daily T-A profile. Different location similarity matrices may have a certain level of impact on the general T-A pattern discovered. These results indicated that the proposed modeling technique could effectively identify a certain number of (including 0) representative daily T-A sequences of an individual, and it was sensitive to the dimension of daily T-A sequences and the location similarity measurements at a certain level. As the existence of multiple or none representative daily T-A sequences of an individual indicates day-to-day variations of the individual's daily T-A patterns, the proposed modeling technique could effectively identify temporal variations in individual daily T-A patterns.

Connections with Urban Studies

Information and communication technologies enabled the collection of "big data", such as GPS trajectory data, smart phone records, social media posts, etc. Many recent studies in urban geography explored the potential of such data in improving urban planning and urban management. As spatial and temporal information can both be derived from such data, these studies provide us a dynamic understanding of urban lifestyles rather than the traditional static views. Cities empowered with real-time knowledge and intelligence for better decision making and management by digital technology are regarded as "smart cities" (Steenbruggen, Tranos, and Nijkamp 2015). This section illustrates how the modeling and analysis of collective and individual T-A patterns can support the establishment of "smart cities".

Identifying urban functional regions (e.g., residential, commercial, business, etc.) and land use types is essential to urban planning and land use management. Filion (2000) defined urban "monofunctional areas" and "multifunctional areas" based on activities occurred within a 24-hour span. "Monofunctional areas" are areas with a single land-use type and only active for one period of time during a day (e.g., suburban residential areas, industrial areas, etc.), while "multifunctional areas" are areas with mixed land-use types and active for multiple periods of times during a day (e.g., downtown office, shopping, and entertainment areas, Bianchini 1995). This concept has been used in urban planning and development strategies to help stimulate the economy of a city (Lovatt and O'Connor 1995). Recent studies extended the types of activities in this concept from physical only to social media and mobile phone activities (Becker et al. 2011; Zhan, Ukkusuri, and Zhu 2014; Crooks et al. 2015; Dunkel 2015; Wang et al. 2016; Chen et al. 2017; Gao, Janowicz, and Couclelis 2017). These studies also used the differences in activity patterns in a daily cycle, a weekly cycle, a monthly cycle, and seasonal cycles in identifying different functional regions and land use types. Urban population collective level T-A pattern analysis across space and over time help us understand how people use different types of urban space over different temporal cycles; and the individual level analysis helps explain how friends and random strangers end up sharing the same public space. These types of analysis can in turn support more accurate classification of urban functional regions and land use types.

Urban segregation is another important topic in urban planning and development. Segregations can be cultural (e.g., ethnic, language, religion, etc.) or socio-economic (e.g., income, education, professions, etc.) within urban areas. Many studies examined

the issue by focusing on but not limited to inequalities in access to groceries, healthy food, green space, and other urban amenities. Others focus on political recruitment and student performance among school districts. However, these approaches on segregation are mostly place-based. Recent studies took on a new approach, people-based approach, by examining individuals' T-A spaces using GPS or mobile phone data, and relating the differences with their cultural and socio-economic characteristics or urban neighborhoods (Jarv et al. 2015; Shelton, Poorthuis, and Zook 2015; Xu et al. 2015). The differences in individuals' T-A spaces include number of activity locations, the spatial distribution of activity locations, and the spatial extent of activity spaces. Collective and individual level T-A pattern analysis helps to reveal the differences in individual and group T-A spaces and explore how these differences change as various observation period is adopted.

Identifying spatial interaction (movement) patterns within urban areas is essential for travel demand modeling, transportation planning, and traffic management. Most previous studies used the household travel survey data, which includes a smaller sample of the urban population (usually one to two thousand individuals) over one or two days. The spatial interaction patterns extracted from such data may not be representative of the whole urban population, neither can it reveal the changes of such patterns over time. Recent studies modeled spatial interaction patterns using GPS data, mobile phone records, or social media data (Becker et al. 2011; Calabrese et al. 2013; Alexander et al. 2015; Zhou et al. 2016). A large number of people (tens of thousands) are usually included in the sample, and their data collection period spans from weeks to months. Home, work, and other activity locations can be inferred from such data. Individual and

collective T-A pattern analysis proposed in this research helps to determine an appropriate temporal cycle for the spatial interaction patterns.

Location-based services is one of the strategies of stimulating urban economic growth. Collective T-A pattern modeling helps urban planners answer questions like how people use urban space spatially and temporally, such as which group of people are likely to return to downtown for activities during the weekends; which group of people are likely to stay downtown for dinner or entertainment after work during weekdays (Becker et al. 2011). Individual T-A pattern modeling helps to establish a personalized profile for each person, thus urban service providers can identify individuals that are likely to present at certain places during certain hours of a day and a week. Individual and collective T-A pattern modeling both ensures the accurate delivery of location-based services (Noulas et al. 2012; Hasan and Ukkusuri 2015).

Urban sustainability is also one of the objectives of "smart cities". Individual and collective T-A pattern modeling can help compute individually and collectively carbon dioxide emissions and energy consumptions on a daily level, a weekly level, a monthly level, and a yearly level (Becker et al. 2013). This types of modeling can also help compute individual exposure to air pollutions during different periods of time. This helps cities to create a healthy living environment and a sustainable management of natural resources.

Future Studies

Collective and individual T-A pattern modeling techniques provide transportation geographers tools to study travel activity behaviors of urban people. Recall the

fundamental assumption in transportation geography that each individual repeats the same T-A every day and the T-A behaviors of people from a sample on one randomly chosen day is representative of the long-term T-A behaviors of the population. This assumption can be tested with the proposed modeling techniques in this research, as soon as long-term T-A data from a sufficient and unbiased sample of urban population became available. The case study results in this research were somewhat interesting. It showed that collective activity patterns differed significantly from a weekday to a weekend day, and the prediction of upcoming collective activity patterns based on the assumption of a weekly cycle was pretty accurate. This clearly contradicted the assumption that one-day data was enough for modeling collective T-A patterns. Furthermore, the individual level analysis found one of the two participants had two equally important daily T-A patterns and the other participant had no obvious pattern. This also contradicted the assumption that each individual had one daily routine of T-A. Future studies may focus on testing the assumption, as well as exploring the relationship between individual patterns and collective patterns. Results from such studies may change the design of transportation surveys, transportation demand analysis, and urban planning.

This research used GPS trajectory data to study individual and collective T-A behaviors, as it is easy to collect for a long period of time and contains precise location and time information compared to traditional travel surveys. However, the case study GPS data were not perfect. The San Francisco dataset may only represent urban commercial, tourism, and entertainment activities as it contained the trajectories of taxi cabs instead of regular people. The Beijing dataset contained a very limited number of people who might come from the same workplace and lacked any other information

about them. Thus, the trajectory data may be limited or biased for generating any prototypes of individual daily T-A patterns for the general urban population, neither was it able to build any link between the prototypes and the individual-level characteristics. Moreover, some of the T-A behaviors were not recorded by the participants due to privacy concerns or technical issues. Data incompleteness and inconsistency were challenges. Bigger and better datasets are needed. As GPS and location-based services grow faster, future studies will be able to use trajectory data sets from a variety of sources. Automobiles are equipped with a GPS device and smart phones are inserted with a GPS chip. Data collection does not have to be limited to certain types of service vehicles (e.g., taxi cabs, buses, etc.) or certain groups of people (e.g., from a research institution) anymore. Data sources other than GPS trajectory data may also be used to study urban population T-A behaviors. Mobile phone records may contain more accurate location information as the accuracy of cell tower triangulation improves. Social media data may be used to study people's T-A behaviors as they constantly post their locations and activities online.

There were a few concerns with the proposed modeling procedures. For the individual level modeling, the process of identifying T-A events and lining them up chronologically to create sequences might have introduced some errors. These may be related to particular algorithms used or the spatial and temporal scales adopted. For example, how trip end points were identified, clustered, and assigned to an anchor location might impact the constructing of sequence elements. This study used ClustalTXY for analyzing individual’s daily T-A sequences. Its sequence alignment and clustering algorithms may have room to improve. For example, it seemed to be pre-

determined that a set of sequences were always clustered into three groups, even with two of the three groups sharing the same consensus sequences. Moreover, the scale of similarity between corresponding elements were fixed to be between 0 and 10. However, to the authors' knowledge, ClustalTXY was the only available software package to handle multi-dimensional sequence alignment analysis. Future work could seek to develop assessment methods on the representativeness of consensus or median sequences.

Another concern was the potential impact of MAUP (MTUP) in the detection of collective activity hot spots. The spatiotemporal unit used in the analysis was the combination of census tracts and one-hour intervals. This was appropriate for the analysis in this research considering the size of the study area and the expected number of activities in each census tract per hour. Future studies may incorporate census data with the dynamics of the activity hot spots, in order to link the collective activity patterns with neighborhood characteristics. Other sets of grids combining with other time intervals may also be investigated in the future to build a full spectrum of the patterns under investigation. Approaches of this type would help test for the impact of other space-time scales on the collective activity patterns found.

Another concern was related to the spatial accuracy of the activity locations that were inferred from the taxi trajectory dataset. Both the accuracy of the census tract boundary and the accuracy of the derived activity locations might create chances for an activity not being placed into the correct census tract. Methods to improve the spatial accuracy of mapping the activity locations from this kind of dataset should be investigated in the future.

The last concern was related to a lack of effective visualization techniques for displaying the life cycle and development stages of activity hot spots and the space-time daily T-A patterns of individuals. Snapshots in two-dimensional space were used for displaying collective activity patterns and a map of anchor locations and T-A sequences were used for displaying individual daily T-A patterns. Future works may explore a 3D approach that can effectively incorporate the time dimension into the T-A patterns.

REFERENCES

Abbott, A. 1995. Sequence analysis: new methods for old ideas. *Annual Review of Sociology* 21: 93–113.

Abraham, S., and P. S. Lal. 2010. Trajectory similarity of network constrained moving objects and applications to traffic security. In *Intelligence and security informatics - Pacific Asia workshop, PAISI 2010, proceedings*, ed. H. Chen, M. Chau, S. Li, S. Urs, S. Srinivasa, and G. A. Wang, 31-43. Germany: Springer.

Alexander, L., S. Jiang, M. Murga, and M. C. Gonzalez. 2015. Origin-destination trips by purpose and time of day inferred from mobile phone data. *Transportation Research Part C: Emerging Technologies* 58: 240-50.

Alvares, L. O., V. Bogorny, J. A. Fernandes de Macedo, B. Moelans, and S. Spaccapietra. 2007a. Dynamic modeling of trajectory patterns using data mining and reverse engineering. In *ER '07 Tutorials, posters, panels and industrial contributions at the 26th international conference on conceptual modeling*, ed. J. Grundy, S. Hartmann, A. H. F. Laender, L. Maciaszek, and J. F. Roddick, 149-54. Darlinghurst, Australia: Australian Computer Society.

Alvares, L.O., V. Bogorny, B. Kuijpers, J. A. Fernandes de Macedo, B. Moelans, and A. Vaisman. 2007b. A model for enriching trajectories with semantic geographical information. In *Proceedings of the 15th ACM international symposium on advances in geographic information systems, GIS 2007*, 162-9. New York, NY: ACM.

Alvarez-Garcia, J. A., J. A. Ortega, L. Gonzalez-Abril, and F. Velasco. 2010. Trip destination prediction based on past GPS log using a Hidden Markov Model. *Expert Systems with Applications* 37 (12): 8166-71.

Anderson, R. P., D. Lew, and A. T. Peterson. 2003. Evaluating predictive models of species' distributions: criteria for selecting optimal models. *Ecological Modelling* 162 (3): 211-32.

Andrienko, G., and N. Andrienko. 2008. Spatio-temporal aggregation for visual analysis of movements. In *2008 IEEE symposium on visual analytics science and technology*, ed. D. Ebert and T. Ertl, 51-8. Los Alamitos, CA: IEEE.

Anselin, L. 1995. Local indicators of spatial association-LISA. *Geographical Analysis* 27 (2): 93-115.

———. 2005. *Exploring spatial data with GeoDa^{TM}: a workbook*. The center for spatial data science - The university of Chicago. https://s3.amazonaws.com/geoda/software/docs/geodaworkbook.pdf (last accessed 15 January 2018).

Ashbrook, D., and T. Starner. 2003. Using GPS to learn significant locations and predict movement across multiple users. *Personal and Ubiquitous Computing* 7 (5): 275-86.

Becker, R. A., R. Caceres, K. Hanson, J. M. Loh, S. Urbanek, A. Varshavsky, and C. Volinsky. 2011. A tale of one city: using cellular network data for urban planning. *IEEE Pervasive Computing* 10 (4): 18-26.

Becker, R., R. Caceres, K. Hanson, S. Isaacman, J. M. Loh, M. Martonosi, J. Rowland, S. Urbanek, A. Varshavsky, and C. Volinsky. 2013. Human mobility characterization from cellular network data. *Communications of the ACM* 56 (1): 74-82.

Bianchini, F. 1995. Night cultures, night economies. *Planning Practice and Research* 10 (2): 121–26.

Bianco, A. 2004. The vanishing mass market. *Business Week* 12 July: 61-72.

Biljecki, F., H. Ledoux, and P. Van Oosterom. 2013. Transportation mode based segmentation and classification of movement trajectories. *International Journal of Geographical Information Sciences* 27 (2): 385-407.

Black, W. R. 2004. Recent developments in US transport geography. In *Handbook of transport geography and spatial systems*, ed. D. A. Button, K. J. Haynes, and K. E. Stopher, 13-26. Oxford, UK: Elsevier Ltd.

Bogorny, V., B. Kuijpers, and L. O. Alvares. 2009. ST-DMQL: a semantic trajectory data mining query language. *International Journal of Geographical Information Science* 23 (10): 1245-76.

Bogorny, V., C. A. Heuser, and L. O. Alvares. 2010. A conceptual data model for trajectory data mining. In *Geographic information science - 6th international conference, GIScience 2010, proceedings*, ed. S. I. Fabrikant, T. Reichenbacher, M. Van Kreveld, and C. Schlieder, 1-15. Germany: Springer.

Bohte, W., and K. Maat. 2009. Deriving and validating trip purposes and travel modes for multi-day GPS-based travel surveys: a large-scale application in the Netherlands. *Transportation Research Part C* 17 (3): 285-97.

Bolbol, A., T. Cheng, I. Tsapakis, and J. Haworth. 2012. Infering hybrid transportation modes from sparse GPS data using a moving window SVM classification. *Computers, Environment, and Urban Systems* 36 (6): 526-37.

Brakatsoulas, S., D. Pfoser, and N. Tryfona. 2004. Modeling, storing, and mining moving object databases. In *International database engineering and applications symposium, 2004, IDEAS '04, proceedings*, 68-77. Los Alamitos, CA: IEEE.

Buchin, K., M. Buchin, J. Gudmundsson, M. Loffler, and J. Luo. 2011. Detecting commuting patterns by clustering subtrajectories. *International Journal of Computational Geometry & Applications* 21 (3): 253-82.

Buliung, R. N., M. J. Roorda, and T. K. Remmel. 2008. Exploring spatial variety in patterns of activity-travel behavior: initial results from the Toronto Travel-Activity Panel Survey (TTAPS). *Transportation* 35 (6): 697-722.

Calabrese, F., M. Diao, G. D. Lorenzo, J. Ferreira Jr., and C. Ratti. 2013. Understanding individual mobility patterns from urban sensing data: a mobile phone trace example. *Transportation Research Part C* 26: 301-13.

Chang, J., R. Bista, Y. Kim, and Y. Kim. 2007. Spatio-temporal similarity measure algorithm for moving objects on spatial networks. In *Computational science and its applications - ICCSA 2007 - international conference, proceedings*, ed. O. Gervasi, 1165-78. Germany: Springer.

Chen, C., H. Gong, C. Lawson, and E. Bialostozky. 2010. Evaluating the feasibility of a passive travel survey collection in a complex urban environment: lessons learned from the New York City case study. *Transportation Research Part A* 44: 830-40.

Chen, J., S.-L. Shaw, H. Yu, F. Lu, Y. Chai, and Q. Jia. 2011. Exploratory data analysis of activity diary data: a space-time GIS approach. *Journal of Transport Geography* 19 (3): 394-404.

Chen, Y., X. Liu, X. Li, X. Liu, Y. Yao, G. Hu, X. Xu, and F. Pei. 2017. Delineating urban functional areas with building-level social media data: a dynamic time warping (DTW) distance based *k*-medoids method. *Landscape and Urban Planning* 160: 48-60.

Cheng, H. K., and K. Dogan. 2008. Customer-centric marketing with internet coupons. *Decision Support Systems* 44: 606-20.

Chung, E., and A. Shalabay. 2005. A trip reconstruction tool for GPS-based personal travel surveys. *Transportation Planning and Technology* 28 (5): 381-401.

ClustalTXY. Version: 2.0. C. Wilson. Ottawa, ON, Canada.

Collia, D. V., J. Sharp, and L. Giesbrecht. 2003. The 2001 national household travel survey: A look into the travel patterns of older Americans. *Journal of Safety Research* 34 (4): 461-70.

Crooks, A., D. Pfoser, A. Jenkins, A. Croitoru, A. Stefanidis, D. Smith, S. Karagiorgou, A. Efentakis, and G. Lamprianidis. 2015. Crowdsourcing urban form and function. *International Journal of Geographical Information Science* 29 (5): 720-41.

De Almeida, V. T., and R. H. Guting. 2005. Indexing the trajectories of moving objects in networks. *GeoInformatica* 9 (1): 33–60.

Diggle, P. J., A. G. Chetwynd, R. Haqqkvist, and S. E. Morris. 1995. Second-order analysis of space-time clustering. *Statistical Methods in Medical Research* 4 (2): 124-36.

Ding, Y., H. Lu, and L. Zhang. 2016. An analysis of activity time use on vehicle usage rationed days. *Transportation* 43 (1): 145-58.

Dodge, S., R. Weibel, and E. Forootan. 2009. Revealing the physics of movement: Comparing the similarity of movement characteristics of different types of moving objects. *Computers, Environment, and Urban Systems* 33 (6): 419-34.

Dodge, S., P. Laube, and R. Weibel. 2012. Movement similarity assessment using symbolic representation of trajectories. *International Journal of Geographical Information Science* 26 (9): 1563-88.

Dunkel, A. 2015. Visualizing the perceived environment using crowdsourced photo geodata. *Landscape and Urban Planning* 142: 173-86.

Elnekave, S., M. Last, and O. Maimon. 2007. Incremental clustering of mobile objects. In *2007 IEEE 23rd international conference on data engineering workshop*, 585-92. Los Alamitos, CA: IEEE.

Ester, M., H.-P. Kriegel, J. Sander, and X. Xu. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the second international conference on knowledge discovery and data mining*, ed. E. Simoudis, J. Han, and U. Fayyad, 226-31. Palo Alto, CA: AAAI Press.

Fang, T. B., and Y. Lu. 2012. Personal real-time air pollution exposure assessment methods promoted by information technological advances. *Annals of GIS* 18 (4): 279-88.

Filion, P. 2000. Balancing concentration and dispersion? Public policy and urban structure in Toronto. *Environment and Planning C: Government and Policy* 18 (2): 163–89.

Fotheringham, A. S., and D. W. S. Wong. 1991. The modifiable areal unit problem in multivariate statistical analysis. *Environment and Planning A* 23 (7): 1025-44.

Gambs, S., M.-O. Killijian, and M. N. Del Prado Cortez. 2010. GEPETO: a GEoPrivacy-Enhancing TOolkit. In *2010 IEEE 24th international conference on advanced information networking and applications workshops (WAINA)*, 1071-6. Los Alamitos, CA: IEEE.

Gao, S., K. Janowicz, and H. Couclelis. 2017. Extracting urban functional regions from points of interest and human activities on location-based social networks. *Transactions in GIS* 21 (3): 446-67.

Gao, Y., B. Zheng, G. Chen, and Q. Li. 2010. Algorithms for constrained k-nearest neighbor queries over moving object trajectories. *Geoinformatica* 14 (2): 241-76.

Giannotti, F., M. Nanni, D. Pedreschi, F. Pinelli, C. Renso, S. Rinzivillo, and R. Trasarti. 2011. Unveiling the complexity of human mobility by querying and mining massive trajectory data. *International Journal on Very Large Data Bases* 20 (5): 695-719.

Goetz, A. R., B. A. Ralston, F. P. Stutz, and T. R. Leinbach. 2003. Transportation geography. In *Geography in America at the dawn of the 21st century*, ed. G. L. Gaile and C. J. Willmott, 221-36. Oxford: Oxford University Press.

Gong, H., C. Chen, E. Bialostozky, and C. T. Lawson. 2012. A GPS/GIS method for travel mode detection in New York City. *Computers, Environment and Urban Systems* 36 (2): 131-39.

Gonzales, P. A., J. S. Weinstein, S. J. Barbeau, M. A. Labrador, P. L. Winters, N. L. Georggi, and R. Perez. 2010. Automating mode detection for travel behaviour analysis by using global positioning systems-enabled mobile phones and neural networks. *IET Intelligent Transport Systems* 4 (1): 37-49.

Graham, B. 1999. TGRG page—a proposed research agenda for transport geography: mobility and social change. *Journal of Transport Geography* 7 (3): 235–36.

Greaves, S., S. Fifer, R. Ellison, and G. Germanos. 2010. Development of a Global Positioning System web-based prompted recall solution for longitudinal travel surveys. *Transportation Research Record* 2183: 69-77.

Grengs, J., X. Wang, and L. Kostyniuk. 2008. Using GPS data to understand driving behavior. *Journal of Urban Technology* 15 (2): 33-53.

Gutiérrez, J., and J. C. García-Palomares. 2007. New spatial patterns of mobility within the metropolitan area of Madrid: Towards more complex and dispersed flow networks. *Journal of Transport Geography* 15 (1): 18–30.

Haight, F. A. 1967. *Handbook of the Poisson distribution*. New York: John Wiley & Sons.

Hanson, S., and P. Hanson. 1981. The travel-activity patterns of urban residents: dimensions and relationships to sociodemographic characteristics. *Economic Geography* 57 (4): 332-47.

Hanson, S., and J. O. Huff. 1988. Systematic variability in repetitious travel. *Transportation* 15 (1-2): 111-35.

Hariharan, R., and K. Toyama. 2004. Project lachesis: parsing and modeling location histories. *Lecture Notes in Computer Science* 3234: 106-24.

Hasan, S., and S. V. Ukkusuri. 2015. Location contexts of user check-ins to model urban geo life-style patterns. *PLoS ONE* 10 (5): 1-19.

Herring, R. J. 2010. Real-time traffic modeling and estimation with streaming probe data using machine learning: Dissertation.

Hu, H., Z. Wu, B. Mao, Y. Zhuang, J. Cao, and J. Pan. 2012. Pick-up tree based route recommendation from taxi trajectories. In *Web-age information management - 13th international conference, WAIM 2012, proceedings*, ed. H. Gao, L. Lim, W. Wang, C. Li, and L. Chen, 471-83. Germany: Springer.

Huff, J. O., and S. Hanson. 1986. Repetition and variability in urban travel. *Geographical Analysis* 18 (2): 97-114.

Hwang, J.-R., H.-Y. Kang, and K.-J. Li. 2005. Spatio-temporal similarity analysis between trajectories on road networks. In *Perspectives in conceptual modeling - ER 2005 Workshops CAOIS, BP-UML, CoMoGIS, eCOMO, and QoIS, Proceedings*, ed. J. Akoka, S. W. Liddle, I.-Y. Song, M. Bertolotto, I. Comyn-Wattiau, S. S. Cherfi, W.-J. van den Heuvel, B. Thalheim, M. Kolp, P. Bresciani, J. Trujillo, C. Kop, and H. C. Mayr, 280-89. Germany: Springer.

Hwang, R. H., Y. L. Hsueh, and H. W. Chung. 2012. A novel time-obfuscated algorithm for trajectory privacy. In *2012 12th international symposium on pervasive systems, algorithms and networks, ISPAN*, 208-15. Los Alamitos, CA: IEEE.

Jarv, O., K. Muurisepp, R. Ahas, B. Derudder, and F. Witlox. 2015. Ethnic differences in activity spaces as a characteristic of segregation: a study based on mobile phone usage in Tallinn, Estonia. *Urban Studies* 52 (14): 2680-98.

Jeung, H., M. L. Yiu, X. Zhou, and C. S. Jensen. 2010. Path prediction and predictive range querying in road network databases. *International Journal on Very Large Data Bases* 19 (4): 585-602.

Jiang, S., J. Ferreira, and M. C. Gonzalez. 2012. Clustering daily patterns of human activities in the city. *Data Mining and Knowledge Discovery* 25 (3): 478-510.

Jiang, S., X. Qian, T. Mei, and Y. Fu. 2016. Personalized travel sequence recommendation on multi-source big social media. *IEEE Transactions on Big Data* 2 (1): 43-56.

Joh, C.-H., T. Arentze, and H. Timmermans. 2001. Pattern recognition in complex activity travel patterns: comparison of Euclidean distance, signal-processing theoretical, and multidimensional sequence alignment methods. *Transportation Research Record* 1752: 16-22.

———. 2005. A utility-based analysis of activity time allocation decisions underlying segmented daily activity-travel patterns. *Environment and Planning A* 37 (1): 105-25.

———. 2007. Identifying skeletal information of activity patterns by multidimensional sequence alignment. *Transportation Research Record* 2021: 81-8.

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

Kim, S., J. Won, J. Kim, M. Shin, J. Lee, and H. Kim. 2007. Path prediction of moving objects on road networks through analyzing past trajectories. In *Knowledge-based intelligent information and engineering systems: KES 2007 - WIRN 2007*, ed. B. Apolloni, R. J. Howlett, and L. Jain, 379-89. Germany: Springer.

Kitamura, R., and T. Van Der Hoorn. 1987. Regularity and irreversibility of weekly travel behavior. *Transportation* 14 (3): 227-51.

Knowles, R. D. 1993. Research agendas in transport geography for the 1990s. *Journal of Transport Geography* 1 (1): 3-11.

Knox, E. G., and M. S. Bartlett. 1964. The detection of space-time interactions. *Journal of the Royal Statistical Society Series C (Applied Statistics)* 13 (1): 25-30.

Krenn, P. J., S. Titze, P. Oja, A. Jones, and D. Ogilvie. 2011. Use of Global Positioning Systems to study physical activity and the environment: a systematic review. *American Journal of Preventive Medicine* 41 (5): 508-15.

Kulldorff, M. 2010. *SaTScan™ user guide for version 9.4*. SaTScan. https://www.satscan.org/cgi-bin/satscan/register.pl/Current%20Version:%20SaTScan%20v9.1.1%20released%20March%209%202011.?todo=process_userguide_download (last accessed 15 January 2018).

Kulldorff, M., R. Heffernan, J. Hartman, R. Assuncao, and F. Mostashari. 2005. A space-time permutation scan statistic for disease outbreak detection. *PLoS Medicine* 2 (3): 216-24.

Kumar, V., and J. A. Petersen. 2005. Using a customer-level marketing strategy to enhance firm performance: a review of theoretical and empirical evidence. *Journal of the Academy of Marketing Science* 33 (4): 504-19.

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

———. 2009. From place-based to people-based exposure measures. *Social Science and Medicine* 69 (9): 1311-3.

Kwan, M.-P., N. Xiao, and G. Ding. 2014. Assessing activity pattern similarity with multidimensional sequence alignment based on a multiobjective optimization evolutionary algorithm. *Geographical Analysis* 46 (3): 297-320.

Lawson, A. B. 2010. Hotspot detection and clustering: ways and means. *Environmental and Ecological Statistics* 17 (2): 231-45.

Lee, J., J. Han, and K. Whang. 2007. Trajectory clustering: a partition-and-group framework. In *Proceedings of the ACM SIGMOD international conference on management of data, SIGMOD 2007*, 593-604. New York, NY: ACM.

Li, X., and H. Lin. 2006. Indexing network-constrained trajectories for connectivity-based queries. *International Journal of Geographical Information Science* 20 (3): 303-28.

Li, Z., J.-G. Lee, X. Li, and J. Han. 2010. Incremental clustering for trajectories. In *15th international conference on database systems for advanced applications, DASFAA 2010, proceedings*, ed. H. Kitagawa, Y. Ishikawa, W. Li, and C. Watanabe, 32-46. Germany: Springer.

Lin, B., and J. Su. 2008. One way distance: for shape based similarity search of moving object trajectories. *Geoinformatica* 12 (2): 117-42.

Liu, X., and H. A. Karimi. 2006. Location awareness through trajectory prediction. *Computers, Environment and Urban Systems* 30 (6): 741-56.

Lovatt, A., and J. O'Connor. 1995. Cities and the night-time economy. *Planning Practice and Research* 10 (2): 127–34.

Lu, Y., and Y. Lui. 2012. Pervasive location acquisition technologies: opportunities and challenges for geospatial studies. *Computers, Environment and Urban Systems* 36 (2): 105-8.

Lv, M., L. Chen, and G. Chen. 2013. Mining user similarity based on routine activities. *Information Sciences* 236: 17-32.

Manso, J. A. R. C., V. C. Times, G. Oliveira, L. O. Alvares, and V. Bogorny. 2010. DB-SMoT: a direction-based spatio-temporal clustering method. In *2010 5th IEEE international conference intelligent systems (IS)*, 114-9. Los Alamitos, CA: IEEE.

Montoliu, R., J. Blom, and D. Gatica-Perez. 2013. Discovering places of interest in everyday life from smartphone data. *Multimedia Tools & Applications* 62 (1): 179-207.

Mouza, C., and P. Rigaux. 2005. Mobility Patterns. *GeoInformatica* 9 (4): 297–319.

Murakami, E., and D. P. Wagner. 1999. Can using global positioning system (GPS) improve trip reporting? *Transportation Research Part C* 7 (2): 149-65.

Nanni, M., and D. Pedreschi. 2006. Time-focused clustering of trajectories of moving objects. *Journal of Intelligent Information Systems* 27 (3): 267-89.

Navarro, G. 2001. A guided tour to approximate string matching. *ACM Computing Surveys* 33 (1): 31-88.

Noulas, A., S. Scellato, N. Lathia, and C. Mascolo. 2012. Mining user mobility features for next place prediction in location-based services. In *2012 IEEE 12th international conference on data mining*, 1038-43. Los Alamitos, CA: IEEE.

Oliveira, M., P. Vovsha, J. Wolf, Y. Birotker, D. Givon, and J. Paasche. 2011. Global positioning system-assisted prompted recall household travel survey to support development of advanced travel model in Jerusalem, Israel. *Transportation Research Record* 2246: 16-23.

Palma, A. T., V. Bogorny, B. Kuijper, and L. O. Alvares. 2008. A clustering-based approach for discovering interesting places in trajectories. In *Proceedings of the 23rd annual ACM symposium on applied computing, SAC'08*, 863-8. New York, NY: ACM.

Pan, G., G. Qi, Z. Wu, D. Zhang, and S. Li. 2013. Land-use classification using taxi GPS traces. *IEEE Transactions on Intelligent Transportation Systems* 14 (1): 113-23.

Pas, E. I. 1983. A flexible and integrated methodology for analytical classification of daily travel-activity behavior. *Transportation Science* 17 (4): 405-29.

———. 1988. Weekly travel-activity behavior. *Transportation* 15 (1-2): 89-109.

Pas, E. I., and F. S. Koppelman. 1987. An examination of the determinants of day-to-day variability in individuals' urban travel behavior. *Transportation* 14 (1): 3-20.

Pas, E. I., and S. Sundar. 1995. Intrapersonal variability in daily urban travel behavior: some additional evidence. *Transportation* 22 (2): 135-50.

Pattison, W. D. 1964. The four traditions of geography. *Journal of Geography* 63 (5): 211-6.

Pfoser, D., and C. S. Jensen. 2005. Trajectory indexing using movement constraints. *GeoInformatica* 9 (2): 93-115.

Pelekis, N., G. Andrienko, N. Andrienko, I. Kopanakis, G. Marketos, and Y. Theodoridis. 2012. Visually exploring movement data via similarity-based analysis. *Journal of Intelligent Information Systems* 38 (2): 343-91.

Preston, J. 2001. Integrating transport with socio-economic activity—a research agenda for the new millennium. *Journal of Transport Geography* 9 (1): 13–24.

Qiao, S., C. Tang, H. Jin, T. Long, S. Dai, Y. Ku, and M. Chau. 2010. PutMode: prediction of uncertain trajectories in moving objects databases. *Applied Intelligence* 33 (3): 370-86.

Recker, W. W., M. G. McNally, and G. S. Root. 1985. Travel/activity analysis: pattern recognition, classification and interpretation. *Transportation Research Part A* 19 (4): 279-96.

Reddy, S., M. Mun, J. Burke, D. Estrin, M. Hansen, and M. Srivastava. 2010. Using mobile phones to determine transportation modes. *ACM Transactions on Sensor Networks* 6 (2): 13-27.

Ripley, B. D. 1976. The second-order analysis of stationary point processes. *Journal of Applied Probability* 13 (2): 255-66.

Roh, G., and S. Hwang. 2010. NNCluster: an efficient clustering algorithm for road network trajectories. In *Database systems for advanced applications - 15th international conference, DASFAA 2010, proceedings*, ed. H. Kitagawa, Y. Ishikawa, Q. Li, and C. Watanabe, 47-61. Germany: Springer.

Roh, G., J. Roh, S. Hwang, and B. Yi. 2011. Supporting pattern-matching queries over trajectories on road networks. *IEEE Transactions on Knowledge and Data Engineering* 23 (11): 1753-8.

Rossmo, D. K., Y. Lu, and T. B. Fang. 2012. Spatial-temporal crime paths. In *Patterns, Prevention, and Geometry of Crime*, ed. M. A. Andresen and J. B. Kinney, 16-42. London and New York: Routledge.

Sadahiro, Y., R. Lay, and T. Kobayashi. 2013. Trajectories of moving objects on a network: detection of similarities, visualization of relations, and classification of trajectories. *Transactions in GIS* 17 (1): 18-40.

Saitou, N., and M. Nei. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution* 4 (4); 406-25.

Scellato, S., M. Musolesi, C. Mascolo, V. Latora, and A. T. Campbell. 2011. NextPlace: a spatio-temporal prediction framework for pervasive systems. In *Pervasive computing - 9th international conference, Pervasive 2011, proceedings*, ed. K. Lyons, J. Hightower, and E. M. Huang, 152-69. Germany: Springer.

Seifert, J. 2004. Data mining and the search for security: challenges for connecting the dots and databases. *Government Information Quarterly* 21 (4): 461-80.

Shaw, S. L. 2006. What about "time" in transportation geography? *Journal of Transport Geography* 14 (3): 237-40.

Shelton, T., A. Poorthuis, and M. Zook. 2015. Social media and the city: rethinking urban socio-spatial inequality using user-generated geographic information. *Landscape and Urban Planning* 142: 198-211.

Sheth, J. N., R. S. Sisodia, and A. Sharma. 2000. The antecedents and consequences of customer-centric marketing. *Journal of the Academy of Marketing Science* 28 (1): 55-66.

Shoval, N. 2008. Tracking technologies and urban analysis. *Cities* 25 (1): 21-8.

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

Shoval, N., B. McKercher, A. Birenboim, and E. Ng. 2015. The application of a sequence alignment method to the creation of typologies of tourist activity in time and space. *Environment and Planning B - Planning & Design* 42 (1): 76-94.

Spaccapietra, S., C. Parent, M. L. Damiani, J. A. Macedo, F. Porto, and C. Vangenot. 2008. A conceptual view on trajectories. *Data & Knowledge Engineering* 65 (1): 126-46.

Steenbruggen, J., E. Tranos, and P. Nijkamp. 2015. Data from mobile phone operators: a tool for smarter cities? *Telecommunications Policy* 39 (3-4): 335-46.

Stopher, P. R., and S. P. Greaves. 2007. Household travel surveys: Where are we going? *Transportation Research Part A* 41 (5): 367-81.

Stopher, P. R., and Y. Zhang. 2011. Repetitiveness of daily travel. *Transportation Research Record* 2230: 75-84.

Stopher, P., C. FitzGerald, and J. Zhang. 2008. Search for a global positioning system device to measure person travel. *Transportation Research Part C* 16 (3): 350-69.

Taaffe, E. J., and H. L. Gauthier. 1994. Transportation geography and geographic thought in the United States: an overview. *Journal of Transport Geography* 2 (3): 155-68.

Tang, J., and L. Meng. 2006. Learning significant locations from GPS data with time window. In *Proceedings of SPIE - the international society for optical engineering, Geoinformatics 2006: GNSS and integrated geospatial applications*, ed. D. Li and L. Xia. SPIE.

Thierry, B., B. Chaix, and Y. Kestens. 2013. Detecting activity locations from raw GPS data: a novel kernel-based algorithm. *International Journal of Health Graphics* 12 (14): 1-10.

Tiakas, E., A. N. Papadopoulos, A. Nanopoulos, Y. Manolopoulos, D. Stojanovic, and S. Djordjevic-Kajan. 2009. Searching for similar trajectories in spatial networks. *The Journal of Systems and Software* 82 (5): 772–88.

Ullman, E. I. 1954. Transportation Geography. In *American Geography: Inventory and Prospect*, ed. E. J. Preston and C. F. Jones, 310-32. Syracuse, NY: Syracuse University Press.

Van Der Hoorn, T. 1979. Travel behavior and the total activity pattern *Transportation* 8 (4): 309-28.

Vanhulsel, M., C. Beckx, D. Janssens, K. Vanhoof, and G. Wets. 2011. Measuring dissimilarity of geographically dispersed space-time paths. *Transportation* 38 (1): 65-79.

Vu, T. H. N., K. H. Ryu, and N. Park. 2009. A method for predicting future location of mobile user for location-based services system. *Computers & Industrial Engineering* 57 (1): 91-105.

Wang, Y., T. Wang, M.-H. Tsou, H. Li, W. Jiang, and F. Guo. 2016. Mapping dynamic urban land use patterns with crowdsourced geo-tagged social media (sina-weibo) and commercial points of interest collections in Beijing, China. *Sustainability* 8 (11): 1-19.

Wartenberg, D. 1985. Multivariate spatial correlation: a method for exploratory geographical analysis. *Geographical Analysis* 17 (4): 263-83.

Wiehe, S., A. E. Carroll, G. C. Liu, K. L. Haberkorn, S. C. Hoch, J. S. Wilson, and J. D. Fortenberry. 2008. Using GPS-enabled cell phones to track the travel patterns of adolescents. *International Journal of Health Geographics* 7: 1-11.

Wilson, C. 1998. Analysis of travel behavior using sequence alignment methods. *Transportation Research Record* 1645: 52-9.

———. 2008. Activity patterns in space and time: calculating representative Hagerstrand trajectories. *Transportation* 35 (4): 485-99.

Wolf, J., S. Hallmark, M. Oliveira, R. Guensler, and W. Sarasua. 1999. Accuracy issues with route choice data collection by using global positioning system. *Transportation Research Record* 1660: 66-74.

Wolf, J., S. Schonfelder, U. Samaga, M. Oliveira, and K. W. Axhausen. 2004. Eighty weeks of Global Positioning System traces: approaches to enriching trip information. *Transportation Research Record* 1870: 46-54.

Wu, J., C. Jiang, D. Houston, D. Baker, and R. Delfino. 2011. Automated time activity classification based on global positioning system (GPS) tracking data. *Environmental Health* 10: 1-13.

Wu, J., C. Jiang, G. Jaimes, S. Bartell, A. Dang, D. Baker, and R. J. Delfino. 2013. Travel patterns during pregnancy: comparison between Global Positioning System (GPS) tracking and questionnaire data. *Environmental Health* 12: 1-12.

Xu, Y., S.-L. Shaw, Z. Zhao, L. Yin, Z. Fang, and Q. Li. 2015. Understanding aggregate human mobility patterns using passive mobile phone location data: a home-based approach. *Transportation* 42 (4): 625-46.

Yalamanchili, L., R. M. Pendyala, N. Prabaharan, and P. Chakravarthy. 1998. Analysis of global positioning system-based data collection methods for capturing multistop trip-chaining behavior. *Transportation Research Record* 1660: 58-65.

Ye, Y., Y. Zheng, Y. Chen, J. Feng, and X. Xie. 2009. Mining individual life pattern based on location history. In *2009 tenth international conference on mobile data management: systems, services, and middleware, MDM '09*, 1-10. Los Alamitos, CA: IEEE.

Yue, Y., Y. Zhuang, Q. Li, and Q. Mao. 2009. Mining time-dependent attractive areas and movement patterns from taxi trajectory data. In *2009 17th international conference on geoinformatics*, ed. L. Di and A. Chen, 1-6. Los Alamitos, CA: IEEE.

Zhan, X., S. V. Ukkusuri, and F. Zhu. 2014. Inferring urban land use using large-scale social media check-in data. *Networks and Spatial Economics* 14 (3-4): 647-67.

Zhao, X., and W. Xu. 2009. A clustering-based approach for discovering interesting places in a single trajectory. In *2009 second international conference on intelligent computation technology and automation, ICICTA '09*, ed. J. E. Guerrero, 429-32. Los Alamitos, CA: IEEE.

———. 2011. A new measurement method to calculate similarity of moving object spatio-temporal trajectories by compact representation. *International Journal of Computational Intelligence Systems* 4 (6): 1140-7.

Zheng, Y., Q. Li, Y. Chen, X. Xie, and W.-Y. Ma. 2008. Understanding mobility based on GPS data. In *UbiComp 2008 - proceedings of the 10th international conference on ubiquitous computing*, 312-21. New York, NY: ACM.

Zheng, Y., L. Zhang, X. Xie, and W.-Y. Ma. 2009. Mining interesting locations and travel sequences from GPS trajectories. In *WWW '09 - proceedings of the 18th international world wide web conference*, 791-800. New York, NY: ACM.

Zheng, Y., Y. Chen, Q. Li, X. Xie, and W. Ma. 2010. Understanding transportation modes based on GPS data for web applications. *ACM Transactions on the Web* 4 (1): 1-36.

Zhou, M., Y. Yue, Q. Li, and D. Wang. 2016. Portraying temporal dynamics of urban spatial divisions with mobile phone positioning data: a complex network approach. *International Journal of Geo-Information* 5 (12): 1-21.

Zhou, Y., Z. Fang, J.-C. Thill, Q. Li, and Y. Li. 2015. Functionally critical locations in an urban transportation network: identification and space-time analysis using taxi trajectories. *Computers, Environment and Urban Systems* 52: 34-47.