DYNAMIC ORIGIN DESTINATION ESTIMATION WITH LOCATION-BASED SOCIAL NETWORKING DATA: EXPLORING URBAN TRAVEL DEMAND SENSOR

By

WANGSU HU

A dissertation submitted to the

School of Graduate Studies

Rutgers, The State University of New Jersey

In partial fulfillment of the requirements

For the degree of

Doctor of Philosophy

Graduate Program in Civil and Environmental Engineering

Written under the direction of

Peter J. Jin

and approved by

New Brunswick, New Jersey

January, 2019

ABSTRACT OF THE DISSERTATION

Dynamic Origin-Destination Estimation with Location-based Social Networking Data: Exploring Urban Travel Demand Sensor

by

Wangsu Hu

Dissertation Director:

Peter J.Jin

The emergence of Transportation Big Data provides rich information for estimating and predicting urban travel demand patterns. The traditional travel demand sensors involve labor-intensive survey data, traffic detector data for assignment model calibration, vehicle re-identification data from scattered Bluetooth, Wifi, or License plate readers, or aggregated cellphone activity data used in existing dynamic Origin-Destination estimation models or applications. With the growing number of mobile devices with GPS units and improvement in WLT technologies, the Location-Based Social Network (LBSN) data is an emerging travel demand data source. LBSN data recorded check-in or tweeting activities of massive users at different points of interests (POIs). The wide-range of POIs ensure the dense coverage of the main urban areas and the user-confirmed POI information provides the much-needed trip purpose information not available in other data sources. Meanwhile, the LBSN data has the advantages of passive secondary data collection usually not for the purpose of travel surveys, and anonymization. Despite the above advantages, LBSN data is not without its limitations for estimating urban travel demand, as well as dynamic OD estimation for proactive urban congestion mitigation and operations. First, LBSN can have a systematic temporal error for estimating travel demand. The LBSN activities do not always mimic travel activities throughout the day. Second, LBSN data includes a sampling bias for different population groups and venue types. Third, the stochastic nature of human activities, especially the POI arriving patterns are critical for travel demand estimation. The existing approaches to the LBSN-based travel demand analysis have suffered from those limitations on deriving the dynamic travel demand patterns.

Recent development in spatial-temporal characteristics provides the opportunities to identify and quantify the correlation between LSBN-based travel activity and urban travel demand pattern. In this dissertation, a novel set of travel demand models based on the LBSN data is proposed and tested. The research starts with a comprehensive review of the existing travel demand data collection methods and the travel demand modeling. Then we introduce a profiling method to infer the functionality of city zones based on the POI categorical distribution and local mobility patterns. By classifying zones by these zone topics, we can now analyze interactions between zones of different functionality. Thirdly, by conducting zonal time-of-day variation modeling on the LBSN check-in arrivals, a new stochastic point process based trip arrivals estimation is developed. The output is applied to the input of a temporal delay based trip distribution model for deriving dynamic OD patterns. And the model calibration and applications are also provided and discussed. The evaluation results illustrate the promising benefits of applying LBSN Data in urban travel demand modeling.

ACKNOWLEDGEMENTS

I would like to start by thanking my advisor, Prof. Jing Jin, for his continuous support, encouragement, and guidance throughout my academic career from year 2014 to now. He has led me to the transportation field with his knowledge, passion, vision and experience. This work would not have been possible without his guidance and tireless commitment to his students, the university, and to the profession.

It has been an honor and a privilege to have Prof. Jie Gong, Prof. Xiang Liu, Prof. Xuan Di and Prof. Hui Xiong serve on my committee and provide me with suggestions on improving the dissertation.

Thanks also to my friend and colleague: Yizhou Wang for his assistance with graduate student life, research work, and friendship throughout the years.

Outside Rutgers University, I would like to extend my gratitude to my mentor Jin Cao and collaborators Sen Yang and Weiqing Ni at Nokia Bell Labs. I had one of my best research experiences there and gained ideas about what my future career will be.

Furthermore, I would like to thank Foursquare for providing the API (Application Programming Interface) and access to their data and also thank the Capital Area Metropolitan Planning Organization (CAMPO) and New York Metropolitan Transportation Council (NYMTC) for providing the 2010 & 2017 Origin-Destination data used in this study. This research was partially supported by National Science Foundation of China (NSFC) Grant No. 61620106002.

Finally, I would like to thank my wife Shanshan Niu for her love, support, and patience these years.

PREFACE

The work conducted in this dissertation has been presented and published in several conferences and journals. Below is the list of publications derived from this proposal:

- **Wangsu Hu**, Peter J. Jin, Dynamic trip attraction estimation with location-based social network data balancing between the time-of-day variations and zonal differences, ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, p193-198. 6p, 2015
- **Wangsu Hu**, Peter J. Jin, An adaptive hawkes process formulation for estimating time-of-day zonal trip arrivals with location-based social networking check-in data, Transportation Research Part C: Emerging Technologies, 79, 136-155, 2017.
- **Wangsu Hu**, and Peter J. Jin, Adaptive Hawkes Process Formulation for Estimating Urban Trip Attraction with Location-Based Social Networkinging Data, 16-4766, TRB 95th Annual Meeting, Washington D.C., January 2016.
- **Wangsu Hu**, Peter J. Jin, A Hawkes Process Based Dynamic Trip Attraction Model using Open Location Based Social Network Data, INFORMS Annual Meeting 2015, Philadelphia, PA, November 2015.
- **Wangsu Hu**, and Peter J. Jin, Dynamic Origin-Destination Estimation based on Time Delay Correlation Analysis on Location-based Social Network (LBSN) Data, 18-03032, TRB 97th Annual Meeting, Washington D.C., January 2018.
- **Wangsu Hu**, Zijun Yao, Sen Yang, Shuhong Chen, and Peter J. Jin. "Discovering Urban Travel Demands through Dynamic Zone Correlation in Location-Based Social Networks." In ECMLPKDD, Dublin, Ireland., 2018.
- Zijun Yao, Yanjie Fu, Bin Liu, **Wangsu Hu**, and Hui Xiong. "Representing Urban Functions through Zone Embedding with Human Mobility Patterns." In IJCAI, pp. 3919-3925. 2018.

TABLE OF CONTENTS

List of Tables

List of Figures

1 Introduction

Travel demand modeling is the attempt of estimating the number of vehicles or people that will use a specific transportation facility. For instance, a travel demand forecasting model may predict the number of vehicles passing through a planned tunnel during AM peak period of a weekday, the ridership on a transit route, or the number of ships departing from and arriving at a seaport. Such prediction starts with the collection of historical/current travel demand data collection, such as population, employment, trip rates, travel costs, etc., to develop a traffic demand model for the desire situation. The result of travel demand modeling can be used for several key purposes in transportation policy, planning, and engineering. For example, we need the predicted numbers/arrival time/waiting time of the travelers to design a public traffic infrastructure e.g., calculate the desire bus routes, the capacity of bus station, and bus schedule. Meanwhile, the current technologies facilitate the access to real-time data and big data to provide the opportunity to develop new algorithms to improve the predictability and accuracy of the travel demand modeling. The existing data sources for dynamic travel demand analysis can be classified in to three major categories, flow calibration data, fixed-location vehicle identification data, and WLT (Wireless Location Technologies) data. Flow data collected from traffic detectors are first used to generate traffic information in the literature. The data requires a baseline matrix, and the matrix is calibrated to reproduce the dynamic flow through traffic assignment. Fixed-location vehicle identification data collected through Bluetooth readers [1], Electronic Toll Tag Readers [2], and License Plate Readers [3] can identify vehicles at different locations. Through re-identification, the dynamic origin-destination (OD) information of vehicles traveling within the covered areas or highway system can then be detected [1]. The key issue with the fixed-location sensor data is its limited geographical coverage and sampling rate. The WLT data use passive or active wireless positioning signals to track user activities within the geographical coverage of cellular network. Analytic models are developed to directly convert cellphone trajectories into dynamic travel time patterns and OD path flow patterns of mobile phone users. The framework has achieved some success but with key limitations related to sampling bias and positioning errors [4]. These errors can lead to difficulty in identifying location types therefore trip purposes. The trip purposes need to be derived from analyzing recurrent trip patterns e.g. commuting patterns.

Origin-destination (OD) estimation is a key step in urban travel demand analysis. It predicts the destination choices of travelers based on the zonal production and attraction abilities and the level of travel impedance between each OD pair [5]. The key to accurate OD estimation is the accuracy of sensing the static or dynamic correlation between origins and destinations [6]. Static OD estimation models estimate trip distribution within a targeted period (e.g. a normal workday) to capture the long-term average trend of travel demand patterns; while dynamic OD estimation reveals not only spatial but also temporal mobility patterns of travel demand. The increasing popularity of social networking services (SNS) and location-based services (LBS) has offered new opportunities for urban mobility patterns analysis. The combination of SNS and LBS leads to a new type of social networking service, Location-based Social Networking (LBSN) service. In LBSN, users can “check-in” with their LBS-enabled mobile devices to a nearby “venue,” or point of interest (POI) to declare their arrivals. Such information can be shared with friends and family, as well as with business owners for potential discounts and promotions. Given the pre-registered location and POI type information of venues, travelers’ trip arrivals are recorded with accurate location and trip purpose information. When aggregated, such data can provide a new secondary data source for the estimation of urban travel demand.

1.1 Problem Statement

LBSN data are not generated for the purpose of travel survey, so the dataset does have its limitations in urban mobility and travel demand estimation, as well as dynamic OD estimation for proactive urban congestion mitigation and operations [7, 8]. First, LBSN can have a systematic temporal error for estimating travel demand. The LBSN activity does not always mimic travel activities throughout the day. LBSN check-in activities tend to be more intensive during afternoons and evenings at social recreational places, as opposed to during morning peak hours when commuters are rushing against time to get to workplaces. The previous studies show that the check-in arrival patterns need appropriate dynamic stochastic processing to infer the underlying trip arrival patterns [9, 10]. Second, LBSN data includes a sampling bias for different population groups and venue types. Third, the stochastic nature of human activities, especially the POI arriving patterns are critical for travel demand estimation. In previous studies [11, 12], it was observed that the accurate estimation of zonal departures and arrivals are critical in reducing the errors in the subsequent OD estimation. Finally, the existing LBSN-based travel demand modeling only focus on static or time-of-day OD estimation for planning applications. Such application does not take full advantage of the high spatial-temporal resolution and large-scale coverage of the LBSN data. Dynamic OD estimation, however, can potentially provide valuable inputs to traffic operational applications such as the Active Transportation Demand Management (ATDM) applications. Existing studies on the time-of-day OD matrices were generated by aggregating the predicted trips over a targeted period of time. The aggregation was to reveal a time-of-day urban travel demand pattern rather than a time-dependent OD matrix.

There are several problems to be addressed in this proposal.

1. How can the real-time LBSN data benefit the dynamic OD estimation, especially

compared to the existing travel demand data collection methods and models?

2. How can we handle the systematic error between LSBN-based travel activity and urban travel demand pattern across different times of day and different locations?

3. How can we describe the stochastic nature of human activity over certain activity types (e.g., shopping and dining)?

4. How can urban dynamic travel demand patterns be identified and quantified to derive dynamic OD estimation using LBSN data?

1.2 Research Objectives and Scope of Work

To estimate the population activity patterns in urban areas from LBSN data, several critical issues need to be addressed. First, the LBSN activity data needs to be interpolated and expanded into the population activity patterns in urban areas. For dynamic activity pattern estimation, such interpolation or expansion needs to be conducted both spatially and temporally. Second, LBSN data includes sampling bias over some population groups (e.g., income and education levels) and certain activity types (e.g., shopping and dining). Third, the stochastic nature of human activities needs to be addressed in the estimation model. To address those issues, stochastic dynamic estimation models need to be considered and the calibration methods should be carefully designed to reduce the impact of sampling bias.

The dynamic travel demand patterns can be characterized by zonal trip intensity patterns and the dynamic OD flow patterns. The zonal trip intensity patterns can be directly generated by aggregating the human activity pattern based on the required spatial resolution (e.g., POI, land use parcel, or Traffic Analysis Zone (TAZ)). However, the estimation of dynamic OD flow patterns requires advanced modeling to identify the spatial-temporal correlation among different zonal trip intensity changes. The existing models, such as the gravity models and more recently the radiation models, have difficulties in dynamic OD estimation. Both models are developed for offline planning purposes and do not account for the variations and interactions between OD flow patterns in different time intervals unless the model parameters are calibrated at each time interval, which may not converge in time for dynamic demand pattern estimation. In this study, to address those limitations, we propose a new approach that estimates OD flows based on the spatial-temporal correlations of the activity intensity changes at different locations (zones) at different time intervals. For example, the drop of Twitter activities at work places and the increase of Foursquare check-in activities at restaurants several time intervals later during lunch hours can be used to estimate the number of trips made from work places to restaurants.

Based on the above status of current LBSN-based travel demand modeling research, the objective of my research is as follows:

1. Bias reduction of social media data based travel demand modeling:

To apply the social media data for travel demand modeling, we first address the issue of sampling bias of the LBSN user. It shows that not every traveler use LBSN service and not every place report LBSN check-in activity at every time period. We first explore the zonal urban functionality based on the local POI distribution and taxi pick-up/drop-off pattern to quantify the relationship between zonal check-in arrival and zonal trip arrival. Secondly, we applied a point process method to model the time-of-day variation of trip arrival using check-in arrival. By doing so, we can address the spatial and temporal bias for the social media data.

2. The zonal functionality profiling using POI data:

For zonal functionality profiling, we treat zonal functionality as a latent topic" variable to discover from POI categorical density. By classifying zones by these zone topics, we can now analyze interactions between zones of different functionality.

3. The zonal time-of-day (TOD) variations modeling of travel activity using LBSN data:

To estimate the travel demand patterns in urban areas from LBSN data, the LBSN activity data needs to be interpolated and expanded into the population activity patterns in urban areas. For dynamic travel demand TOD pattern, such interpolation or expansion needs to be conducted both spatially and temporally. Meanwhile, LBSN data includes sampling bias over some population groups (e.g., income and education levels) and certain activity types (e.g., shopping and dining). A clustering model need to be considered and the calibration methods should be carefully designed to reduce the impact of sampling bias.

4. The dynamic trip arrival modeling using LBSN data:

The stochastic nature of human activities needs to be considered in urban travel demand modeling, especially on trip arrival modeling. Trip arrivals can be considered stochastic point processes where arrivals occur at random time intervals within a given period. Based on the clustered and self-reinforcing characteristics of check-in arrivals from LBSN data, a Hawkes process based state propagation model and a state-space framework are introduced to model the stochastic arrival patterns with sampling error feedback. The output can be used into a trip distribution model that applied for dynamic OD estimation.

5. The dynamic demand-activity correlation modeling:

The existing models, such as the gravity models and more recently the radiation models, have difficulties to derive the dynamic OD estimation. Both models are developed for offline planning purposes and do not account for the variations and interactions between OD flow patterns in different time intervals unless the model parameters are calibrated at each time interval. They may not converge in time for the urban dynamic travel demand pattern. In this study, a new approach that estimates OD flows was proposed based on the spatial-temporal correlations of the activity intensity changes at different locations (zones) and at different time intervals.

Furthermore, the scope of my research is restrained by the following criteria.

1. Focus on traffic analysis zone (TAZ) level in travel demand estimation.

2. Focus on both weekday dynamic trip arrival estimation and dynamic OD estimation.

3. Focus on LBSN check-in data provided by Foursquare and Geo-twitter service. LBSN check-in data from other services are not considered.

4. Focus on mimicking the travel demand pattern reported from local transportation planning agency. Other travel demand collection data sources are not considered as the reference data.

5. Focus on modeling the trip arrival pattern and OD flow pattern of a mixed mode of travel including private vehicle, taxi, and transit traffic.

1.3 Research Contributions

My research contributions are as follows:

1. Identify zonal urban functionality using POI categorical density distribution and local mobility pattern.

2. Develop a new numerical simulation procedure for trip arrivals.

3. Develop a new spatial-temporal correlation modeling procedure for that considers temporal delay and the daily recurrent patterns for LBSN check-in arrivals.

4. Develop a land-use based sampling method that has superior performance than the previous LBSN-based travel demand estimation models.

5. Develop a new dynamic trip distribution model that incorporate time-varying features based on the LBSN dataset. The generated OD patterns are at 15-min time intervals for TAZ level to support targeted planning and operational applications.

1.4 Organization of the Thesis

The research schematic diagram is given in Figure 1. Chapters of my proposal follow the procedure of my research career. In Chapter 2, literature reviews of LBSN data application are given. Also, the existing travel demand modeling is reviewed and summarized. The proposed methodology includes four parts of zonal functionality profiling, zonal time-of-day variation modeling, stochastic arrival pattern modeling, and dynamic OD estimation, which are described in details in Chapter 3. For zonal functionality profiling, in Chapter 3.2, I shall propose a topic modeling approach that treat zonal functionality as a latent topic variable to discover from POI categorical density and mobility patterns. For zonal time-of-day variation modeling, in Chapter 3.3, I shall propose a clustering approach for the intensity of trip arrivals by time and by location. The comparison between the point-of-interest (POI) density and land use types is also discussed. For stochastic arrival pattern modeling, in Chapter 3.4, I shall simulate the trip arrivals under zonal time-of-day variation by introducing a stochastic point process based state-space framework with sampling error feedback. Then, in Chapter 3.5, the calibrated trip arrival estimation results are applied to the input of a time-delay trip distribution model to derive the dynamic OD. The spatial-temporal correlation analysis will be discussed. In Chapter 4, I shall focus on experimental design including model calibration and model evaluation methods. Model calibration includes both the parameters calibration of the dynamic trip arrivals/OD estimation models such as thresholds and other basic model settings. All proposed algorithms are also calibrated using the training dataset and tested against the same testing datasets, which are completely different from the training dataset. And the results are summarized and analyzed on their learning capabilities and also the transferability in Chapter 5. And finally, in Chapter 6, the conclusion are drawn and recommendation for application and deployment of the proposed dynamic OD estimation models will be discussed.

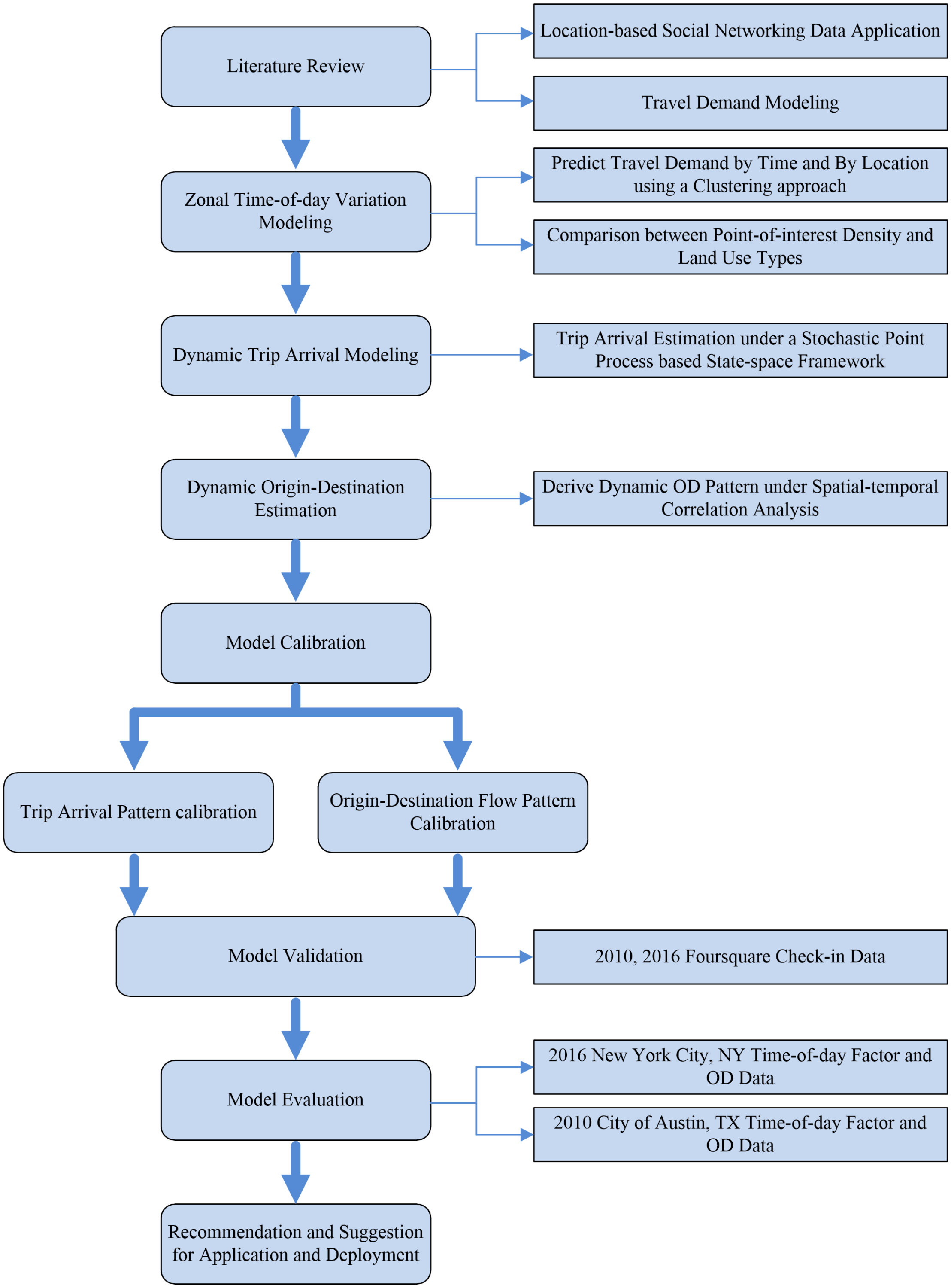

Literature Review
Location-based Social Networking Data Application
Travel Demand Modeling
Zonal Time-of-day Variation Modeling
Predict Travel Demand by Time and By Location using a Clustering approach
Comparison between Point-of-interest Density and Land Use Types
Dynamic Trip Arrival Modeling
Trip Arrival Estimation under a Stochastic Point Process based State-space Framework
Dynamic Origin-Destination Estimation
Derive Dynamic OD Pattern under Spatial-temporal Correlation Analysis
Model Calibration
Trip Arrival Pattern calibration
Origin-Destination Flow Pattern Calibration
Model Validation
2010, 2016 Foursquare Check-in Data
Model Evaluation
2016 New York City, NY Time-of-day Factor and OD Data
2010 City of Austin, TX Time-of-day Factor and OD Data
Recommendation and Suggestion for Application and Deployment

Figure 1: The Flowchart of the Proposed Research.

2 Literature Review

2.1 An Overview of Travel Demand Modeling

The ever-increasing gap between travel demand and transportation system supply causes serious congestion, safety, and environmental issues that spread from large metropolitan areas to medium and smaller size cities and urban areas. In 2010, more than 32,000 fatalities and 2.2 million injuries occurred in more than 5.4 million traffic accidents [13]. According to the TTI Mobility Report [14], on average each U.S. driver loses more than 34.4 hours every year due to traffic congestion. More than 3.9 billion gallons of fossil fuels are wasted in congestion, representing more than 28 of U.S. greenhouse gas emissions and energy consumption [14]. Furthermore, under the widespread budget and resource limitations, expanding the existing infrastructures are no longer viable solutions in many densely populated urban areas. To alleviate the ever-increasing urban congestion issues, transportation researchers and engineers have started to develop the active traffic and demand management (ATDM) strategies [15]. ATDM applies proactive control strategies to effectively divert travel demand and actively manages traffic network before congestion occur. However, ATDM strategies rely on dynamic travel demand information rather than the supply-sided data (e.g., traffic speed, flow, and occupancy) available in existing traffic data sources. The proposed study attempt to address this issue by developing Big Data analytic and transportation models to convert the new data sources into dynamic demand-side information. Such information can enable new ATDM applications that proactively respond to dynamic demand changes so that traffic diversion and harmonization can be executed even before the peak traffic arrives at traffic bottlenecks.

Travel demand modeling was first developed in the 1950s by the metropolitan areas such as Chicago [16] and Detroit [17]. The modeling has followed the sequential four-step model for the conventional urban transportation planning application. The four steps are trip generation, trip distribution, mode choice, and trip assignment. Such structure of the four-step models was later expanded and applied to evaluate the short, medium and long-term consequences of different design and policies [18]. The four-step modelling paradigm is a trip-based approach that uses the individual person trip as the fundamental unit of analysis. The trip-based model that lead to the tour-based scheme and evolved to activity-based model in which individual/household level data is used to model individual/household level travel behavior [19].

Travel demand modeling techniques rely on understanding the travel behavior of people and vehicles. Models are developed based on different levels data sources such as household level, individual level, and regional level. The availability of high-resolution data source provides promising potentials for developing advanced travel demand modeling techniques. The socio-demographic and economic attributes of people [20] can be collected passively and actively to model their travel behavior at different locations and at a different time. The detailed travel diary of a sample of travelers is available to sense the traveling pattern of the whole population. Meanwhile, such individual-level travel diary can be fed into both tour-based and activity-based model for modeling the important travel attributes such as: (a) trip purpose, (b) departure/arrival time, (c) mode of transport, (d) activity duration, (e) activity location, (f) travel route, (g) party composition, and (h) traffic condition.

Table 1 summarizes the characteristics of the latest technology-based primary and secondary data collection methods for travel demand modeling. Primary data collection methods include GPS and smartphone-based travel survey. Secondary data sources include Bluetooth, cell-phone location and location-based social networking (LBSN) data.

Taxi is an important mode of transportation in urban areas, providing parking-free traveling between the pick and drop-off locations. In NYC, taxis serve 172 million

Table 1: Emerging versus Traditional Travel Demand Data Collection Methods.

Attributes	Trad. Survey Method	GPS	Bluetooth	Smart Phone Survey	Cell Phone Signals	Social Media	LBSN Check-in
Spatial Resolution	Low	Low	Low	High	High	High	High
Temporal Resolution	Low	High	High	High	High	High	High
Large-scale Deployment	Yes	No	No	No	Yes	Yes	Yes
Survey/Data Cost	High	Medium	Medium	Medium	Low	Low	Low
Survey Needs	Yes	Yes	No	Yes	No	No	No
Social Demographic Data	Yes	No	No	Inferred	No	Yes	Inferred
Origin-Destination Data	Yes	Yes	Yes	Yes	Yes	Yes	Inferred
Trip Chain	Yes	Yes	Yes	Yes	Yes	Inferred	Inferred
Trip Purpose Confirmation	Yes	Limited	Limited	Yes	Limited	Inferred	Yes
Mode Share	Yes	Inferred	Inferred	Yes	Inferred	Inferred	Inferred
Arrival Time Resolution	Low	High	High	High	High	High	High
Arrival Location Resolution	Low	High	Low	Low	High	Medium	High
Sampling bias	Low	Medium	Medium	Medium	Medium	Yes	Yes
Privacy Concern	No	Medium	No	No	No	Medium	Medium

Non-shaded Characteristics are based on NCHRP Report 735 Table D.2. [21], and previous papers [22, 9].

trips in 2005 made up 11% of trips in the city [23]. According to the 2014 Taxi Fact Book, the annual number of taxi trips has increased to 175 million, and the system has transported 236 million passengers each year from 2008 through 2013.

This demand consists of trips by residents, people who work in the city, tourists, and individuals with disabilities. Modeling taxi demand is necessary to understand how taxi trips are generated and distributed by time and location, and how people choose taxis as their transportation mode. The literature related to taxis demand modeling contributes to congestion effects, customer demand elasticity, policies and regulations on taxi markets and taxi airport ground access. Yang [24] developed a mathematical model to describe taxi movements together with normal traffic in a congested road network. OD demand patterns and traffic assignment procedure for the determination of zone-level taxi and normal traffic movements were discussed. Veloso [25] analyzed taxi-GPS traces and visualize the spatiotemporal variation of taxi services. The relationships between pickup and drop-off location and the waiting time of vacant taxi were also discussed. When applied to the same dataset in this study, research efforts contain urban dynamic [26], travel time variability [27], weather impact on travel time [28], and travel time estimation for different OD pairs [29].

The emerging travel demand data sources can play a significant role to complement the traditional data source that exhausts a large portion of the provided financial resources for planning and operating the transport system. There exist two ways of incorporating with those emerging data source. Researchers and engineers either keep working on innovative approaches to temporally or/and spatially transferring data and models [30] or indirectly imputing the required data from other readily accessible data source [31].

New technologies, especially the LBS and social network technologies, also promote innovations in collected travel demand information. Traditionally, travel demand information is collected through expensive and labor-intensive phone or home interview survey methods that reflect only static travel demand patterns averaging over years. In recent years, new travel demand data collection methodologies have emerged, such as GPS [32], cell phone [33], Bluetooth [1], and now the LBSN [10]. GPS-based methods distribute GPS-survey-enabled devices or applications to volunteers and record their GPS trajectory and daily activities to collect reliable travel demand information [32]. The wireless location technologies (WLT) available through cellular carriers allow passive collection of cellphone positions by tracking wireless signal transition events to identify movement in urban environments. Bluetooth detectors installed at critical locations in a transportation network can also collect dynamic travel demand information [1]. The LBSN-based travel demand analysis is facilitated by the development of the LBS features enabled in smartphones and tablets and the rapid expansion of social networks led by Facebook and Twitter. With billions of people actively updating their personal activities online, the real-time travel patterns can then be derived and lead to more accurate and higher-resolution estimation of urban travel demand than traditional methods. Recently, researchers began to conduct data mining of social networking data to study the spatial pattern of cellphone user behavior. Cheng et al. [34] studied human mobility patterns by analyzing social networking data. The derivation of travel demand information, especially the trip origin and destination information for urban travel, is explored by the investigator team [11, 12, 9, 10]. The radiation model has also been used to model facility choice for non-work trips [35] and, examines human mobility as part of a large-scale spatial-transmission model for infectious [36].

In 1950s, the metropolitan areas like Chicago and Detroit in US first developed and adopted travel demand modelling as the conventional four-step models. Since then, the similar model structure has been used to evaluate the short, medium and long term consequences of different design and policies. The four-step modelling paradigm, which is a trip-based approach, lead to the tour-based scheme in which individual level travel information is regarded for modelling purposes. The tour-based models were later evolved to activity based model in which individual/household level data is used to model individual/household level travel attributes (Activity –based model, eprime). Travel demand modelling techniques target modelling the mobility of people and vehicles in cities to understand their travel behavior. Models are developed based on individual level data sources, in which behavior of travelers is reflected, have been argued to dominate aggregate level model in terms of policy appraisal [30].

The evolution of travel demand modelling techniques developed the need for high resolution databases in which socio-demographic and economic attributes of people are used to model their day-to-day travel behavior. Such data sources encompass travel diary of a sample of people representing the population. Having access to such an individual level travel diary is crucial to develop several components of the advanced behavioral modelling frameworks like tour-based and activity-based.

Data is generally a valuable product which exhausts a large portion of the provided financial resources for planning and operating the transport system. As a result, not necessarily all metropolitan areas can afford collecting data on a monthly or yearly basis. The has resulted in emergent of innovative approaches to temporally or/and spatially transferring data and models [30] or indirectly imputing the required data from other readily accessible data source [31].

Data for demand modelling has been collected using two major methods called: i) revealed preference (RP) surveys and ii) stated preference (SP) surveys. [what is RP, what is SP, what data is collected, what data collection method applied. For example, survey can get counting data, in which GPS or roadside can collected]

2.2 Static and Dynamic OD Estimation

Origin-destination (OD) estimation plays a significant role in studying travel demand pattern; it contains the static or dynamic OD trip patterns. This step develops a matrix that displays trip makers' origin and destinations and the number of trips between each OD pair. OD estimation models predict the best potential destination choices of travelers based on the zonal production and attraction abilities and the level of travel impedance between each OD pair [5]. Therefore, sensing the correlation between origins and destinations is significant to estimate and predict OD patterns. A static OD model is to predict the trip distribution during a targeted time period and do not consider the temporal footnote. Meanwhile, a dynamic OD model contains time-series traffic counts for the static OD pairs. To support Transportation Demand Management applications, dynamic OD estimation models are intended to generate the dynamic travel demand estimation inputs. The dynamic OD estimation models can be classified based on its data sources including flow calibration data, vehicle re-identification data, and GPS survey data. First, one group of the dynamic OD estimation model is derived from the time-series of flow calibration data. Flow data collected from traffic detectors are first used to generate OD information in the literature [37]. The data requires a baseline OD matrix, and the matrix is calibrated to reproduce the dynamic flow through traffic assignment [38]. It explored the correlation between the sequences of entrance flow volumes and the exit flow volumes [6]. The problem can be solved by assuming knowledge of an assignment matrix that defines the temporal and spatial relationships between the link flow and OD volumes [39] or considering split parameters for input-output network relationships [40]. The key of an accurate OD estimation is the accuracy of either the assignment matrix and split parameters based on the intensive historical data.

An alternative dynamic OD estimation data source is the real-time vehicle identification data from automatic vehicle identification (AVI) sensors. Such data collected through Bluetooth readers [1], Electronic Toll Tag Readers [2], and License Plate Readers [3] can identify vehicles at different locations. Through re-identification, the dynamic OD information of vehicles traveling within the covered area or highway systems can then be detected [1]. The key issues with the fixed-location sensor data are the limited geographical coverage, sensor location density, and sampling rate. There are two classes of research efforts defining the dynamic OD estimation problem: the sample OD estimation of AVI tagged vehicle and the population OD estimation using the observed vehicle re-identification data. For AVI tagged vehicle OD demand patterns, studies have developed from using the link-flow proportion matrices [41] to path-flow proportion calculation [42]. For the population OD estimation, it focuses on exploring the sampling rate in either market penetration rates or identification rates that are time-dependent and location-dependent. Asakura et al. [43] applied the least-squares model to estimate the OD demand, the spatial pattern of the identification rate, and the investigated day-to-day OD variations. Zhou and Mahmassani [44] introduced split fractions from AVI tagged vehicle observations and extracted trip distribution information to estimate the population OD estimation using partially observed AVI data.

As the third data source for the dynamic OD estimation, the GPS survey data can have the direct measurement of OD flows with spatial-temporal information [45]. GPS survey-based data [32] can provide positioning information by recruiting travelers to install in-vehicle traffic sensors or mobile applications that keep track of travel activities. However, GPS survey still puts an intensive load on surveyees and require intensives to maintain large sample size and coverage. Similar to AVI data, the GPS-enable devices can provide a fraction of the total number of vehicles, therefore, generate partial information of the population OD flows. Analytic models are developed to convert trajectories of GPS-enable devices into dynamic OD patterns of mobile phone users.

2.3 Social Media Application in Travel Demand Modeling

The capacity of social media platforms such as Facebook, Twitter, LinkedIn, Instagram, Foursquare, and Yelp to provide information on household daily travel has been examined [46, 47]. Tass and Hong [48] presented a wide range of possible ways of using geotagged social media to develop understanding of urban areas instead of using traditional ways of data collection. They categorized the opportunities i) for city planner, ii) for small business owners and iii) for individuals. Social media platforms have a feature known as location-based services, which enable people to share their activity related choices (check-in) in their virtual social networks. Through location0based services, users can share their activity locations when they visit restaurants, shopping malls, movie theatres and so on. Location-based data has received increasing attention, for travel demand modelling as the data can provide further knowledge about travel behaviors. However, the amount of check-in information using such services is less than the geo-tagged associated ‘text’ data available on people’s posts on social media platforms such as Twitter. However, the main challenge before using such rich data is the significant noise existing in them which requires advanced text mining, natural language processing and data mining techniques to extract useful information that can be related to travel behavior of people [49]. Meanwhile, the activity location can be derived from the location-based services, which enable people to share their activity related choices (check-in) in their virtual social networks. The data is named as Location-based Social Networking (LBSN) Data. For example, Foursquare, one of the most popular LBSN services, records users’ check-in when they visit Point-of-Interest named “venue”. The venue can be one restaurant, one shopping mall, or even one bus route. Although the amount of check-in information using such services is less than the geo-tagged associated “text” data. Such location information can be accurately validated against the Foursquare POI database. Furthermore, it can be categorized to generate a confirmed trip purpose information for travel behavior study. We summarized the existing social media data application in travel demand modeling based on the important travel attributes.

First, the trip purpose is one of the most essential travel attributes in travel demand modeling. It can be derived from social media data through either extracting information about the purpose of the activity from the text of tweets or directly review the POI categories of check-in arrivals. The former one required linguistic mining techniques such as Latent Dirichlet Allocation (LDA) method is used widely in the literature (Kosala and Adi 2012, Steiger, Albuquerque et al. 2015). The issue to be addressed is the mismatch between the tweet content and the actual visit. For example, one Twitter user may tweet about a POI such as one restaurant but not necessarily meaning to have an outdoor recreational eating activity there. One solution is to incorporate each tweet with the potential geolocation of the mentioned POI, it can facilitate extracting the purpose of the trip. Similar to trip purpose identification, mode choice can be determined using text mining and natural language processing approaches. Other than looking at words used in a tweet, data sources from LBSN services such as Yelp and Foursquare includes the categorized POI information based on their archived POI database. Researchers and engineers can validate the check-in arrivals' location against the POI location to generate an accurate travel diary. Using the hierarchy category of POIs, the trip purpose can be determined.

Second, determining arrival time, given the fact that tweets and check-in arrivals have a time tag provides the departure time directly. One bias exists due to the fact that an activity happening after or before the time the tweet or the check-in is posted. Considering the travel route and traffic condition, the prediction significantly relies on the individual's usage frequency of social media services. Through analyzing each location or the preceding and succeeding point of tweets and check-in arrivals, it is possible to generate user's traveling route, traffic condition, and travel time as the same way GPS technology is applied.

2.4 LBSN Application in Travel Demand Modeling

LBSN data has several key advantages as a secondary data source including the large urban spatial and continuous temporal coverage, passive data collection, confirmed trip purposes and locations, and anonymization. First, the LBSN data are self-sustained data generated twenty-four hours a day and seven days a week. It involves both the user interests in exploring new POIs and the business owner interests in attracting and maintaining their customer base. Meanwhile, LBSN services have a large-scale dynamically-maintained urban spatial coverage. Users' check-in activity types and timestamps are recorded in the LBSN platform. Users confirm their check-in location with mobile devices or online posts with location tags. Each check-in is linked to a Foursquare venue whose category is defined by venue owners with a three-level Foursquare venue classification system. It can enable the LBSN check-in data to be tightly copied with the trip purpose and land use type. Table 2 and Table 3 lists the trip purposes and land use determined based on the Foursquare venue types. Such extensive spatial-temporal coverage makes the LBSN data useful to understand and model human activity behavior. Secondly, compared to the traditional data collection methods, the LBSN data is a relatively low-cost secondary planning data source. LBSN services are tightly integrated with personal smart mobile devices through mobile applications. The only cost incurred is a data subscription fee. Finally, LBSN service providers implemented a comprehensive privacy protection mechanism. The privacy concern can be addressed through the hiding user ID in public check-in data, POI side aggregation (i.e. only counts at business are posted) and the user consents public information sharing (the Foursquare – Twitter bridge that allows the user to share their Foursquare check-in through Twitter).

Table 2: Trip Purpose Classification.

Type of visited location – Level 1	Trip purpose category
Arts & Entertainment, Food, Outdoors & Recreation, Nightlife Spot	Recreation
College & University	Education
Shop & Service	Retail
Professional & Other Places	Work
Travel & Transport	Transport
Residence	Home
Event	Special event

Type of visited location – Level 2 (e.g. under level 1 category "Outdoors & Recreation")	Trip purpose category
Athletics & Sports, Bathing Area, Bay, Beach, Bike Trail, Botanical Garden, Bridge, Campground, Canal Lock, Canal, Castle...	Recreation

Type of visited location – Level 3 (e.g. under level 1 category "Outdoors & Recreation" and level 2 category "Athletics & Sports")	Trip purpose category
Badminton Court, Baseball Field, Basketball Court, Bowling Green, Curling Ice, Golf Course, Gym / Fitness Center...	Recreation

2.4.1 Existing Trip Arrival Estimation Models based on the LBSN Data

State-of-the-art approaches to predicting trip arrivals fall into three categories. First, in [51], a hierarchical mixture model operates both on spatiotemporal and social information, such as who are the user's social connections, to infer the latent patterns of individual weekly trips, user-specific trips and the trips that were not observed by social media data. The topics of the activities were extracted from the identified trip purpose of the check-in data. Based on the topic modeling, an individual activity

Table 3: Foursquare Venue Classification.

Selected venue in NYC	Venue Type – Level 3	Venue Type – Level 2	Venue Type – Level 1	Trip purpose category	Land use category
West End Apartments	Residential Building (Apartment / Condo)	Building	Residence	Residence / Recreation	Residence
Time Square	Shopping Plaza	Plaza	Shop & Service	Retail / Recreation / Work	Commercial Use
Penn Station	Station	Train Station	Travel & Transport	Transport	Transportation / Utility
Bellevue Hospital Center	Hospital	Medical Center	Professional & Other Places	Work / Maintenance	Public Facilities and Institutions
Central Park	Park	-	Outdoors & Recreation	Recreation	Open Space & Recreation
Icon Parking	Parking	-	Professional & Other Places	-	Parking

Characteristics are based on of New York City zoning and land use data [50]

arrival pattern can be predicted as a probability distribution.

$$P(a_l|w) = \sum_{j=}^{K} P(a_l|z_l = k) * P(z_l = k|w) \tag{1}$$

where K is all latent activity patterns represented by day of week, hour of day, and activity categories (e.g. home, work, eating, entertainment, recreation, shopping, social service or education). The latent variable z_l is the an indicator of activity a_l following a given activity arrival pattern k. For example, $P(z_l = k|w)$ would be, given week w, the probability of an activity a_l following activity arrival pattern k. $P(a_l|z_l = k)$ would be, given an activity a_l following activity arrival pattern k, the probability that the individual has the activity a_l. The results can help to establish the activity-based diaries and show the feasibility of using LBSN data to predict human activity patterns.

Another alternative approach to modeling check-in arrivals is the radiation model [52]. The radiation model was initially proposed to model mobility and migration patterns [35]. It is applied to calculate the intensity of flow T_{ij} from location i to location j as the following.

$$T_{ij} = T_i * \frac{m * n}{(m + po_{ij}) * (m + n + po_{ij})} \tag{2}$$

where T_i is the total number of people that start their journey from location i, m and n represent the population of location i and location j respectively, po_{ij} is the total population in the circle of radius r_{ij} whose centroid is at i (the population of location i and location j is excluded). T_{ij} can indicate the OD flow between two locations. McArdle et al. [52] explored the latent destination of individual movement using the radiation model and LBSN check-in data. Tarasov et al. [53] extended the research and combined the radiation model with social interaction between LBSN users that can be inferred by the reciprocal following on Twitter. In [53], the probability of one check-in arrival at the target location (venue i) is as the following.

$$P(i(t) = i|z(t) = So) = \frac{m * n(venue)}{(m + So) * (m + So + n(venue))} \tag{3}$$

where $i(t)$ is the latent venue of user's check-in arrival at time t, $z(t) = So$ indicates the check-in caused by social influence, and m denotes the number of check-ins made at user's work/home venue, and n is the number of check-ins made at venue i. The venue with the biggest probability is returned as the predicted user location.

Third, the prediction of trip arrivals is the former step of the traditional trip distribution process that normally uses gravity modeling or regression modeling. The check-in arrival data is combined with land use factors and social-demographic information to generate zonal trip production and zonal trip attraction. The trip arrivals can be inferred as trip attraction, which is the input of trip distribution process. Jin [12] converted the daily zonal LBSN check-ins into trip production and attraction fed into a trip distribution model that applied the gravity model. The trip attraction within a zone is a function of the LBSN check-in arrivals and the trip production residual not accounted for by LBSN data. For specific trip purpose p at location i, the estimated trip arrivals $\tilde{A}_{i,p}$ can be formulated as follows:

$$\tilde{A}_{i,p} = \epsilon * x_{i,p} + \frac{x_{i,p}^{\rho}}{\sum_i x_{i,p}^{\rho}} * \sum_i (\omega - \epsilon) * x_{i,p} \tag{4}$$

where $x_{i,p}$ is the total check-ins at location i for trip purpose p, ρ is the power of location factor, ω and ϵ are adjustment factors to zonal trip attraction and production separately from Foursquare check-in counts, and $\frac{x_{i,p}^{\rho}}{\sum_i x_{i,p}^{\rho}}$ redistributes the residual based on the zonal check-in counts. The estimated trip attraction can be used as the input of gravity model to generate the OD matrix.

2.4.2 Urban function

Urban function study is an important research topic for city planners and urban designers to support decision making of city development. Early studies mainly rely on classic theory and case-by-case survey for investigation. Goddard [54] revealed functional regions within the central area of London by measuring the relationship between the taxi flow patterns and the location of human activities. Yuan et al. [55] applied a topic-modeling-based approach based on city-scale positioning data to cluster the segmented regions into functional zones through discovering the users' socioeconomic activities. Litman [56] examined ways to learn the relationship between transportation decision and urban land use patterns including the impact on land use for transportation facilities and the resulting economic, social and environment impacts.

A city consists of a variety of zones providing different functions to support diverse demands of urban residents, such as working, recreation, and residence. Studying the urban functions of city zones provides indispensable information which is useful in solving many urban challenges, therefore plays a critical role in urban analytics. Recent years, the advent of sensing technologies and mobile computing has accumulated a variety of data related to human mobility in urban areas. As a result, data-driven approaches have been increasingly applied to explore urban functions of cities.

While the literature has shown promising effectiveness of analyzing massive positioning data for urban exploration [57, 34], there are limited studies aiming to provide an integrated and principled approach to the representation learning of city zones in terms of urban functions. In this paper, we aim to propose an effective solution to learn the distributed and low-dimensional embeddings of city zones. Zones with similar urban functions are geometrically closer in the embedding space. Using zone embeddings, we are able to identify functional regions of cities which consist of several zones with similar functions. Furthermore, many analytic models can use these extracted representations as enhanced inputs.

Urban function study is an important research topic for city planners and urban designers in a long time for supporting decision making of city development. Early studies mainly rely on classic theory, long-term observation, and case-by-case survey for investigation. The work in [Goddard, 1970] surveys the taxi flow to analyze complex linkage system exists in center of London to study the location of activities. The work in [Putnam, 2001] discusses the change of community in a perspective of people's social interactions. More recently, a series of work [57, 58, 59] use large-scale positioning data to perform data-driven urban function analysis.

2.4.3 Trip Arrival and Stochastic Point Process Model

Trip arrivals can be considered stochastic point processes where arrivals occur at random time intervals within a given period. The existing studies of check-in arrival patterns reveal two key features. First, the time periods between consecutive check-in arrivals are dependent. Check-in arrivals are “clustered” in time and one check-in most likely “excited” the other check-ins afterward [60, 61]. Such characteristics are in-line with the reasons for the thriving development of social networking. Social relations among people who share similar interests, activities, backgrounds or real-life connections bring great social bonding to the LBSN user activities [62]. Second, the repeated behavior of check-ins can be captured by a self-reinforcing process reflected in a user’s most recent behavior [63]. Recently, the above clustered and self-reinforcing characteristics are found to be well-represented by a dynamic model called Hawkes process. Hawkes [64] proposed a self-exciting process model. It was later found to be accurate in modeling earthquake occurrence [65], birth process [66], financial markets [67], seismology [68], and more recently social networks [61]. In Cho et al. [61], the proposed Hawkes point process N for the check-in activities is formulated as follows:

$$\lambda(t) = \mu + \int_{-\infty}^{t} g(t-s) dN(\tau) \tag{5}$$

where (t) is the check-in rate function with respect to t, μ is the background rate of the process N, s is the points at time occurring prior to time t, g is the excitation kernel which parameterizes the self-exciting behavior, and τ is the time interval between two trip arrivals.

The numerical simulation of Hawkes process uses Monte Carlo (MC) simulation due to its various stochastic distributions of arrival times. MC simulation generates random sequences of arrivals to reproduce the cumulative distribution function of the Hawkes process. The thinning algorithm that was introduced by Ogata [69], was used to consider the case that the arrival rate decreases if no more points occur. Table 3 describes the notation used in the simulation. The detailed MC-based Hawkes process simulation algorithm is as follows:

Table 4: Notation Description.

Variables	Description
μ	The minimum value of the arrival rate function
Λ_t^*	Values of a piecewise constant function such that $\lambda(t\|t_1, t_2, \dots, t_n) \leq \Lambda_t^*$ for $t_n \leq s_i \leq t < s_{i+1} \leq t_{n+1}$
s	The points at time occurring prior to time t
t	Time step $t \in T$
T	The simulation period
n	The iteration index
θ	The maximum jump size at each point
τ	The time interval between two arrivals
U	The random value generated from the predefined distribution

1) Set $\Lambda_0^* = \mu$ and put $s_0 = 0$.

2) Generate U_0 and put $\tau_0 = -log_{\frac{U_0}{\Lambda_0^*}}))$.

3) If $\tau_0 \leq T$ then put $t_1 = \tau_0$. Otherwise stop.

4) Set $n_1 = n_2 = n_3 = 0$ and $n_4 = 1$.

5) Set n_3 equal to $n_3 + 1$ and put $\Lambda_{n_3}^* = \lambda(t_{n_4}|t_1, t_2, t_3, ..., t_{n_4-1}) + \theta$.

6) Set n_2 equal to $n_2 + 1$ and generate U_{n_2}.

7) Set n_1 equal to $n_1 + 1$ and generate $\tau_{n_1} = -log(\frac{U_{n2}}{\Lambda_{n_3}^*})$.

8) Put $s_{n_1} = s_{n_1-1} + \tau_{n_1}$. If $s_(n_1) > T$, stop.

9) Set n_2 equal to $n_2 + 1$ and generate U_{n_2}.

10) If $U_{n_2} \leq \lambda(s_{n_1}|t_1, t_2, t_3, ..., t_{n_4-1})$, set n_4 equal to $n_4 + 1$, put $t_{n_4} = s_{n_1}$ and go to step 5.

11) Set n_3 equal to $n_3 + 1$, put $\Lambda_t^* = \lambda(s_{n_1}|t_1, t_2, t_3, ..., t_{n_4-1})$, and go to step 6.

To illustrate the characteristics of the activity patterns generated by the Hawkes process, one sample profile of event arrivals is given by simulating a hypothetical self-exciting point process in Figure 2. The excitation kernel $g(t)$ takes the following

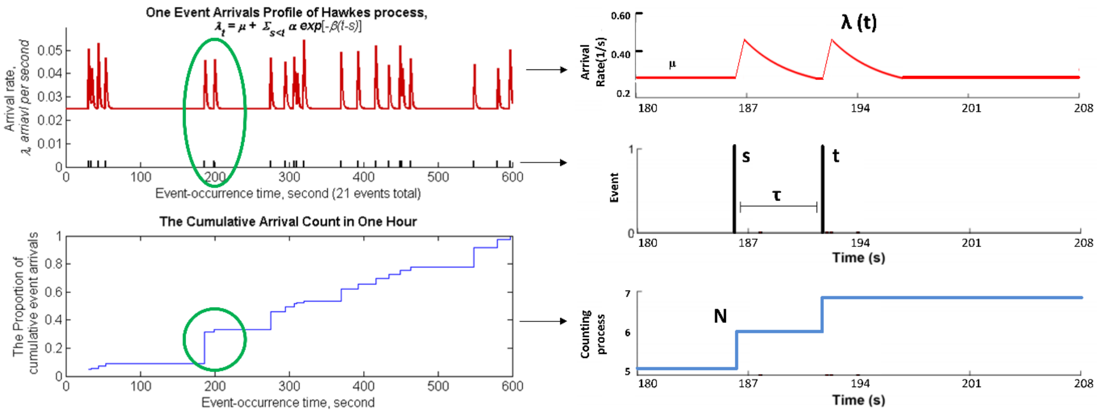


Figure 2: One Sample Profile of Event Arrivals Generated by Hawkes Process.

exponential form:

$$g(t) = \alpha * exp(-\beta * t) \tag{6}$$

where α is a scaling factor for self-excitation $\alpha > 0$ or self-decaying $\alpha < 0$, and β is a positive parameter describing the dependence of recent arrival events on future events. With parameters $\mu = 0.025, \alpha = 0.03, \beta = 0.8$, the "clustering" feature at time is observed and will be used to model demand pulses occurring at trip destinations.

2.4.4 Existing OD Estimation Models based on the LBSN Data

The existing approaches of LBSN-based OD estimation normally use gravity models or regression models. The LBSN data is combined with land use factors and social-demographic information to generate zonal trip production and attraction. Research has been conducted to study the relationship between LBSN data based trips and trip-based travel demand model trips. Jin et al. [12] used Foursquare check-in counts to estimate the urban trip distribution for the city of Austin, Texas. The modeled Origin-Destination matrix was evaluated against the ground truth OD data generated by a trip-based trip distribution model from the planning agency. Yang et al. [22] examined a gravity model to estimate an OD matrix of non-commuting trips based on Foursquare check-in data in the Chicago urban area. Another trip distribution research study based on Twitter data was given by Gao et al. [70]. They were able to detect the regional OD pairs on weekdays from geo-tagged Twitter data compared with the commuting trips from survey data for the Greater Los Angeles metropolitan area. The individual-based trajectory was extracted based on geo-tagged Twitter data to build the peak-hour OD trips at TAZs. Such trips were aggregated at the county level and compared with the survey data for validation regarding the weekday mobility flows. Lee et al. [71] extended the OD estimation algorithm and validated the Twitter-based trips with the four-step travel demand model trips. Regression models with land use factors and social-geographic information were proposed to test the correlations between Twitter-based ODs and traditional travel demand model ODs. The contribution of one Twitter OD trip to the traditional travel demand model OD trips was also discussed. All these research efforts focus on trip distribution of the trip-based models.

2.5 Gravity Model

As the second step of the four-step model, trip distribution is crucial for travel demand modeling. Trip attraction, trip production, and the friction function of the relationship between origins and destinations are applied as the inputs to generate the OD matrix as the output. It represents the travel demand patterns when changes happened within the network. The assumption is about group trip making behavior and the way this is influenced by external factors such as total trip ends and distance traveled [72]. One of the most popular trip distribution models is the gravity model. It is developed by Casey [72] based on Newton's gravitational law in the 1950s and later widely adopted by statewide planning agencies in the US since 1970s. The gravity model was originally formulated as follows:

$$d_{od} = k * \frac{p_o * p_d}{Dist_{od}^2} \tag{7}$$

where d_{od} is the number of trips between zones o and d, P_o,P_d represent the populations at zones o and d respectively, $Dist_{od}$ is the distance between o and d, and k is a proportionality factor. Huff [73] developed the constrained model based on (1). Meanwhile, travel time was applied to indicate the relationship between each ODs rather than distance. Gravity models have been used also to allocate activities. The most elementary form of such model is the residential component of the Lowry model [74]:

$$d_{od} = E_d * f(c_{od}) \tag{8}$$

where d_{od} is the number of workers who live at zones o and d, E_d is the number of workplaces in zones d, and $f(c_{od})$ represents the cost function. In this case, workplaces and residences interact with each other to determine residential location. Subsequent improvements of the model included the use of total trips with origin o and total trips with destination d (O_o and D_d respectively) instead of populations and it was assumed that modelling the effect of distance could be improved by using the generalized travel cost between the zones c_{od}:

$$d_{od} = A_o * B_d * O_o * D_d * f(c_{od}) \tag{9}$$

where A_o and B_d are two constants necessary to guarantee the satisfaction of the constraints on the totals of the trips generated and attracted from each zone:

$$
\begin{aligned}
\sum_{d=1}^{N} d_{od} &= O_o \\
\sum_{o=1}^{N} d_{od} &= D_d \\
A_o &= \frac{1}{\sum_{d=1}^{N} B_d * D_d * f(c_{od})} \\
B_d &= \frac{1}{\sum_{o=1}^{N} A_o * O_o * f(c_{od})}
\end{aligned}
\tag{10}
$$

The equations for A_o and B_d are solved iteratively.

The gravity model is by far the most commonly used aggregate trip distribution model. However, the gravity model has also been criticized: it does not use any explicit individual behavioral theory, it assumes that all information lies in the constraints, its specification does not consider any perception attribute, and it uses an aggregate calibration procedure. Furthermore, it is interesting to note that the O_o's and D_d's are part of the trip generation process [75] and they need to be modeled themselves, as they are a function of some other attributes. Finally, the gravity model assumes knowledge of cost function that defines the spatial relationships between each OD pairs using travel distance; however, it does not consider the time-of-day variation and may not cope with the spatial-temporal correlation to support dynamic OD estimation.

2.6 Temporal Delay Correlation Model

The spatial-temporal correlation model has been extensively studied from three approaches in check-in data-based research. Firstly, studies applied collaborative filtering techniques on check-in data [76, 77, 78]. These studies focused on measuring similarities between locations, such as the visit popularity of a geographic region, and the hierarchical properties of geographic spaces. The user-based collaborative filtering techniques have been extensively applied to support individual location recommendation applications. Secondly, the spatial influence modeling has been widely utilized to improve spatial-temporal correlation analysis. These studies [79, 80, 81, 82] consider spatial information of current locations and the travel distance of visited locations to determine the travelers' potential destination choice. Meanwhile, temporal influence modeling has been widely used to identify the temporal periodic patterns of check-in behaviors. Some research efforts [83, 84, 85] proposed discrete time slots, then separately modeled the temporal influence for each slot based on collaborative filtering techniques. Some research dynamically integrated both spatial and temporal influence models. Cho et al [63] proposed a time-aware Gaussian Mixture model combining periodic short-range movements and sporadic long-distance travels. Wang et al [86] provided a Regularity Conformity Heterogeneous (RCH) model to predict user location at specific times, considering both regularity and conformity. Lian et al. [87] incorporated temporal regularity into a Hidden Markov framework to predict regular user locations. Finally, taking advantage of sequential patterns in human movement [63], various sequential mining techniques [88, 86, 89] have been developed for location predictions based on the sequential pattern of individual's visit. Chong et al explored Latent Dirichlet Allocation (LDA) topic models for venue prediction given users' history of other visited venues.

The accuracy of trip distribution modeling relies on the accurate representation of the relationship between each OD pair. In gravity model, the cost function indicates the travel impedance from one location to another. For dynamic OD estimation, it required an advanced technology to cope with the spatial-temporal correlation. Recently, the above characteristics are found to be well-represented by a called Pearson product-moment correlation coefficient. that the model was developed by Pearson

(Pearson 1900) and was later found to be accurate in measuring the linear dependence between two variables to capture the spatiotemporal dynamics of internet traffic [90]. In Bourke [91], when applied to a sample containing two arrays $\{X|x_1, x_2, ..., x_n\}$ and $\{Y|y_1, y_2, ..., y_n\}$, the Pearson product-moment correlation coefficient r can be obtained as follows:

$$r = \frac{\sum_n (x_i - m_x) * (y_j - m_y)}{\sigma_x * \sigma_y} \tag{11}$$

where x_i and y_j represents the ith and jth element of the array X and Y respectively, m and σ are the mean value and stand deviation of the arrays respectively. r gives a value between 1 and -1 inclusive, where 1 is total positive linear correlation, 0 is no linear correlation, and -1 is total negative linear correlation.

Meanwhile, the LBSN data can passively collect the detailed trip arrival timestamp at the destination location, however, it does not provide the trip departure timestamp. The time difference between two check-in arrivals can be regarded as a time delay that separates the occurrence of two activities. As the critical input for the dynamic OD estimation, it is needed to consider both the travel time and activity duration to generate the time delay. Hamed and Mannering [92] made a hypothesis that the travel time to the targeted activity and the duration of that activity are interrelated; that is, travelers who travel greater distances to participate in a particular activity are more likely to spend more time in that activity than those who travel shorter distances. Bowman and Ben-Akiva [93] made the categorization of time of day decision and destination choice of the primary activity as the explanatory variables in the activity duration estimation equation. Therefore, the activity duration model can be expressed as a function of activity duration τ_{ad}, travel time τ_{ad} from the last activity destination to the current activity destination and the activity type indicator AT for POI category.

$$\tau_{ad} = f(\tau_{tr}, AT) \tag{12}$$

2.7 Summary of LBSN-based Travel Demand Modeling

LBSN-based travel demand modeling relies on travel behavior modeling to sense the travel demand pattern. In dynamic trip arrival modeling, depending on the number of LBSN check-in arrivals, travel demand pattern can be decided spatially and temporally with respect to zonal LBSN check-in arrival data. In dynamic trip distribution modeling, the key is to identify a time-varying characteristic from the LBSN check-in data for dynamic OD estimation.

3 Proposed Methodology

3.1 Preliminary Definitions

$Definition 1.(Zone correlation)$. Given an origin zone and a destination zone, a zone correlation is a quantity measuring the extent of interdependence between the origin's outflow and destination's inflow.

$Definition 2.(Trip arrival)$. For a zone at a time slot, trip arrivals are the number of trip counts that arrived in this zone; aggregated arrivals can describe general human mobility patterns.

$Definition 3.(Check-in arrival)$. For a zone at a time slot, check-in arrivals represent the number of mobile users who visited a POI at this zone reported by check-in data.

3.2 Zonal Functionality Profiling

To infer the zonal functionality, we introduced M zonal types determined by analyzing the POI distributions $D_i = \{d_{i,c}\}$. First, we calculated the POI density $d_{i,c}$ of each POI category c at zone i:

$$p_{i,c} = \frac{numbers\ of\ POI_{c \in C}}{area\ of\ zone\ i} \tag{13}$$

Secondly, we integrated the zonal human mobility events given the dataset of taxi departure and arrival records. Each trip contains the passenger travel with the information of locations and timestamps for the departure and the arrival. From taxi trips, we extract a set of time-of-day zonal human mobility distribution $H_i = \{h_{i,t}\}$, where each zonal human mobility pattern $h_i out_t$,$h_i in_t$ includes the time-of-day taxi departures and arrivals from/at zone i calculated as follows:

$$\begin{aligned} h_i in_t &= \frac{numbers\ of\ taxi\ trips\ departured\ from\ zone\ i\ at\ time\ t}{daily\ taxi\ trips\ departured\ from\ zone\ i} \\ h_i out_t &= \frac{numbers\ of\ taxi\ trips\ arrived\ at\ zone\ i\ at\ time\ t}{daily\ taxi\ trips\ arrived\ at\ zone\ i} \end{aligned} \tag{14}$$

where each timeslot is converted from a timestamp to a $< hourofday, dayofweek >$ combination. The objective is to extract the distributed and low-dimensional embeddings $\{d_{i,h_i}\}$ of city zones based on the spatiotemporal human mobility patterns for representing their urban functions in a city.

Using Latent Dirichlet allocation (LDA) method, we treat the zone functionalities $m \in \{1, 2, ..., M\}$ as the document topics, the zones as documents (each zonal POI distribution D_i and time-of-day human mobility patterns H_i as one documentation), and POI categorical density $d_{i,c}$ and time-of-day mobility patterns $h_{i,t}$ as words in the documentation. Then the zone functionality can be uncovered from the POIs and taxi trip dataset. We construct a hierarchical topic model following:

$$\begin{aligned} \theta_i \sim Dir(\eta_1); z_{i,d_{i,c},h_{i,t}} \sim Multinomial(\theta_i); \\ \varphi_m \sim Dir(\eta_2); \omega_{i,d_{i,c},h_{i,t}} \sim Multinomial(\varphi_m) \end{aligned} \tag{15}$$

where η_1 and η_2 are the prior Dirichlet parameters on the per-document topic distribution and word distribution, θ_i represent the topic distribution for zone i, φ_m is the word distribution for topic m, $z_{i,d_{i,c},h_{i,t}}$ and $\omega_{i,d_{i,c},h_{i,t}}$ are the chosen topic and word for zone i. Then the probability of POIs within one zone being covered by zone type

m is:

$$Pro.(m|P_i) = (\prod_m Pro.(\varphi_m|\eta_2))(\prod_i Pro.(\theta_i|\eta_1) \\ \prod_c Pro.(z_{i,d_{i,c},h_{i,t}}|\theta_i) Pro.(\omega_{i,d_{i,c},h_{i,t}}|\varphi_{1:M}, z_{i,d_{i,c},h_{i,t}})) \tag{16}$$

3.3 Zonal Time-Of-Day Variation Modeling

In a previous study [12], the research group identified that the accuracy of trip arrival estimation is the key for the accurate estimation of OD matrices with LBSN data. When considering the model used as equation (4), the concern of directly deriving such a time-of-day model that is a daily model with parameter variations for different trip purposes is the large set of parameters for different times of day. First, a zonal time-of-day variation model is generalized from equation (4) by ignoring the production residual balancing term $\frac{x^{\rho}_{i,p}}{\sum_i x^{\rho}_{i,p}} * \sum_i (\omega - \epsilon) * x_{i,p}$. Meanwhile, we modify the static parameter ω as the time-dependent one $\sigma_p(t)$, which is built for different hours of the day. Given 24 hours within a day, the proposed statistics model with 24 parameters representing the TOD variations is as follows:

$$\tilde{A}_{i,p}(t) = \sigma_p(t) * x_{i,p}(t) \tag{17}$$

where $\tilde{A}_{i,p}(t)$ is the trip arrival estimation at location i for trip purpose p at time t, $x_(i,p)(t)$is the check-in counts at location i for trip purpose p at time t, and $\sigma_p(t)$ is the ratio of trip arrivals to Foursquare check-ins for trip purpose p at time t.

The analysis needs to solve the following equations for balancing the LBSN check-in activities and trip arrivals between time of day variations and zonal differences,

$$Min. \sum_i abs(\sum_t \tilde{A}_{i,p}(t) - A_{i,p}) + \sum_t abs(\sum_i \tilde{A}_{i,p}(t) - A_p(t)) \qquad (18)$$

Where $\tilde{A}_{i,p}(t)t$ is the estimated trip arrival at location i for trip purpose p at time t, $A_{i,p}$ is the reference trip arrival at location i for trip purpose p aggregated in one day, and $A_p(t)$ is the reference trip arrival for trip purpose p aggregated in studied area at time t.

Two limitations exist in this model. First, the model may result in the need for calibrating as many parameter values $\sigma_p(t)$ as the number of the TOD regimes. Second, a model over-simplifies the relationship between LBSN check-in counts and trip arrivals as proportional rather than as two correlated random point processes.

3.4 Dynamic Trip Arrival Modeling

The direct use of the Hawkes model will bring a static pattern across an entire day. To capture the dynamic time-of-day trip arrival patterns, we propose a time-dependent trip arrival estimation model as follows:

$$\tilde{A}_{i,p}(t) = F(\tilde{A}^H_{i,p}, \tilde{A}^c_{i,p}, t, \alpha, \beta, \gamma, \delta) \qquad (19)$$

where $F(\tilde{A}^H_{i,p}, \tilde{A}^c_{i,p}, t, \alpha, \beta, \gamma, \delta)$ is a function of the trip arrivals $\tilde{A}_{i,p}t$ at location i for trip purpose p at time t, the estimated trip arrivals $\tilde{A}^H_{i,p}$ through the Hawkes process using the trip arrival estimation in the previous time interval $(t - \triangle t, t)$ at time t as the input, the estimated trip arrivals through LBSN observation $\tilde{A}^c_{i,p}$, and the set of parameters $(\alpha, \beta, \gamma, \delta)$ to be calibrated. The trip purpose p can be tied to one trip purpose category in the origin-destination data of planning agencies. Each zonal trip arrival is an event occurrence following the Hawkes point process where the arrival rate can be formulated similar to equation (4),

$$\lambda(t) = \mu + \sum_{s:s<t} g(t-s) \tag{20}$$

where s represents the points at time occurring prior to the check-in arrivals time t. The trip arrival rate from the previous time interval is applied as the background rate μ in the adaptive point process model. The excitation kernel $g(t)$ takes the same exponential form shown in equation (6). Therefore, the arrival rate function of Hawkes process $\lambda_{i,p}(t)$ for check-ins at location i with trip purpose p at time t takes the following form:

$$\begin{aligned} \lambda_{i,p}(t) = \mu_{i,p}(t-\triangle t) + \alpha * \sum_{s:s<t} exp(-\beta * (t-s)) \\ \mu_{i,p}(t-\triangle t) = \frac{\tilde{A}_{i,p}(t-\triangle t)}{\triangle t} \end{aligned} \tag{21}$$

where $\mu_{i,p}(t-\triangle t)$ is the trip arrival rate from the previous time interval between $t-\triangle t$ and t at location i for trip purpose p. Considering the case at one location for one trip purpose, the standard MC algorithm for the Hawkes process is then modified accordingly to simulate the proposed adaptive process as follows.

1) Set $\Lambda_0^* = \mu$ and $s_0 = t = 0$.

2) Generate an exponential random number τ_0 with the mean arrival rate $1/\Lambda_0^*$.

3) Put $t = t + \tau$ If $t > T$, stop; otherwise, set $s_1 = t$

4) Set $n_1 = n_2 = n_3 = 0$ and $n_4 = 1$.

5) Set $n_3 = n_3 + 1$and put $\Lambda_{n_3}^* = \lambda(t|t_1, t_2, t_3, ..., t_{n_4-1}) + \theta$

6) Set $n_2 = n_2 + 1$ and generate a standard uniform random number U_{n_2}

7) Set $n_1 = n_1 + 1$ and generate an exponential random number τ_{n_1} with mean arrival rate $1/\lambda_k$

8) Put $s_{n_1} = s_{n_1-1} + \tau_{n_1}$. If $s_{n_1} > T$, stop;

9) Set $n_1 = n_1 + 1$ and generate a standard uniform random number U_{n_2}

10) If $U_{n_2} \leq \lambda(s_{n_1}|t_1, t_2, t_3, ..., t_{n-1})/\Lambda^*_{n_3}$, set $n_4 = n_4 + 1, t_{n_4} = s_{n_1}$ and go to step 5

11) Otherwise, set $n_3 = n_3 + 1$, put $\Lambda^*_t = \lambda(s_{n_1}|t_1, t_2, t_3, ..., t_{n_4-1})$ and go to step 6.

12) All the s_{n_1} are the arrival times wanted.

The modifications compared with the origin Hawkes process simulation algorithm are as follows:
Step 2, we use the mean arrival rate of previous time interval as the initial value of arrival rate function $\Lambda^*_0 = \mu$ to generate the first arrival time $t = t + \tau$ in current time interval.
Step (5), we remove the maximum jump size θ for the generation of Λ since we use the mean arrival rate to avoid extreme value of τ.
Step (6), U was generated as a standard uniform random number.

After the simulation, the number of trip arrivals estimated through the Hawkes process between time intervals t and $t + \triangle t$ can then be calculated by

$$\tilde{A}^H_{i,p}(t) = \int_{-\infty}^{t+\triangle t} \lambda_{i,p}(\tau) d\tau \tag{22}$$

where $\tilde{A}^H_{i,p}(t)$ is the trip arrival estimation through the Hawkes process at location i for trip purpose p at time t, and τ is the estimated sequence of arrival times. Given the TOD variations in daily demand fluctuation, the above models are built for different hours of day. Within each hour, the proposed Hawkes process model simulates all actual trip arrivals. To reduce the number of Hawkes models to be calibrated, hours of day with similar characteristics will be further clustered into groups that share the same Hawkes process parameters.

The LBSN check-ins are only a fraction of the total actual arrivals. To formulate such an "observation" process, we propose a model that describes how actual arrivals are captured by LBSN check-in counts. To account for potential sources of the sampling bias, we introduce a clustering analysis to build the zonal LBSN check-ins observation model as follows.

$$\tilde{A}^{c}_{i,p}(t) = G(x_{i,p}(t), t) \tag{23}$$

where $G(x_{i,p}(t), t)$ is a function of the trip arrival estimation through LBSN check-ins observation $x_{i,p}(t)$ at location i for trip purpose p at time t, the number of check-in arrivals $x_{i,p}(t)$ observed at location i for trip purpose p at time t, and the set of parameters to be calibrated. The equation can be expanded as follows:

$$\tilde{A}^{c}_{i,p}(t) = \gamma_p(t) * x_{i,p}(t) \tag{24}$$

Where $\gamma_p(t)$ is the converting factor of Foursquare check-in counts to actual trip arrivals for trip purpose p at time t.

Compared to the equation (13), the proposed models introduce the potential to use different point process techniques for different regimes and improves the accuracy and resolution of the estimation. However, to reduce the number of parameters to be calibrated, we introduce a time of day clustering and recalibration process. First, the model for each hour of a day is calibrated separately against the planning trip arrivals statistics. Then, hours of day with parameters, i.e. similar sampling bias characteristics, are grouped together and recalibrated to share the same parameter set. An objective function is introduced to find the optimal clustering based on the initial individual hour-of-day model calibration.

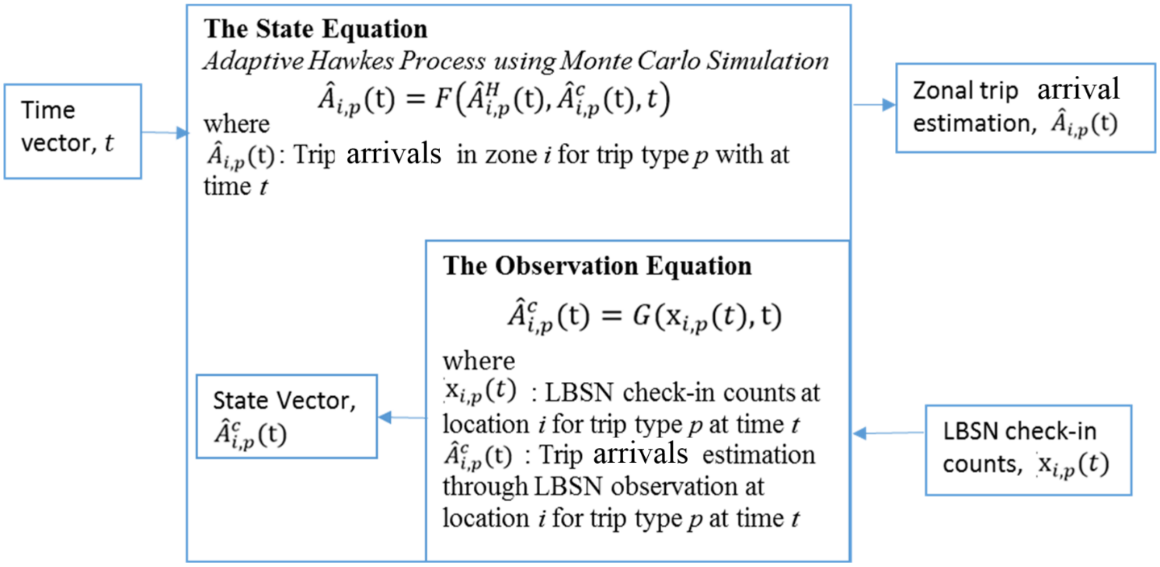

Time vector, t
The State Equation
Adaptive Hawkes Process using Monte Carlo Simulation
$\hat{A}_{i,p}(\mathrm{t}) = F\left(\hat{A}^{H}_{i,p}(\mathrm{t}), \hat{A}^{c}_{i,p}(\mathrm{t}), t\right)$
where
$\hat{A}_{i,p}(\mathrm{t})$: Trip arrivals in zone i for trip type p with at time t
Zonal trip arrival estimation, $\hat{A}_{i,p}(\mathrm{t})$
The Observation Equation
$\hat{A}^{c}_{i,p}(\mathrm{t}) = G(\mathrm{x}_{i,p}(t), \mathrm{t})$
where
$\mathrm{x}_{i,p}(t)$: LBSN check-in counts at location i for trip type p at time t
$\hat{A}^{c}_{i,p}(\mathrm{t})$: Trip arrivals estimation through LBSN observation at location i for trip type p at time t
State Vector, $\hat{A}^{c}_{i,p}(\mathrm{t})$
LBSN check-in counts, $\mathrm{x}_{i,p}(t)$

Figure 3: Block Diagram Representation of the State-space Equations.

$$Min.abs(\frac{h * (\overline{\gamma(t|t \in H)} - \overline{\gamma_p(t)})^2}{(\gamma_p(t) - \overline{\gamma_{H,p}})^2} - 1) + \\ abs(\frac{\sum_i x_{i,p}(t) * \gamma(t|t \in H, p)}{\sum_t \sum_i x_{i,p}(t) * \gamma(t|t \in H, p)} - \frac{\sum_i A_{i,p}(t)}{\sum_t \sum_i A_{i,p}(t)}) \quad (25)$$

where H is the index of clusters, h represents the number of TOD regimes (hours of day) in cluster H, $\overline{\gamma_{H,p}}$ denotes the mean of scaling factor with respect to TOD regime t, $\sum_i x_{i,p}(t)$ and $\sum_i A_{i,p}(t)$ denotes the hourly check-in arrivals statistics and the CAMPO data for hourly trip arrivals in the study TAZs respectively.

The proposed model works in a two-step process following the state-space framework. The observation equation converts the number of LBSN check-ins into the number of trip arrivals. The state equation uses the estimated trip arrivals through the Hawkes process in the previous time interval $\tilde{A}^{H}_{i,p}$) and the feedback from the estimated trip arrivals through LBSN check-in observation in the current time interval $\tilde{A}^{c}_{i,p}$) to estimate trip arrivals in the new time interval. Figure 3 shows the state-space framework of the proposed Hawkes process model.

3.5 Dynamic Origin-Destination Estimation

The PPMC analysis was applied to measure the similarity between check-in arrival patterns at two locations. Let vectors o_i and d_j represents check-in arrival sequence that consist of a set of check-in arrivals $\{X|x_i^t, x_i^{t+1}, ..., x_i^{t+w}\}$ and $\{Y|y_i^t, y_i^{t+1}, ..., y_i^{t+w}\}$. The o_i and d_j are normalized as follows.

$$\begin{aligned} f_i^t(w) &= \frac{(o_i^t(w) - m_o)}{\sigma_o} \\ f_j^s(w) &= \frac{(d_j^s(w) - m_d)}{\sigma_d} \end{aligned} \tag{26}$$

where t is the time slot and $s = t + \tau$ indicates the time slot after the time delay τ, w is the selected sequence length which represents the daily recurrent patterns of check-in arrivals, m and σ are the mean and standard deviation of the corresponding sequences respectively. The equation of computing time delay correlation coefficient takes the following form similar to equation (3).

$$r_{ij}^{ts} = f_i^t(w) * f_j^s(w) \tag{27}$$

where r_{ij}^{ts} is a vector that contains $(2 * w - 1)$ elements and $s = t + \tau$ indicates the time slot after the time delay τ. The correlation vector r_{ij}^{ts} for OD pair ij at different time delays τ can be used to identify the most likely OD flow potentials. The activity duration may vary according to the individual activities and the travel time. Regarding equation (2), the total time delay τ and the duration of activity is represented by the multivariate regression equation of the following form:

$$\begin{aligned} \tau &= \tau_{tr} + \tau_{ad} \\ \tau_{ad} &= a + b * AT + c * \tau_{tr} \end{aligned} \tag{28}$$

where τ_{tr} is the travel time from location i to location j, τ_{ad} is the activity duration that represents the dwelling time at the check-in arrival location, AT is the vector of activity type indicator for different venue types of the check-in arrivals, and a, b, c are parameters needs to be calibrated. Furthermore, the model considered both the positive and negative correlations. The positive correlation indicates the check-in arrival patterns at destination may replicate the ones at origin after a time delay. The negative correlation is considered to indicate the potential OD flow when the check-in arrivals' reduction at the origin may contribute to one increase at the destination. When the correlation is calculated between a sequence and a lagged version of itself, an autocorrelation will be produced. A high correlation coefficient indicates a periodicity of check-in arrivals' pattern within the location in the corresponding time interval. The autocorrelation coefficient is computed as follows.

$$r_{ii}^{ts} = f_i^t(w) * f_i^s(w) \tag{29}$$

Therefore, the PPMC modeling procedure is to generate the correlation matrix as follows.

Table 5: Notation Description.

Variables	Description
i,j	Location index
s,t	The time slot index
τ	The time delay value based on the sum of the travel time between the origin and destination locations and the activity duration
o_i, d_j	The check-in arrival sequence at location i and at location j
T	The experiment time period
N	The total number of TAZs
r	The correlation coefficient vector

1) set $i = 1$ as origin location number, $j = 1$ as destination location number, $t = 1$.

2) If $i > N$, stop; otherwise set τ as the time delay and set $s = t + \tau$

3) Put vector $o_i = \{x_i^t, x_i^{t+1}, ..., x_i^{t+w}\}$ and vector $d_j = \{x_i^s, x_i^{s+1}, ..., x_i^{s+w}\}$.

4) If $i = j$, Generate auto correlation coefficient r_{ii}^{ts} with mean and stand deviation of sequence; Otherwise, generate correlation coefficient vector r_{ij}^{ts} with mean and stand deviation of sequence.

5) If $j > N$, set $i = 1 + i$, $j = 1$ and go to step (2); otherwise, set $j = 1 + j$ and go to step (2).

6) All the r are the coefficients wanted.

The modifications compared with the origin Pearson product-moment correlation algorithm are as follows:

Step (1), we add the timestamp index t to indicate that the model involves a dynamic estimation.

Step (2), we add the time delay value τ to represent that the comparison is between the different time-of-day trip patterns.

Step (3), we add the length value w of selected sequence that considers the daily recurrent patterns of check-in arrivals.

Consider both the mixture of the HPSS formulation and PPMC coefficient, we jointly predicted dynamic OD flow patterns by incorporating the predicted dynamic trip arrivals A_i^t,A_j^d and PPMC coefficient rcc_{ij}^{td} as follow.

$$Pr_{ij}^{td} = \frac{A_i^t A_j^d g(rcc_{ij}^{td})}{\sum_j A_j^d g(rcc_{ij}^{td})} \tag{30}$$

where Pr_{ij}^{td} stands for the probability of a trip made from zone i at time slot t to zone j at time slot d, A_i^t are predicted trip arrivals of zone i at time t, and $g(rcc_{ij}^{td})$ is the travel cost function which considers the PPMC coefficient.

Given the reference OD flow matrix $F = \{f_{ij}\}$, we sample $N = \sum_{i,j} f_{ij}$ trips following the predicted probability Pr_{ij}^{td} and different types of constraints to generate the OD flow matrix $\tilde{F} = \{\tilde{f}_{ij}\}$. We consider four different types of constraints of the proposed joint PPMC-GM model:

Unconstrained model (UM). The only constraint of UM is to ensure the total number of predicted trips $\tilde{N} = \sum_{i,j,t,d} \tilde{f}_{ij}^{td}$ is equal to the total number of trips N in the reference data. The N trips are randomly sampled from the multinomial distribution.

$$Multinomial(N, (Pr_{ij}^{td})) \tag{31}$$

Singly-production-constrained model (PCM). PCM is to ensure the total number of predicted origin zone's trips $O_i = \sum_j f_{ij}$ is preserved. For each origin zone i, the O_i trips are randomly sampled from the multinomial distribution.

$$Multinomial(O_i, \frac{\sum_{t,d} Pr_{ij}^{td}}{\sum_{j,t,d} Pr_{ij}^{td}}) \tag{32}$$

Singly-attraction-constrained model (ACM). ACM is to ensure the total number of predicted destination zone's trips $d_j = \sum_i f_{ij}$ is preserved. For each destination zone j, the D_j trips are randomly sampled from the multinomial distribution.

$$Multinomial(D_j, \frac{\sum_{t,d} Pr_{ij}^{td}}{\sum_{i,t,d} Pr_{ij}^{td}}) \tag{33}$$

Doubly-constrained model (DCM). DCM is to ensure the number of both origin's and destination zone's trips is preserved. For each origin zone i and destination zone j, the N trips are randomly sampled from the multinomial distribution.

$$\tilde{f}_{ij}^{td} = B_i B_j Pr_{ij}^{td}; \sum_{j,t,d} \tilde{f}_{ij}^{td} = O_i; \sum_{i,t,d} \tilde{f}_{ij}^{td} = D_j; \tag{34}$$

$$Multinomial(N, \frac{\sum_{t,d} \tilde{f}_{ij}^{td}}{\sum_{i,j,t,d} \tilde{f}_{ij}^{td}}) \tag{35}$$

where B is the calibrated balancing factors with the Iterative Proportional Fitting procedure.

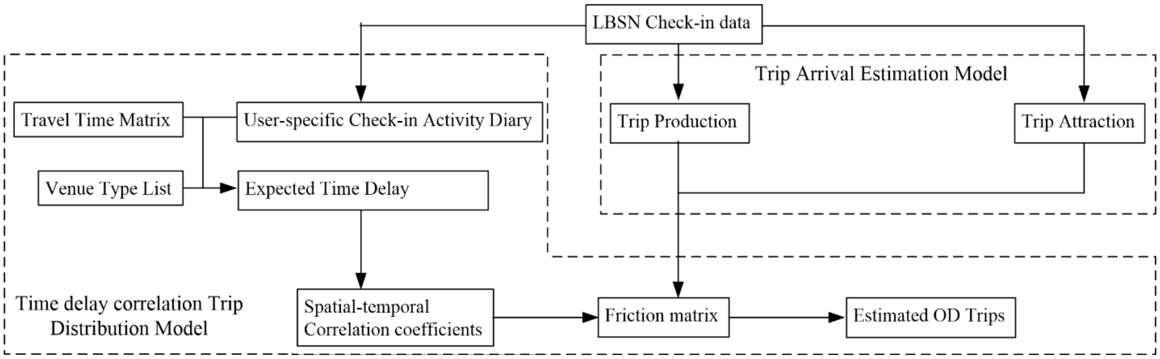

LBSN Check-in data
Trip Arrival Estimation Model
Travel Time Matrix
User-specific Check-in Activity Diary
Trip Production
Trip Attraction
Venue Type List
Expected Time Delay
Time delay correlation Trip Distribution Model
Spatial-temporal Correlation coefficients
Friction matrix
Estimated OD Trips

Figure 4: Block Diagram Representation of the Time-delay-correlation Gravity Model.

Finally, the predicted dynamic OD flows are aggregated based on the modelled zone functions of origin and destination zones for each OD pair and the time of day. Figure 4 shows the two-stage framework of the proposed methodology.

4 Experimental Design

The experimental design includes the detailed design for model validation, calibration, and evaluation.

4.1 Performance Measures

The performance measure includes two parts. First, we predicted the dynamic trip arrivals and compared the result of the spatial and temporal pattern with the reference data. The reference data contains the zonal trip arrivals derived from OD matrices and the time-of-day factor. Based on existing trip arrival estimation and machine learning literature, there are several important performance measures. These measures are critical in the optimization procedure for model parameters during validation, calibration and evaluation procedure. We used two common objective functions, including the Root Mean Square Error (RMSE) (Pohlmann and Friedrich 2013) and the Mean Absolute Error (MAE) between the estimated and agency trip arrivals to conduct geospatial and statistical performance comparisons statistically and while working with the RMSE to indicate the variation in the errors. Since the overall performance of models may have issues with the high variance of a small percentage of the high trip arrival counts, the MOEs were generated based on the trip arrival counts with different levels (e.g. 10000 trips per day with 1000 RMSE value versus 10 trips per day with 1 RMSE value).

The two indicators are defined as follows:

$$\begin{aligned} RMSE &= \sqrt{\frac{\sum_i (\bar{A}_i - A_i)^2}{I_{level}}} \\ MAE &= \frac{\sum_i abs(\bar{A}_i - A_i)}{I_{level}} \end{aligned} \qquad (36)$$

Where I_level is the number of Traffic Analysis Zones in the study area with respect to the level of zonal trip arrival counts.

Given the calibrated result of dynamic trip arrival estimation, we predicted the dynamic OD and compared the result of the spatial and temporal pattern with the reference data. We compared our proposed approach with the following baseline methods.

Normalized Gravity Model with exponential distance decay (NGravExp). In this popular form of the gravity model, the probability of a trip between zone i and zone j is proportional to the outflow of the origin zone O_i and the inflow of the destination zone D_j, and is inversely proportional to the travel cost $cost_{ij}$ between the two zones, which is modeled with an exponential distance decay function:

$$Pr_{ij} = \frac{O_i D_j g(cost_{ij})}{\sum_j D_j g(cost_{ij})}; f(cost_{ij}) = exp(-\beta distance_{ij}) \tag{37}$$

Normalized Gravity Model with power distance decay (NGravPow). Unlike the NGravExp model, the NGravPow considers travel cost modeled with a power distance decay function:

$$Pr_{ij} = \frac{O_i D_j g(cost_{ij})}{\sum_j D_j g(cost_{ij})}; f(cost_{ij}) = (distance_{ij})^{-\beta} \tag{38}$$

Schneider Intervening Opportunity Model (Schneider). In this model, the probability of a trip from zone i to zone j is proportional to the outflow of the origin zone and to the conditional probability that a traveler departure from zone i with outflow O_i is attracted to zone j, given that there are S_{ij} populations in between:

$$Pr_{ij} = \frac{exp(-\beta S_{ij}) - exp(-\beta(S_{ij} + O_i))}{\sum_j exp(-\beta S_{ij}) - exp(-\beta(S_{ij} + O_i))} \tag{39}$$

Radiation Model (Rad). Simini et al. [35] reformulated the intervening opportuni-ties model in terms of radiation and absorption processes:

$$Pr_{ij} = \frac{O_i D_j}{(O_i + S_{ij})(O_i + D_j + S_{ij})} \tag{40}$$

We used Mean Absolute Error (MAE), Normalized Root Mean Square Error (NRMSE), and Coincidence Ratios (CR) as metrics to evaluate the performance of zonal OD estimation:

$$MAE = \frac{\sum_{i,j} abs(f_{ij} - \tilde{f}_{ij})}{\sum_{i,j} 1} \tag{41}$$

$$NRMSE = \frac{\sum_{i,j} abs(f_{ij} - \tilde{f}_{ij})^2}{\sum_{i,j} f_{ij}} \tag{42}$$

$$CR = \frac{\sum_k min(\tilde{tl}_{distance_k}, tl_{distance_k})}{\sum_k min(\tilde{tl}_{distance_k}, tl_{distance_k})} \tag{43}$$

where $tl_{distance_k}$ represents the percentage of trips in interval k of trip length distance, CR measures the common area of the trip length distribution for the predicted and ground truth OD matrices. The result takes the value in $[0, 1]$. When $CR = 0$, two distributions are completely different; while $CR = 1$, two distributions are identical.

4.2 Data Source and Preliminary Analysis

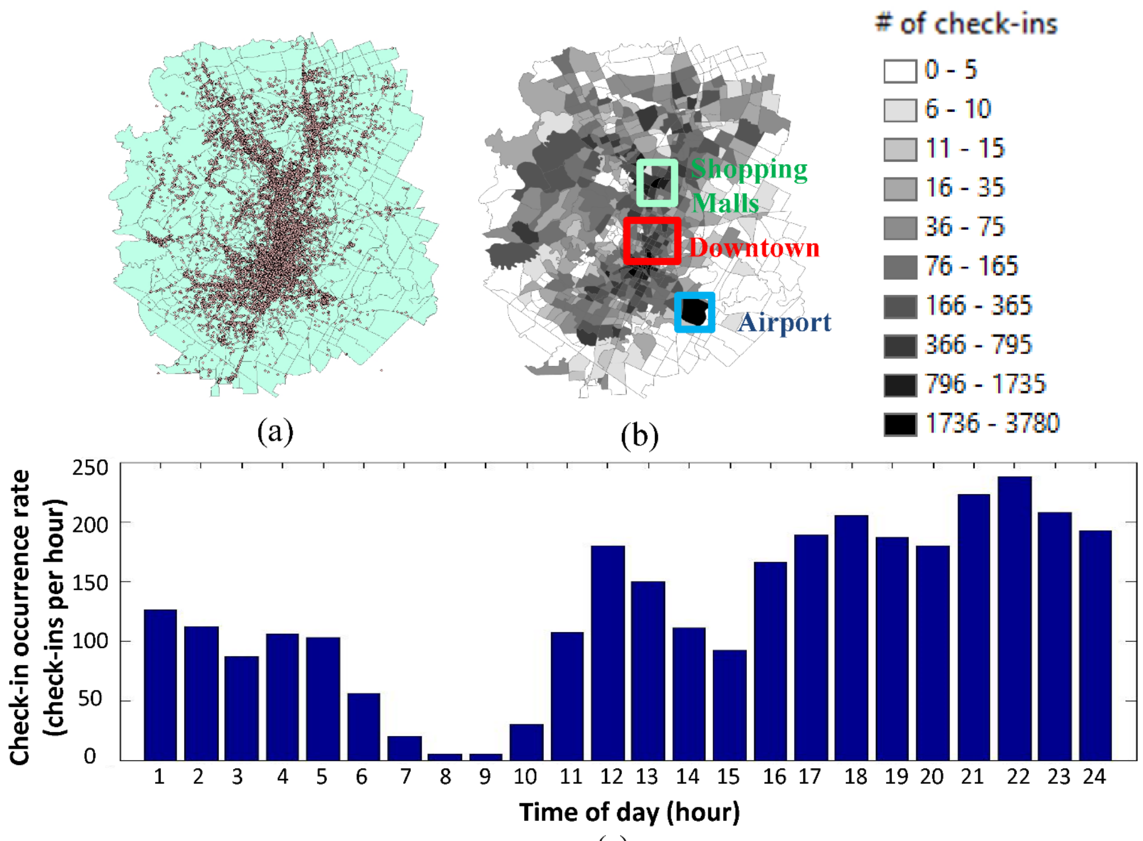
In order to test the proposed algorithms, the City of Austin, Texas and the Manhattan Island of the New York City were selected as the study area. As the LBSN data can provide detailed census-level data, the proposed model was applied to explore the TAZ-aggregated travel demand patterns in two study areas. For the City of Austin, Texas, we obtained the Geographic Information System (GIS) data of TAZ, land use patterns in the Austin area, and the personal daily trip data including OD matrices and TOD factors from Capital Area Metropolitan Planning Organization (CAMPO).

We collected approximate one-year Foursquare check-in data from February 26th, 2010 to January 21st, 2011, which is posted, on Twitter. The GIS data of TAZ boundaries are used to identify the TAZ ID for each Foursquare venue. As illustrated in Figure 2, a total of 730 CAMPO TAZs located in the study area. The CAMPO OD matrix data include the estimated daily trip tables for 17 detailed trip purposes. Six trip purpose categories are used in this study, including home-based work trips (direct/strategic/complex), home-based non-work retail, home-based non-work others, and non-home based work. The TOD factors used by CAMPO to generate TOD OD matrices from daily trip OD are used as reference data for the observation model calibration in the dynamic trip arrival estimation model.

When permitted, the check-ins may include spatial-temporal information demonstrating where and when the check-ins are generated. The individual activity types are categorized by the restarted venue, type of check-in records. Foursquare records users' arrival at their POIs whose location types are given in comprehensive three-level classifications (Li et al., 2013). Therefore, various kinds of trip purposes are identified regarding the POI location types. Since each type has its own spatial and temporal distribution within the urban area, such confirmation allows researchers to compare the impact of different destination type to the model estimation. The URL of a check-in refers to the restarted venue that contains its geographic information. In this study, we focus on the check-ins regarding three particular trip purposes: work trips, retail trips and recreation trips as shown in Table 6.

Table 6: Trip arrival category classification based on Foursquare venue types.

Trip purpose category	Foursquare venue types
Work trips	Professional building, business distinct, and other work-related places
Retail trips	Shopping, grocery activity, and other retail-related places
Recreation trips	Food, entertainment, and other recreation-related places



(a) (b)

(c)

Figure 5: Spatial-temporal Distribution of the LBSN Check-in Arrivals.

Preliminary analysis is conducted on the spatial-temporal characteristics of the check-in patterns in Austin, Texas. Locations of the 124,611 check-ins are collected in the experiment, and are plotted in Figure 5(a). A heat map representing the check-in densities at each TAZ is shown in Figure 5(b). The check-ins are most densely distributed at the airport, downtown and north-central area of the city. To illustrate the temporal characteristics of Foursquare check-ins, we aggregated check-in counts within one downtown TAZ for each hour in a day in Figure 5(c). It is observed that the peaks of check-in counts do not necessarily coincide with regular AM/PM peak hours. This further indicates the need for the observation equations to reduce the sampling bias at different time periods by LBSN data.

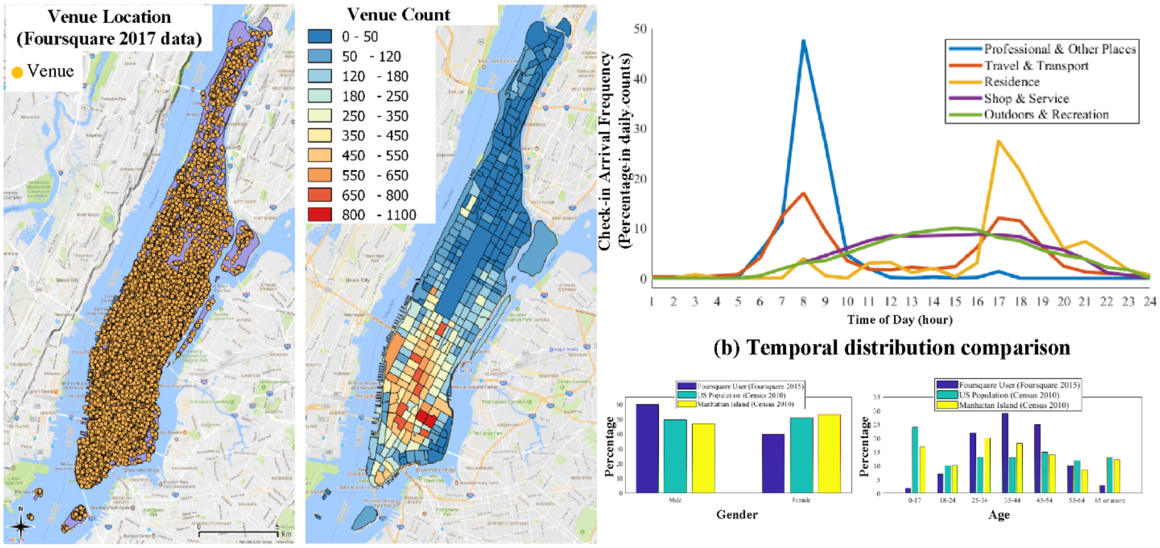
For the Manhattan Island of New York City, The proposed model is evaluated based on LBSN check-in data and NYMTC OD data. LBSN check-in data can be aggregated by different spatial schemes such as census tracts, TAZs, and neighborhood. In this experiment, LBSN is aggregated by TAZ specified in the NYMTC OD matrices. The model input includes the venue-level LBSN check-in data during the weekdays between August 1st, 2016 and March 31st, 2017. Each check-in arrival record contains the venue ID with confirmed location and venue types and the timestamp of the arrival. The venue types are structured in a comprehensive three-level classification system defined by Foursquare with the first level including 10 categories, second level with 100 categories, and the third level consisting of more than 250 categories. To avoid overfitting, we choose the first level that has the number of levels similar to the trip purpose categories used by NYMTC. Excluding the "event" category, a total of 10 venue categories are considered including "Nightlife Spot", "Food", "Shop & Service", "College & University", "Arts & Entertainment", "Travel & Transport", "Professional & Other Places", "Outdoors & Recreation", "Residence", and "Event". Meanwhile, the shape of NYMTC TAZs are used to determine the TAZ ID for each venue. The reference data includes the 2017 weekday OD matrices and TOD factors of personal daily trips from NYMTC. To estimating the dwelling time at the destinations, two datasets, a geo-tagged twitter dataset and the typical travel time profiles from Google Traffic are used. The geo-tagged twitter data provides trip chain information since all geo-tagged activities can be linked together with a Twitter user ID. The typical travel time profiles provide an estimation of the route travel time between ODs. Both datasets are used to initialize the parameters in a dwelling time estimation model.

The LBSN check-in data and the reference OD matrix are separated into two datasets for model calibration and evaluation, respectively. The calibration dataset contains randomly-selected 50 TAZs out of the 318 TAZs in Manhattan. Then the calibrated parameters were applied to evaluate all 318 TAZs in the evaluation. The original 50 TAZs are included to ensure complete visualization and analysis of the full mobility pattern in Manhattan.

Preliminary analysis is first conducted on the spatial and temporal characteristics of the check-in arrival patterns in Manhattan Island. Locations of the 892,970 check-in arrivals were collected in the experiment, and are plotted in Figure 1. A heat map representing the zonal daily trips from both reference data and the LBSN data is shown. The check-in arrivals are most densely distributed in the central park, Midtown, and lower Manhattan area of the city while few observed in the ring area of Central Park and Upper Manhattan area. Meanwhile, hourly check-in counts were aggregated to illustrate the time-of-day variation of check-in arrivals. It is observed that the peaks of check-in counts do not necessarily coincide with regular AM/PM peak hours. The results indicate the need for reducing the spatial and temporal systematic error of the LBSN data.

In the second stage, some initial parameter values are obtained for Equation (2) with a special LBSN dataset. One limitation of the Foursquare data is the lack of dwelling time information at the venues. This can leads to a poor estimation of the activity duration and causing error in estimating the actual time lapse between two consecutive destinations. To address this problem, we used a particular category of the LBSN data from the “Foursquare – Twitter bridge” as analyzed in (Hasan and Ukkusuri, 2014). Through Foursquare-Twitter bridge, users consent to not only letting a check-in counted towards the venue check-in counts but also to posting their arrivals on Twitter with Geo-tags. One such geo-tagged twitter data contains the same details as one Foursquare check-in data but also comes with the Twitter User ID. The temporal difference between two consecutive check-ins of one user is then used to obtain the sum of the activity duration at the first venue location and the travel time between those two venues. The automobile (e.g., Taxi and Uber), transit, and walking travel time information can be extracted from google map API based on the spatial locations of two check-ins. To simplify the model, only the shortest travel time from different modes are used in the dwelling time estimation. We collected around 20,000 geo-tagged tweets to generate OD pairs of different trip purposes and at different time of day. Each check-in arrival records were sorted by the arrival time during the whole day. Meanwhile, users may or may not always check-in at all venue locations. To decide if two check-in arrivals are most likely to be consecutive destinations, we applied the empirical four-hour threshold found in [70] for the OD estimation in the Greater Los Angeles metropolitan area. That is, if a person made two check-in arrivals at different locations within 4 hours, it is considered to be one OD-trip. The latitudes and longitudes of the Geotags are used to generate the travel time between origin and destination locations for one OD-trip through Google Map. Based on equation (2), travel time and activity type are the independent variables and the activity duration is the dependent variable. The initial parameters of the equation (2) are then obtained through multivariate linear regression. Those parameters will be further calibrated in the next step with other model parameters.

Figure 6 depicts some preliminary analysis of the spatial and temporal characteristics of the check-in arrival patterns in Manhattan. The spatial distribution of Foursquare venues studied in the experiment is plotted in Figure 6(a). The venues are most densely distributed in the Midtown and Lower Manhattan area of the city while sparsely in the ring area of Central Park and Upper Manhattan area. A heat map representing the zonal daily arrivals from both reference data is shown in Figure 2b. Meanwhile, we selected five venue categories to show the time-of-day variation of the check-in arrival frequency. In Figure 6(b), it is observed that the peaks of check-in arrivals do coincide with regular AM/PM peak hours of urban travel demand pattern. The results indicate the potential of using check-in arrival to represent the urban travel demand. Figure 6(c) shows the demographic characteristics of foursquare users regarding gender and age. The Foursquare user's information comes



(b) Temporal distribution comparison

(a) Spatial distribution comparison

(c) Demographic Distribution

Figure 6: Spatial-Temporal Distribution and Demographic Distribution of the Obtained LBSN Check-in Data.

from the worldwide statistics of Foursquare users in 2015. It can be observed that the demographic distribution of Foursquare users generally matches that of the general population. Some inconsistencies can be found among population groups under 17 years old and greater than 55 years old. These population groups may have less number of trips made and miles traveled than the other population groups.

As one critical role in the proposed time-delay correlation model, activity duration estimation is to determine the potential time-of-day correlated OD pairs. We collected around 20k check-in arrival records, which posted in Tweet, to generate individual trajectory data. Such data then was applied to extract the OD pairs within the study area. Each check-in arrival records were sorted by the arrival time during the whole day. In essence, if a person made two check-in arrivals at different locations within 4 hours, it is considered to be one OD-trip. The latitude and longitude information is used to generate the travel time between origin and destination locations for one OD-trip through Google Map.

4.3 Model Calibration

The model calibration was conducted in two stages. First, the proposed Hawkes process based trip arrival estimation model was calibrated against the reference trip arrivals data and TOD factors from the reference dataset. Secondly, the proposed time-delay-correlation gravity model was compared with reference OD data from. In each model calibration stage, the genetic algorithm is used to obtain the optimal model parameters, and the objective function is to minimize the MAE (Mean Absolute Error) between the modeled and reference data. The genetic algorithm (GA) is a search heuristic that mimics the process of natural selection (Deb, Pratap et al. 2002). By mimicking the evolution process in nature, GA applies “mutation” and “crossover” to the initial population and select the best “offspring” in each generation. Such repetition stops when a relative optimal that satisfies the termination condition is reached.

The experiment consists of two components of the dynamic trip arrival estimation and dynamic OD estimation. As the calibrated trip arrivals will be fed into the trip distribution modeling to generate the OD matrices, it expects that an accurate trip arrival estimation leads to an accurate OD estimation. Therefore, we focus on testing the algorithm of dynamic trip arrival estimation only in the City of Austin, Texas and testing the algorithm of dynamic OD estimation in Manhattan Island of New York City.

For the City of Austin, Texas, the trip arrivals from CAMPO OD matrices are aggregated into matrices for work trips (home and non-home based), retail trips (home-based) and recreation trips (home-based and non-home-based) respectively. To guide the GA optimization, a specially designed objective function is used to 1) avoid giving too much weight to TAZs with high check-in frequencies, and 2) efficiently account for TAZs with low check-in frequencies without numerical overflows. By filtering the

TAZs with null check-in occurrence, we use the Angle of Difference (AOD) value as the fitness value of our models. It is an indicator of the similarity between the calibrated results and CAMPO data, which is defined as follows:

$$AOD = \frac{\sum_i^I abs(\pi/4 - atan2(\bar{A}_i, A_i))}{I} \tag{44}$$

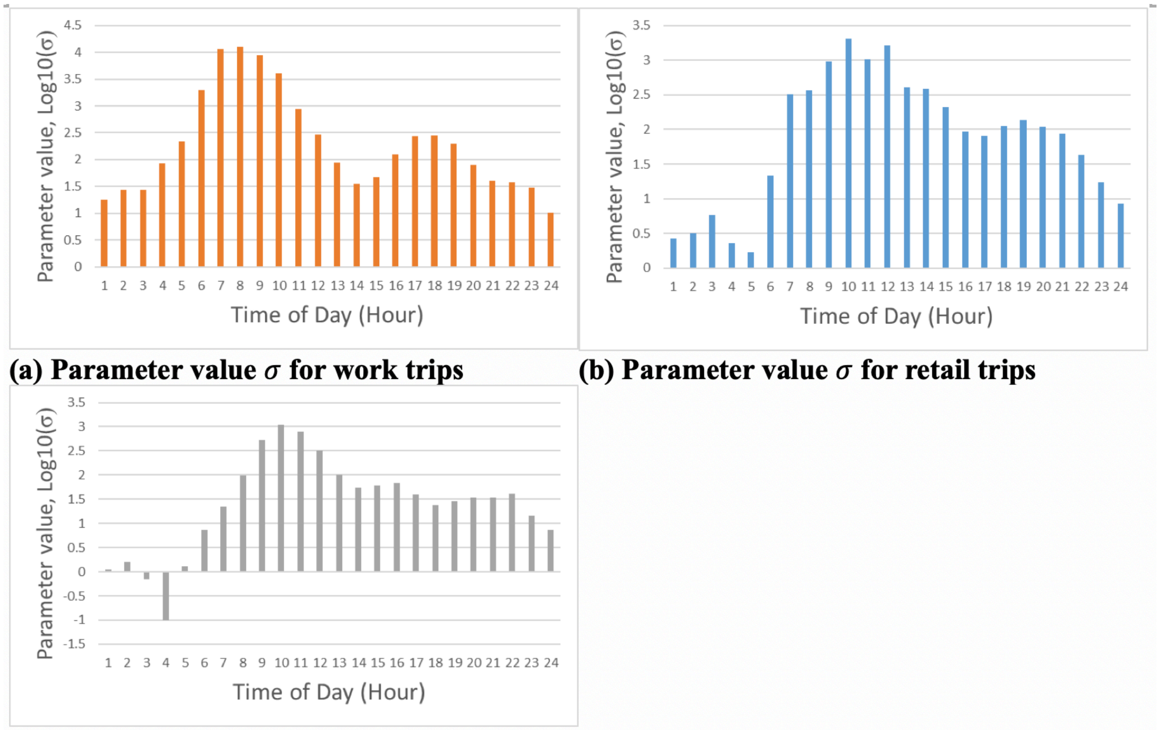
where $\bar{A}_i$ and A_i are the calibrated trip arrival estimation from the proposed model and the CAMPO trip arrivals for location i respectively, $atan2(x, y)$ is the four-quadrant inverse tangent function of a point (x, y) which will return zero degree if $\bar{A}_i = 0$, and I is the number of Traffic Analysis Zones. AOD is a straightforward way of describing the deviations. It gives the number of degrees by which the average data points deviated from the $y = x$ line. When AOD is close to zero, the $(\bar{A}_i, A_i)$ pair is close to the $y = x$ line, thus the model output is similar to the agency OD.

Table 7 shows the description of parameters used in the baseline statistics model and proposed Hawkes process model. Figure 7, Figure 8, and Figure 9 provide the model calibration results from the genetic optimization for all work trips, retail trips, and recreation trips models. The proposed Hawkes process models outperform the baseline statistics model with fewer parameters: 8 in the proposed Hawkes process model versus 24 in the baseline statistics model.

Table 7: Notation Description.

ParametersDescription	
α	A scaling factor for self-excitation (α¿0) or self-decaying (α¡0)
β	A positive parameter describing the dependence of recent arrival events on future events
γ, σ	The converting factor of Foursquare check-ins to trip arrivals
δ	A positive parameter for the selected percentage of trip arrivals estimated through the LBSN check-ins observation

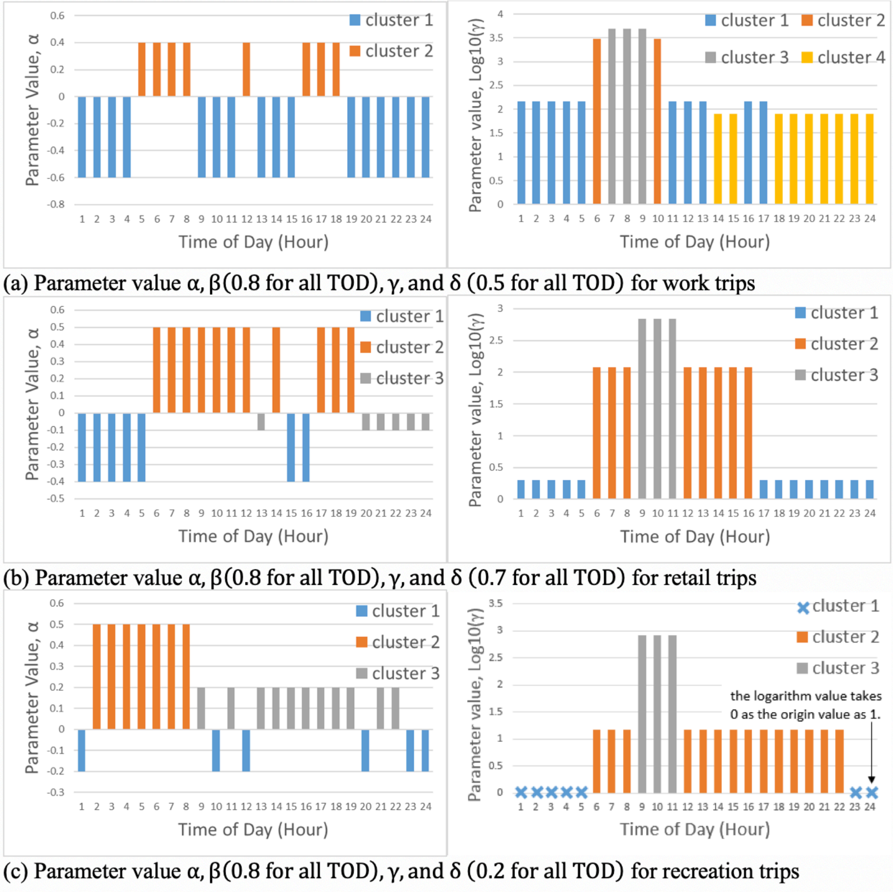
For Manhattan Island of New York City, in each model calibration stage, the genetic algorithm is used to obtain the optimal model parameters, and the objective function



(a) Parameter value σ for work trips **(b) Parameter value σ for retail trips**

(c) Parameter value σ for recreation trips

Figure 7: Model Calibration Results for the Baseline Statistics Model. (The logarithm values of parameter σ with respect to base 10 are used due to the high range of the parameter value).

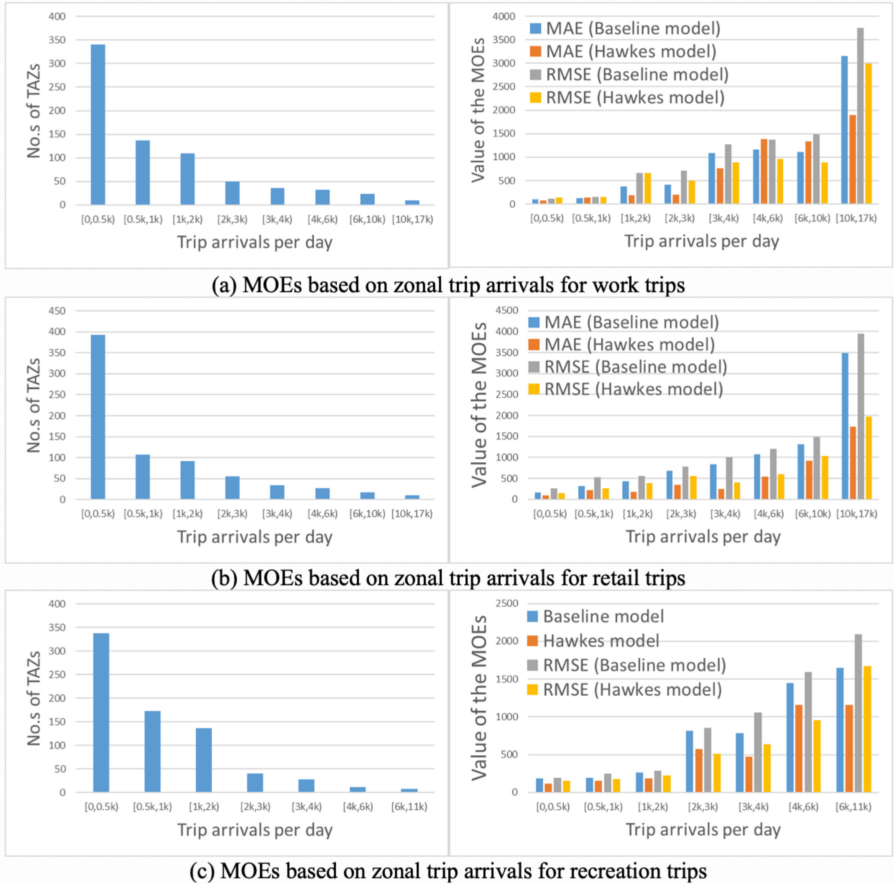


(a) Parameter value α, β(0.8 for all TOD), γ, and δ (0.5 for all TOD) for work trips

(b) Parameter value α, β(0.8 for all TOD), γ, and δ (0.7 for all TOD) for retail trips

(c) Parameter value α, β(0.8 for all TOD), γ, and δ (0.2 for all TOD) for recreation trips

Figure 8: Model Calibration Results for the Proposed Hawkes Process Model . (The logarithm values of parameter γ with respect to base 10 are used due to the high range of the parameter value)



(a) MOEs based on zonal trip arrivals for work trips

(b) MOEs based on zonal trip arrivals for retail trips

(c) MOEs based on zonal trip arrivals for recreation trips

Figure 9: MOEs for the Baseline Statistic Models and the Proposed Hawkes Process Models.

is to minimize the MAE (Mean Absolute Error) between the modeled and reference data. Table 8 listed the model calibration results. It is found that when the length of selected sequence equals to 14, the model performs better according to the overall MAE and CR values. The reason may be related to the mechanism of Pearson product-moment correlation analysis which is the comparison between two sequences. Under 1 min resolution of the data feed, the longer a sequence selected, the larger noise may be brought into the correlation analysis.

The estimation result of activity duration using multivariate regression model was shown in Table 9. When $p - value > 0.05$, there is no strong evidence to indicate that the variable is statistically significant. From the test results, variables for activity types *Event* and $Travel\&Transport$ contained coefficients found to be insignificant based on its P-value. Such variables may not be considered in the activity duration estimation model. Meanwhile, the positive sign of coefficient for variables $Traveltimetoandfromactivity$ indicates that people may stay longer at one location where he/she spent a long time traveling to. It responded to the hypothesis proposed by [92]. Regarding the magnitude of coefficient value, we found that activity type $Arts\&Entertainment$ and $Professional\&OtherPlaces$ contains the longest activity duration and activity type $College\&University$ and *Residence* contains the shortest activity duration. After filtering out the insignificant activity type indicator of linear regression equation (10) for activity duration estimation, eight different activity types are considered to estimate the activity duration.

Table 8: Model Calibration Results for the Dynamic OD Estimation Model.

Notation description	Parameter	Value
Hawkes equations $(4 \sim 5)$	α	-0.5 (hour of day: $1 \sim 5, 20 \sim 24$); 0.3 (hour of day: $6 \sim 10, 12, 15 \sim 19$); 0.1 (hour of day: $11, 13 \sim 14$)
	β	4 (hour of day: $1 \sim 3, 15, 23 \sim 24$); 7.6 (hour of day: $4 \sim 5, 10 \sim 14, 16 \sim 22$); 34 (hour of day: $6 \sim 9$)
	γ	0.8
	δ	0.5
Selected sequence length (min) in time-delay-correlation equations $(6 \sim 8, 11)$	w	14
The constant of linear regression equation (10) for activity duration estimation	a	5.22
Coefficients for activity type indicator of linear regression equation (10) for activity duration estimation	b(AT)	78.54 (Food); 58.89 (Professional & Other Places); 104.47 (Arts & Entertainment); 82.26 (Nightlife Spot); 60.06 (Shop & Service); 90.12 (Outdoors & Recreation); 64.22 (Residence); 68.89 (College & University)
Coefficients for travel time of linear regression equation (10) for activity duration estimation	c	0.84
Threshold coefficient of indicator function $(14 \sim 15)$ for determinant of the potential correlated OD pairs	$r_{threshold}$	0.71
MOE for trip arrival estimation	MAE	23.1
MOE for time-of-day OD matrices comparison	MAE	5.2
MOE for trip length distribution comparison	CR	0.86

Table 9: Regression analysis result of activity duration.

Variable	Coefficient value	P-value
Constant	6.87	#N/A
Travel time to and from activity (min)	0.68	0.00
Activity type indicator (if 1, other 0)		
Food	78.54	0.00
Travel & Transport	73.77	> 0.05
Professional & Other Places	158.89	0.00
Arts & Entertainment	104.47	0.00
Nightlife Spot	82.26	0.00
Shop & Service	80.06	0.00
Outdoors & Recreation	90.12	0.00
Residence	64.22	0.00
College & University	68.89	0.00
Event	93.22	> 0.05

Note: Dependent variable is the activity duration (in minutes). Number of observations = 24812 and the system $R^2 = 0.650$.

5 Experiment Results

5.1 Evaluation Result for the Dynamic Trip Arrival Estimation of the City of Austin, Texas

5.1.1 Temporal Trip Arrival Patterns Comparison

For the dynamic trip arrival estimation of City of Austin, Texas, we review the predicted spatial-temporal pattern. First, we compared the calibrated models and CAMPO trip arrivals matrix by examining the temporal distribution of estimated trip arrivals based on the proposed model. The predicted trip arrivals patterns within each hour are used to plot the temporal distribution of estimated trip arrivals. Figure 10 shows the calibrated results from the genetic algorithm for all three Hawkes process models.

As shown in Figure 5.1, the predicted TOD variations from the proposed Hawkes pro-

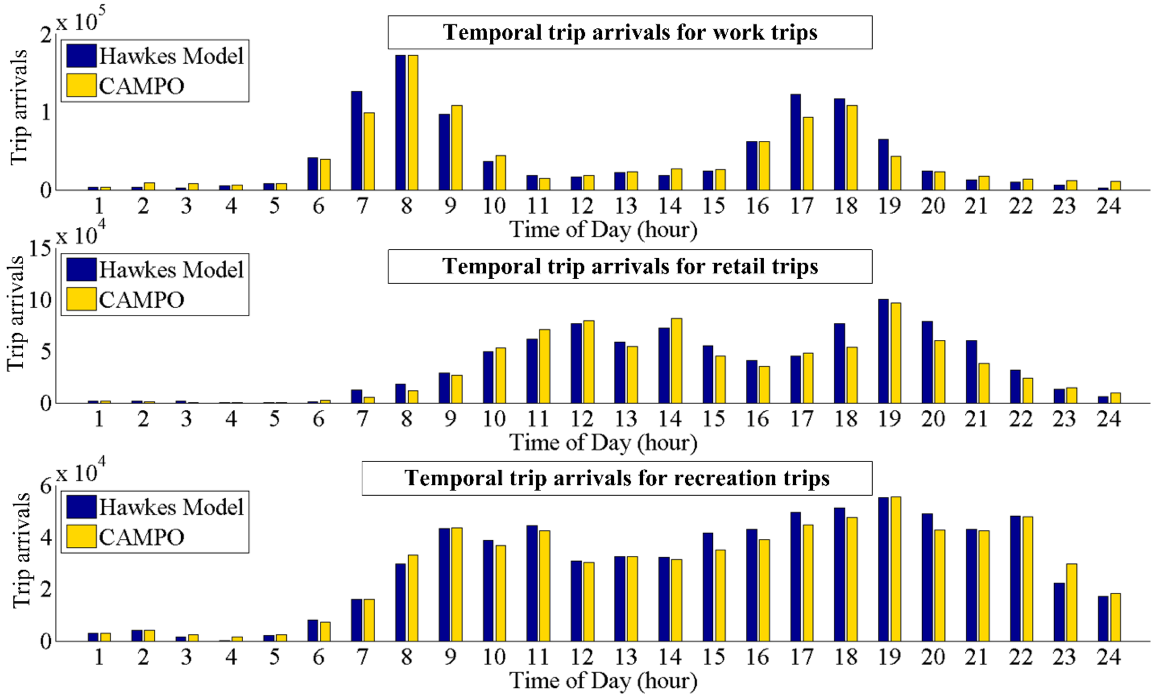


Figure 10: The Time-of-day Distribution of the Estimated Trip Arrivals.

cess models are similar to those reflected in the CAMPO data. The predicted patterns from the proposed model match well with the time-of-day variations indicated in the empirical time-of-day factors from CAMPO with only a slight inconsistency during the early hours of AM and PM peak (work trips), the midday (retail trips), and the evening (recreation trips). Each model reproduces the trip arrivals trend of particular trip type. For work trips, there are two distinct peaks during the AM/PM periods and relatively few trips during the midday and nighttime. For retail trips, the first peak happens around noon when people from work places conduct shopping trips during their lunch breaks. Furthermore, the retail trips reach another peak in the early evening when people shop for groceries and other items. Finally, the recreation trips have an average distribution from late morning to midnight.

The statistical test also was applied over the course of the day as shown in Table 10. The paired t-test is conducted for comparing one-on-one temporal trip arrivals for each hour of the day. The hourly trip arrivals during one day of the total TAZs for each selected trip purpose has been aggregated into 24 observations. The hypothesis is that there is zero estimation error for the hourly trip arrivals within the overall study area for each selected trip types between the model results and the observed trip arrivals from the CAMPO data. The t-value and P-value from the tests demonstrate that we do not reject the null hypothesis.

Table 10: Statistical Test Results for Hourly Trip Arrival Estimation.

t-Test: Paired Two Sample for Means	Work Trips		Retail Trips		Recreation Trips	
Indicator	CAMPO	Hawkes	CAMPO	Hawkes	CAMPO	Hawkes
Mean	4.2E+04	4.1E+04	3.7E+04	3.6E+04	2.9E+04	2.9E+04
Variance	2.4E+09	1.9E+09	9.9E+08	9.1E+08	3.4E+08	3.1E+08
Observations	24		24		24	
Pearson Correlation	0.98		0.97		0.99	
Hypothesized Mean Difference	0		0		0	
Df	23		23		23	
t Stat	0.51		1.14		0.74	
$P(T <= t)$ one-tail	0.31		0.13		0.23	
t Critical one-tail	1.71		1.71		1.71	
$P(T <= t)$ two-tail	0.62		0.27		0.47	
t Critical two-tail	2.07		2.07		2.07	

5.1.2 Spatial Trip Arrivals Pattern Comparison

The model evaluation of spatial trip arrival estimation is shown in Figure 11. The angle between the trend line of the result and line “y=x” can illustrate how well the model output matches the CAMPO data. The close scattering of points to the y=x line represents the accurate estimation of zonal trip arrivals. The horizontal axis represents the design trip arrivals for each zone that used the sorted Zone ID as the value, while the vertical axis is the calibrated result for the corresponding zones. Due to the high variations of the zonal trip arrival estimates among 738 TAZs, we selected the X-Y plot of the modeled result versus CAMPO data to indicate estimation accuracy. The coordinates of each dot in the diagram display the calibrated trip arrivals for zone i, which is defined as follows:

$$(x, y) = (i, i * (\bar{A}_i/A_i) \tag{45}$$

Furthermore, in the color scale for the density of dots distribution used, the darker color indicates the higher density. The comparison points of the proposed Hawkes models distribute more closely to the y=x line which is consistent with lower AOD value. More distant points can be found from the baseline model plots than the Hawkes process model plots. Additionally, the proposed model has better coverage for all three types of trip estimation. The baseline plots display greater error of null estimation as there are more dots along the line “y=0”. It is found that those points with large deviations are usually residential areas where few check-in activities occur. This demonstrates the need for bias reduction across different types of venues. A regression analysis has been conducted in each model result. An R2 indicator was used to indicate the Pearson correlation between the reference data and the modeled result. The regression line and the equation have shown that the proposed model outperforms the baseline model for three selected trip purpose.

Figures 12 describe trip arrival estimation within the involved TAZs regarding the City of Austin’s land use classification. We have identified five area types: CBD, urban intense, urban residential, suburban residential, and rural area. Figure 13, 14, 15 compare the geospatial patterns of zonal trip arrivals with grayscale showing the estimated trip arrivals within each TAZ. The color scale highlights where the model

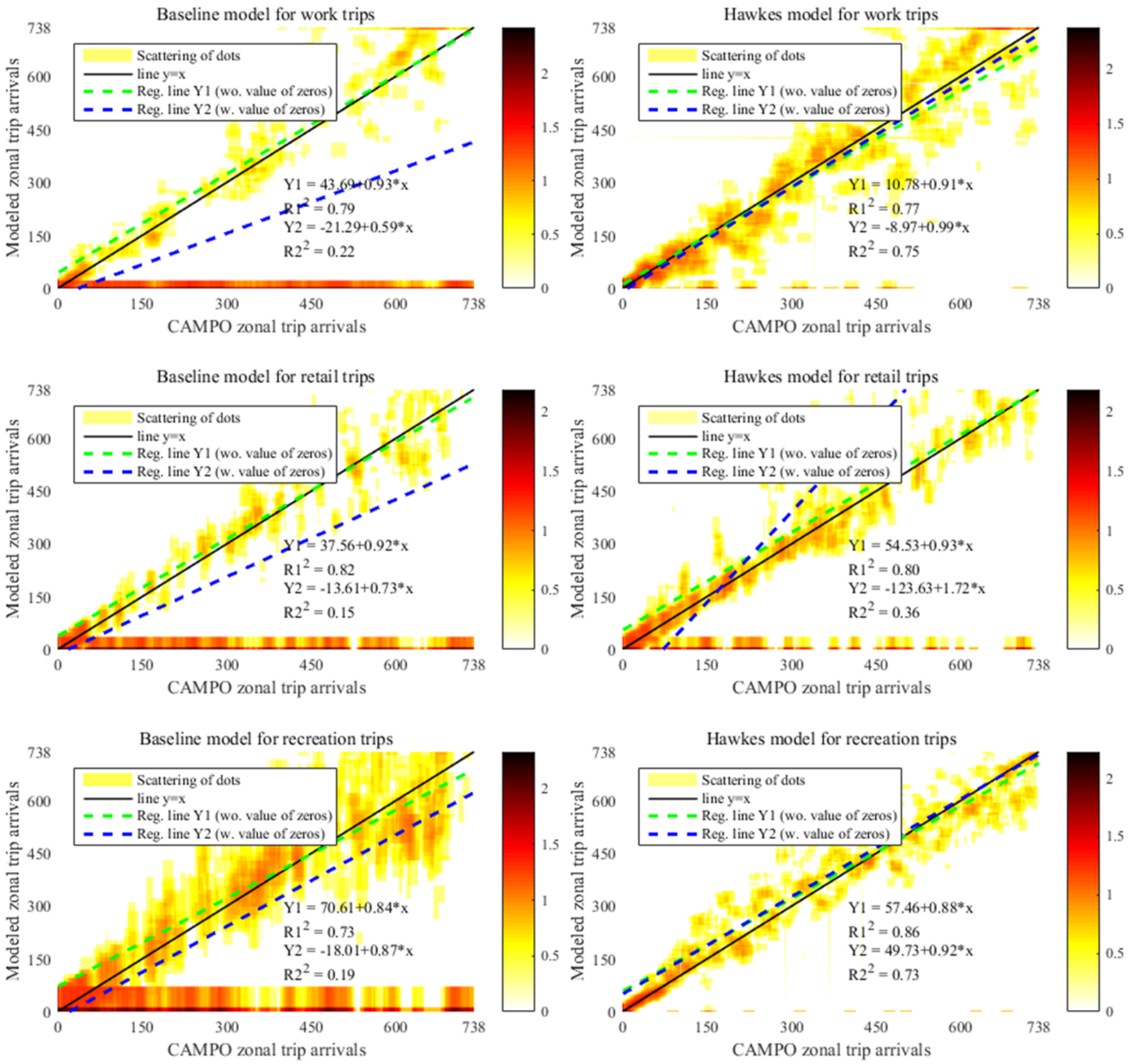
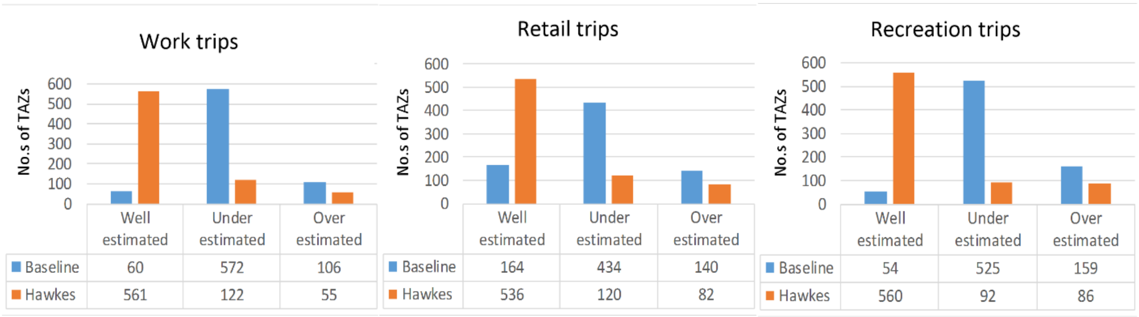


Figure 11: Zonal Trip Arrivals Pattern Comparison.



Work trips	Well estimated	Under estimated	Over estimated
■ Baseline	60	572	106
■ Hawkes	561	122	55

Retail trips	Well estimated	Under estimated	Over estimated
■ Baseline	164	434	140
■ Hawkes	536	120	82

Recreation trips	Well estimated	Under estimated	Over estimated
■ Baseline	54	525	159
■ Hawkes	560	92	86

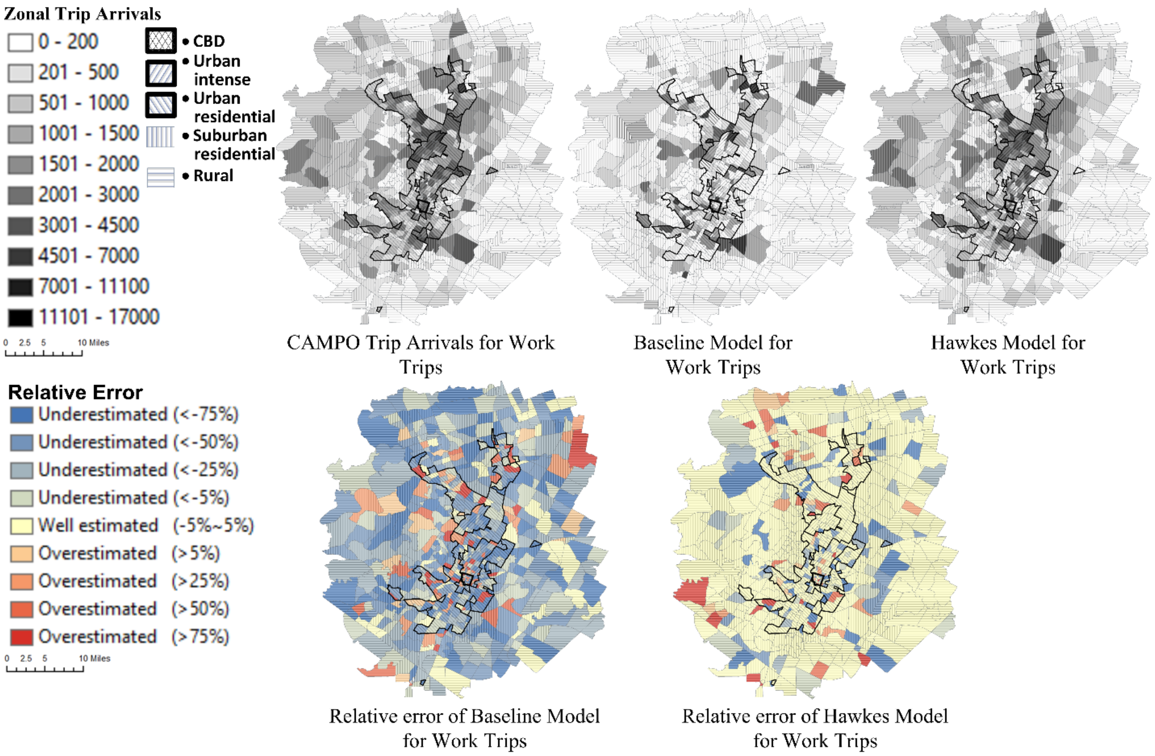
Figure 12: Comparison of Estimation Results for Work, Retail, and Recreation Trips.

overestimates or underestimates the number of trip arrivals. Five land use types are indicated by five different line fill symbols. The estimated trip arrivals from the proposed model share the more consistent spatial pattern reflected in the CAMPO data. Figure 12 indicates that the overall performance of the proposed Hawkes process model is better than the one of the baseline statistics model. Furthermore, Table 11 indicates that we do not reject the null hypothesis for zonal trip estimation comparison between estimated result and observed data.

In Figure 13, the underestimation of work trips can be found for the baseline model in the suburban area indicating compensation needed for residential trips. The spatial pattern of work trips around the downtown area estimated by the Hawkes model shows a close resemblance to that of the CAMPO data.

For the retail trips pattern indicated in Figure 14, both models experience high observations at many industrial locations, which shows the limitation of Foursquare data. So there is a need for a more detailed bias reduction process towards urban and suburban employment centers though the Hawkes process model does provide larger coverage than the baseline statistics model.

As shown in Figure 15, the proposed model also outperforms the baseline model in the estimation of recreation trip arrivals regarding the similarity of the spatial patterns in the CAMPO data.



CAMPO Trip Arrivals for Work Trips

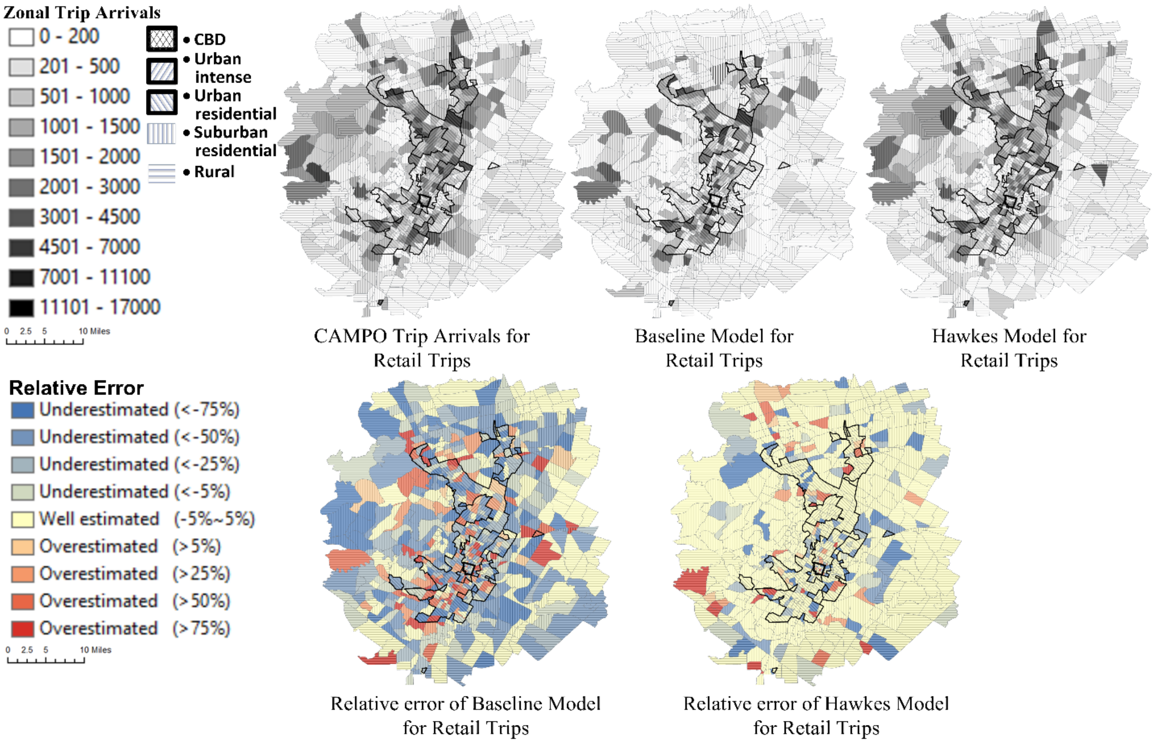
Baseline Model for Work Trips

Hawkes Model for Work Trips

Relative error of Baseline Model for Work Trips

Relative error of Hawkes Model for Work Trips

Figure 13: Daily Zonal Trip Arrivals Heat Maps for Work Trips.



CAMPO Trip Arrivals for Retail Trips

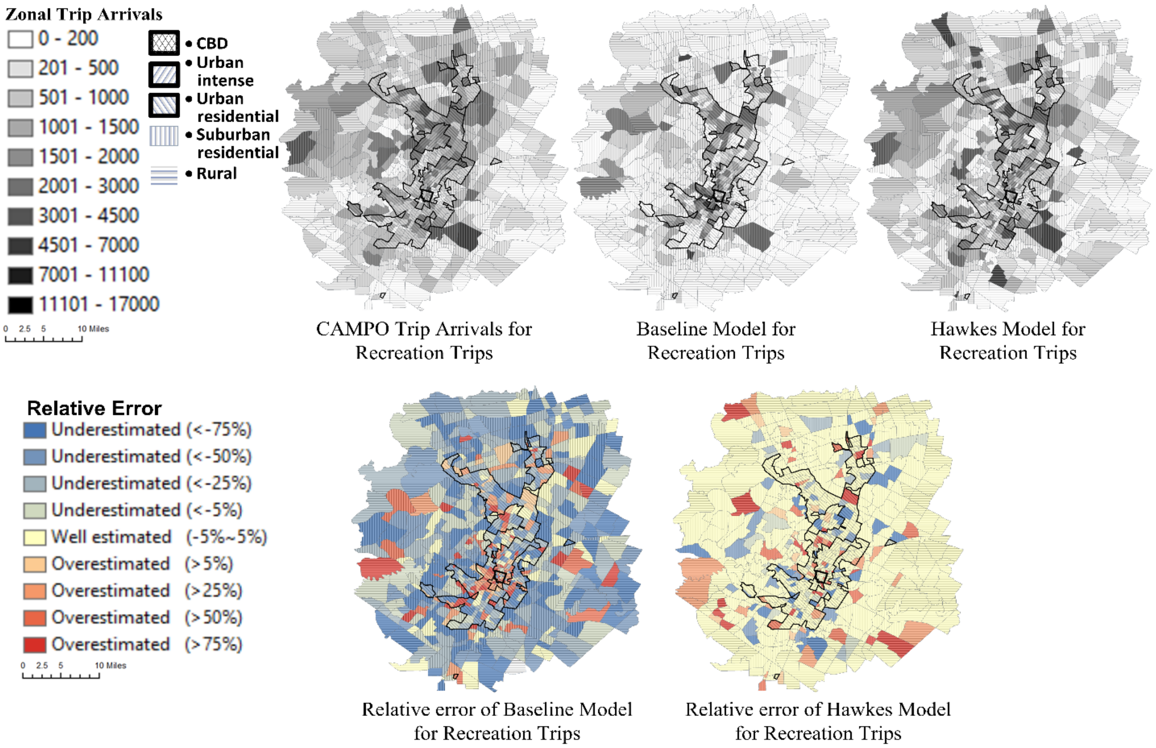
Baseline Model for Retail Trips

Hawkes Model for Retail Trips

Relative error of Baseline Model for Retail Trips

Relative error of Hawkes Model for Retail Trips

Figure 14: Daily Zonal Trip Arrivals Heat Maps for Retail Trips.



CAMPO Trip Arrivals for Recreation Trips

Baseline Model for Recreation Trips

Hawkes Model for Recreation Trips

Relative error of Baseline Model for Recreation Trips

Relative error of Hawkes Model for Recreation Trips

Figure 15: Daily Zonal Trip Arrivals Heat Maps for Recreatiion Trips.

Similar to Table 10, the paired t-test had also been conducted to compare the difference between paired means of the spatial distribution. The statistical test applied to the study area was conducted using the paired t-test for comparing one-on-one spatial trip arrivals within each TAZ. The daily trip arrivals of each TAZ for each selected trip purpose has been aggregated into 738 observations. The hypothesis is that there is zero estimation error for the daily zonal trip arrivals between the model results and the observed trip arrivals from the OD data for each selected trip type. Similar to the regression analysis on Figure 11, the zeros values in both the reference and estimated data have been removed to satisfy the assumption of the paired t-test. As shown in Table 10, the t-value and P-value from the tests demonstrate that we do not reject the null hypothesis.

Table 11: Statistical Test Results for Zonal Trip Arrival Estimation.

t-Test: Paired Two Sample for Means	Work Trips		Retail Trips		Recreation Trips	
Indicator	CAMPO	Hawkes	CAMPO	Hawkes	CAMPO	Hawkes
Mean	1252.05	825.60	1629.26	1444.47	970.52	1226.55
Variance	3.9E+06	3.6E+06	5.9E+06	3.6E+06	1.4E+06	5.1E+06
Observations	643	643	461	461	698	698
Pearson Correlation	0.67		0.62		0.55	
Hypothesized Mean Difference	0		0		0	
Df	642		460		697	
t Stat	6.80		2.03		-3.34	
$P(T <= t)$ one-tail	1.2E-11		0.02		4.4E-04	
t Critical one-tail	1.65		1.65		1.65	
$P(T <= t)$ two-tail	2.3E-11		0.04		8.8E-04	
t Critical two-tail	1.96		1.97		1.96	

5.1.3 Dynamic Trip Arrival Patterns Generation

To further illustrate the potentials of the proposed model in generating dynamic trip patterns, the temporal travel arrival patterns estimated by the proposed model were analyzed in Figure 16, 17, 18, 19. The pattern fits well with the expected daily activities in the Austin area. We used a bar diagram and color maps to describe both temporal and spatial characteristics of the trip arrivals inferred by LBSN data.

Figure 16 shows the different trends regarding three trip types. First, for all three trip types, trip intensity during the nighttime is less than that during the daytime. The activity level reaches a minimum between 0:00 to 4:00. Second, work trips are captured during the AM peak and PM peak for commuting activities. For retail trips, most of the daily trips began in the late morning. Finally, while the first activity peak

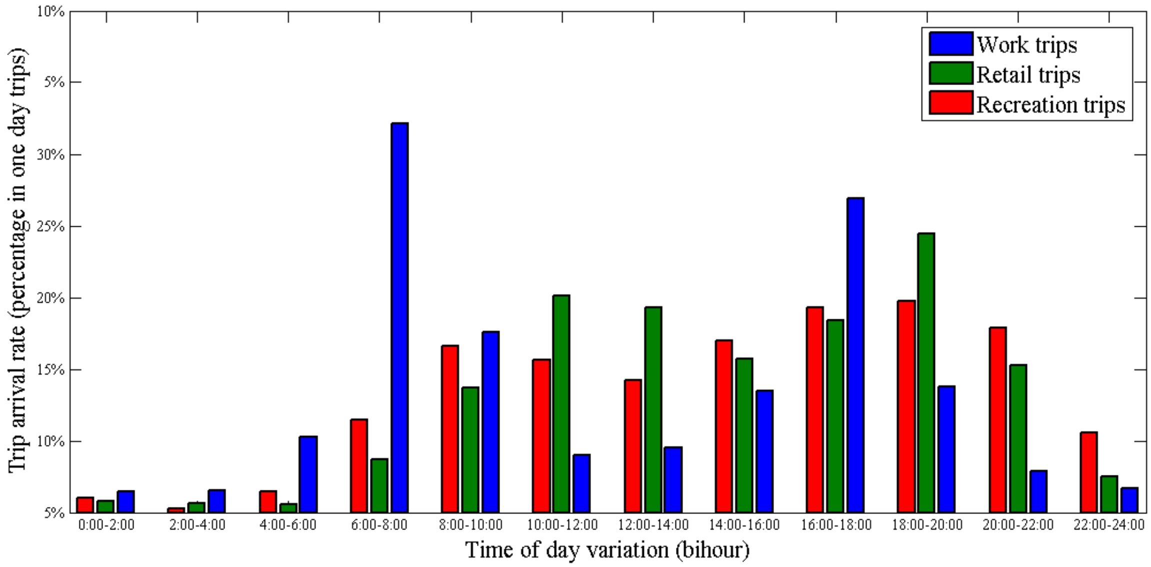


Figure 16: The Bi-hourly Trip Arrival Pattern.

of recreation trips can be found around 10:00 to 14:00, a noon activity peak is also observed, which is consistent with lunch times. Another activity peak is found around 18:00 to 20:00, when most dining, shopping, and entertainment activities may often occur.

Figure 17 shows the temporal zonal trip arrivals for work trips. We combined those time windows that share the similar spatial variation in the study area. For example, among the three different time intervals of 22:00-24:00, 0:00-2:00, 2:00-4:00, few work trips were observed due to the regular pattern of commuting activities. The mean value of zonal trip arrivals among the above time windows was generated to represent the zonal characters. For work trips, we found the AM peak and PM peak for commuting activities in the downtown areas and industrial locations in the west.

Consistent with the temporal pattern observed from the bar chart in Figure 18, for retail trips, most daily trips began in the late morning in the downtown area and specific shopping mall locations in the north and west. On the other hand, few trip arrivals were observed in the east, where most of the city's wetlands and agricultural areas are found. Moreover, the variation of the retail trips' demand fits the regular

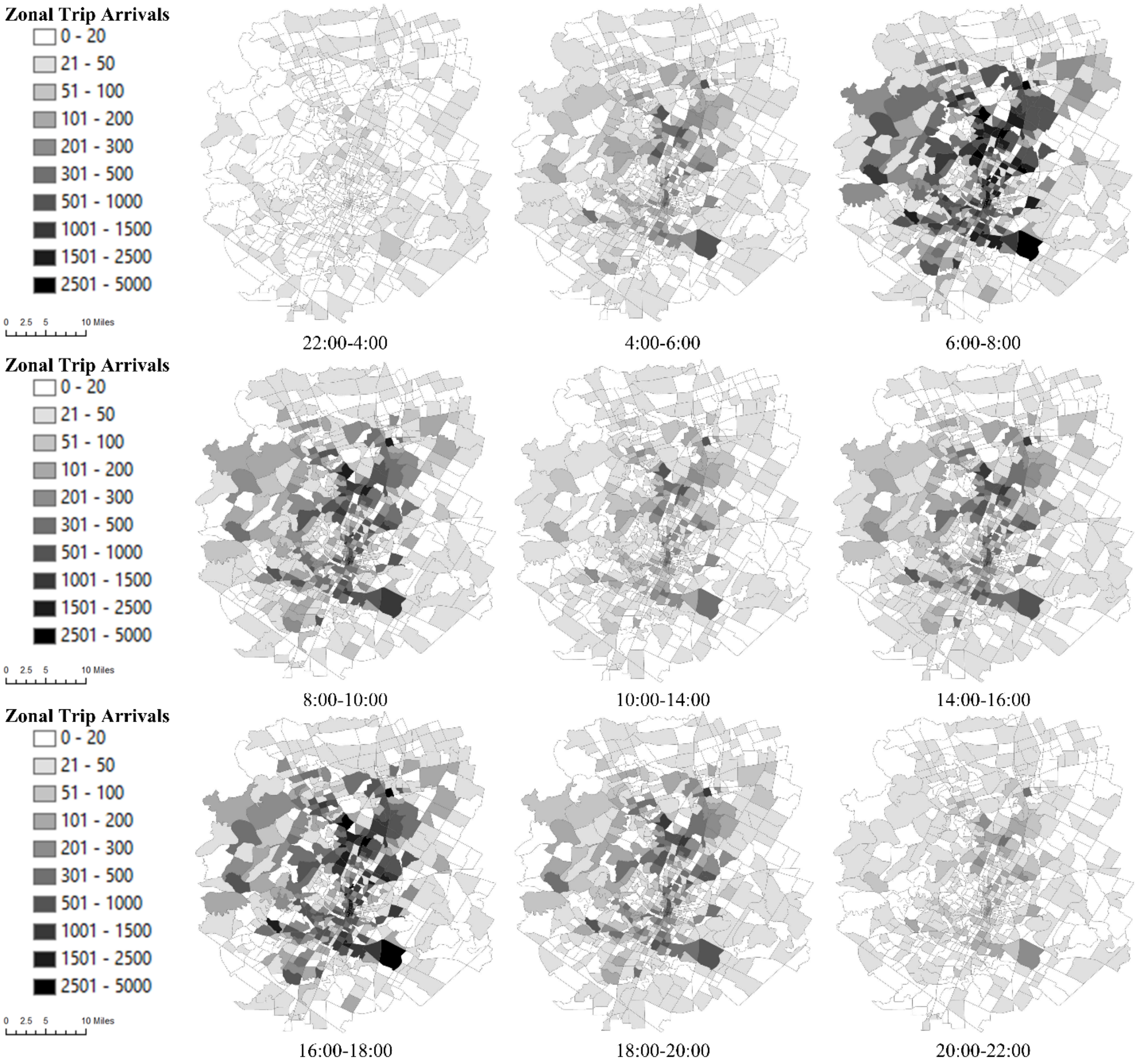


Figure 17: Temporal Zonal Trip Arrivals Heat Map for Work Trips.

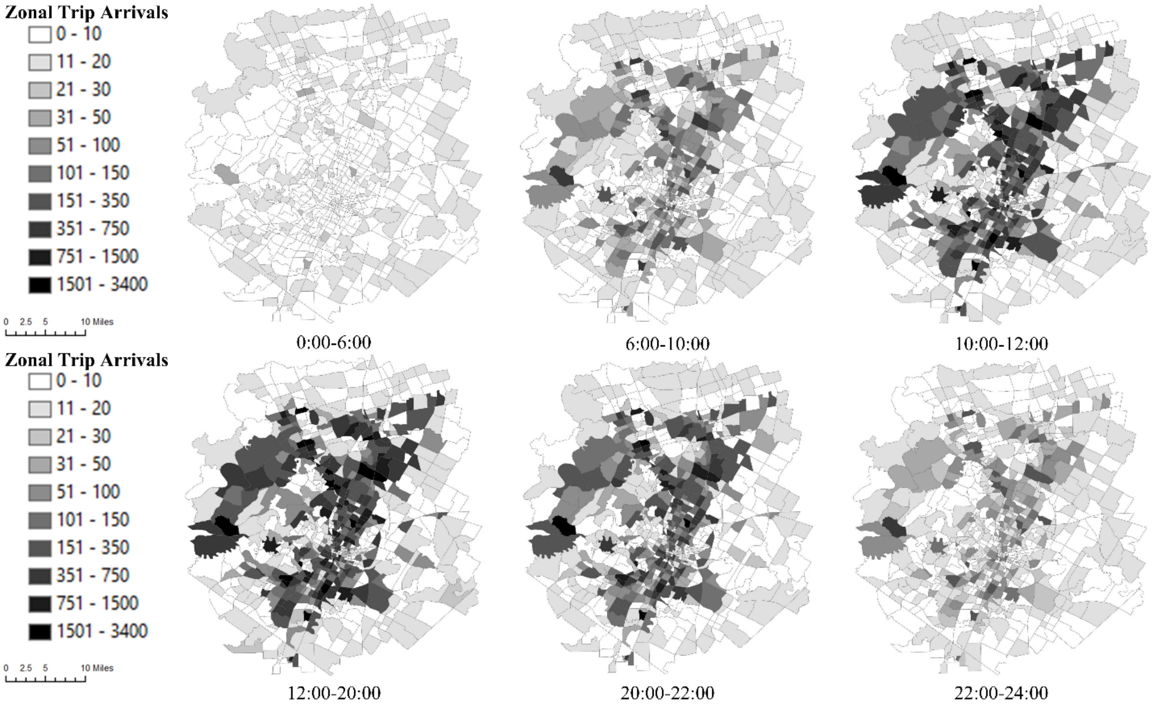


Figure 18: Temporal Zonal Trip Arrivals Heat Map for Retail Trips.

opening hours of the retail services in the urban area.

For recreation trips, Foursquare data exhibits good coverage for the nighttime activities in the City of Austin. In Figure 19, two dining activity peaks are observed around noon and in the evening. The time periods coincide with the lunch and dinner rush times when people leave their morning and afternoon activities' locations for dining or entertainment places. Meanwhile, some residual travel can also be observed between 22:00 to 24:00, indicating the times people are returning home from their late night activities.

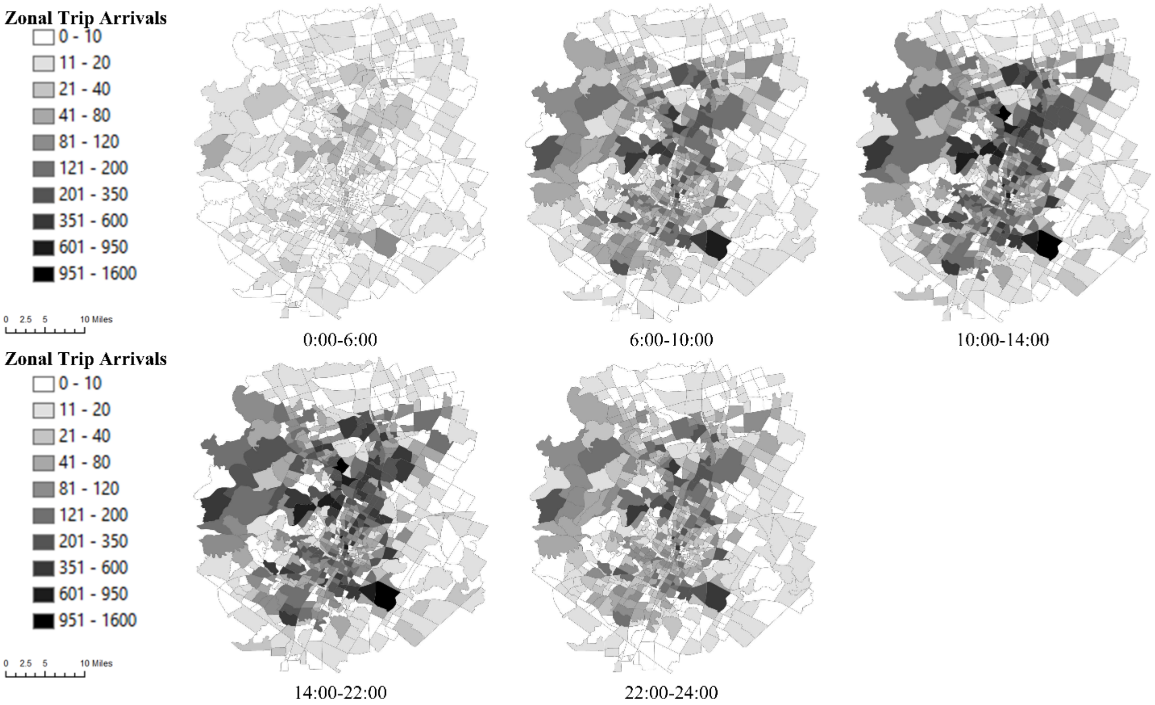


Figure 19: Temporal Zonal Trip Arrivals Heat Map for Recreation Trips.

5.2 Evaluation Result for the Dynamic OD Estimation of Manhattan Island of NYC

5.2.1 Zonal Functionality Profiling

To uncover the zone functions, 5 latent zone types based on the POIs and taxi datasets were generated as shown in Table 1. It can be found that the POIs such as "Shop & Service" and "Food" had a relatively high rank within different topics compared to other POI categories. This reflects the fact that most trips reported by POI check-ins are discretionary trips such as social/recreational activities. Meanwhile, the human mobility pattern reflected by the taxi pickup and dropoff data contributed to generate different topic. Five land use types guided by the New York City zoning and land use data [50] were selected as zonal type labels: "Commercial-Retail", "Commercial-Work", "Residence", "Transportation Hub", and "Open Space". We mapped out in Fig.5a the distribution of zonal types. The zonal types discovered indeed resemble the

Table 12: Zonal Topic Classification.

Topic 1	Prob.	**Topic 2**	Prob.	**Topic 3**	Prob.	**Topic 4**	Prob.	**Topic 5**	Prob.
S	0.315	S	0.201	S	0.114	S	0.222	F	0.218
Sa9(D)	0.187	Mo09(D)	0.180	R	0.113	F	0.216	S	0.211
N	0.121	Th12(D)	0.160	Fr21(D)	0.113	Tu08(P)	0.122	N	0.201
F	0.098	Tu09(D)	0.156	Mo08(P)	0.112	T	0.121	A	0.195
Sa13(D)	0.080	F	0.134	Mo09(P)	0.112	Tu08(D)	0.084	O	0.178
We08(D)	0.077	N	0.104	Tu11(D)	0.109	Sa11(P)	0.044	Sa16(D)	0.017
We20(P)	0.067	P	0.038	F	0.108	A	0.019	C	0.017
Commercial-Retail		Transportation Hub		Residence		Open Space		Commercial-Work	

S-Shop & Service; N-Nightlife Spot; T-Travel & Transport; F-Food; R-Residence; A-Art & Entertainment; P- Professional & Other Building; O- Outdoor & Recreation; C- College & University; DDHH(D) – day-of-week time-of-day taxi drop-offs; DDHH(A) – Day-of-week time-of-day taxi pickups.

functional diversity of Manhattan's census tracts: commercial-work area for "Financial District", open space area "Central Park", and Transportation Hub area "Penn Station".

5.2.2 Dynamic Travel Demand Estimation

Regarding the trip arrival patterns, the predicted trip arrivals from HPSS formulation were aggregated hourly to generate the trip arrival patterns in Figure 20 blue bar indicates the ground truth temporal distribution of trip arrival over the study area; yellow bar indicates the predicted one. The calibrated results show that the predicted trip arrival patterns from the proposed model match well with the ground truth time-of-day trip arrival under the aggregation of 318 total zones. There are two distinct peaks during the AM/PM periods and relatively few trips during the midday and nighttime. Meanwhile, an average distribution can be found during the lunch break. Furthermore, given the high variations of trips among 318 total zones, we plot the modeled result versus the ground truth data to visualize estimation accuracy. Each dot represents the reference value as x coordinate and the predicted value as y coordinate. The regression line has a slope of 0.66 and the $R^2 = 0.80$ under statistically significant level $P = 0.00$.

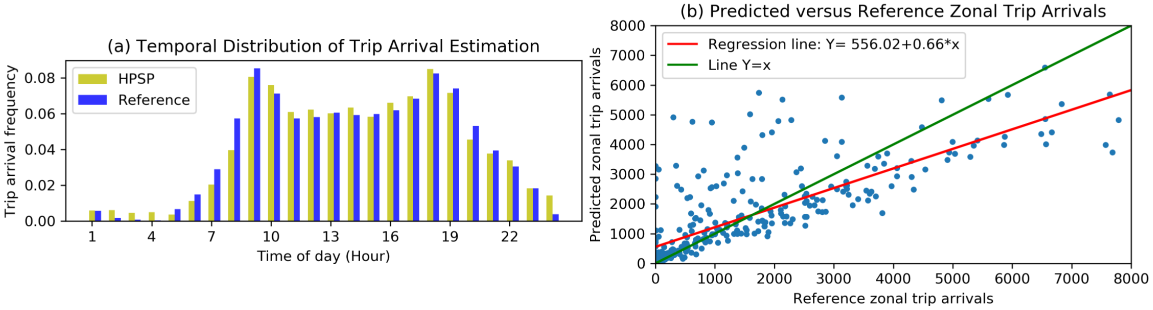


Figure 20: Model Performance of Trip Arrival Estimation.
(a) blue bar indicates the reference temporal distribution of trip arrival frequency over the study area; yellow bar indicates the the predicted one. (b) Regression analysis for the zonal trip arrival prediction over the entire day. each dot represents the reference value as x coordinate and the predicted value as y coordinate.

5.2.3 Correlation Composition Analysis

For the dynamic OD estimation in Manhattan Island of New York City, one example of the correlation analysis between check-in arrivals among TAZs is firstly illustrated in Figure 21. The 60-min sequence of check-in arrivals collected from one commercial area containing "Time Square" was selected. Its 14-min correlated sequence (shown as the red curve) was used as one comparison sequence applied correlation analysis. Three scenarios of compared sequences of check-in arrivals at origin locations collected from one transportation area containing "Penn Station", one residential area, and one open space area containing "Central Park"(shown as the blue curve) was sorted by the time delay value. The first scenario represents a strong positive correlation with 0.96 coefficient value and 24-min time delay. The sequence of check-in arrivals at the destination location was to replicate the sequence at the origin location. It may be expected that the AM commuting trips arrived in the commercial area contain one stop located in the transportation hub in the early period. The second scenario shows a strong negative correlation with -0.79 coefficient value and 25-min time delay. The check-in arrivals' reduction at the origin location contributed to one increased at the destination location. People were leaving their home for the day time activities. It generated a decreasing trip intensity locally and an increasing outflow to the destina-

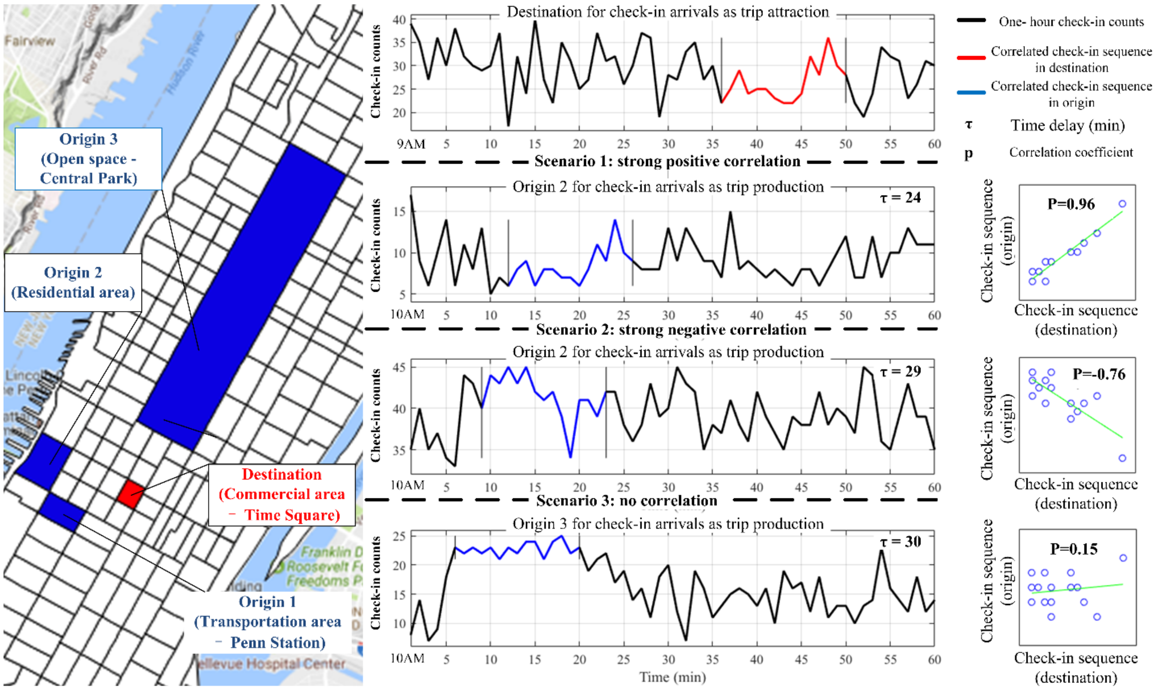


Figure 21: Correlation Composition Profile.

tion. Finally, the correlation of the check-in arrival sequence between the selected OD pairs was lower than the calibrated threshold value in the third scenario. Therefore, no correlation was considered between two locations.

5.2.4 OD Flow Patterns Comparison

Two indicators of the model performance were applied for OD estimation including OD flow patterns and the regression analysis. First, the TAZ-level daily OD flow patterns are evaluated. Figure 22 compares the OD flow patterns between the modeled OD matrix and the NYMTC OD matrix. Each grid (i, j) in the diagram displays the adjusted OD flow intensity I_ij from TAZ i to TAZ j defined as the following.

$$I_{ij} = log_{10}\frac{\bar{T}_{ij}}{\sum_i \sum_j \bar{T}_{ij}} \tag{46}$$

where I_ij was colored based on the value. Dark color represents high OD flow, and light color suggests low OD flow. Meanwhile, the relative error was applied to be the measure of effectiveness to evaluate the model's fitness. The color scale highlights where the model overestimates (relative error ¿0) or underestimates (relative error ¡0) the number of OD trips. It can be found that the model exhibits higher estimation error within the higher 200 TAZs which are the ring area surrounding area "Central Park". The functional diversity in Manhattan Island consequences particular travel demand patterns of areas such as business district for "Financial District", recreation area "Central Park", and residential area in "Upper East & West Side". While those higher 200 TAZs share the similarity of relatively large size and various functional characteristics (e.g. resident, school, work, recreation). It suggests the model may need to consider the land use based travel demand modeling for the mixed functionality within one TAZ. In general, the above comparisons indicate significant similarity between the OD matrix generated from the model and the NYMTC OD matrix. Furthermore, one regression analysis was conducted under the daily OD trips. Due to the high variations of trips among the total of $318 * 318 * 96$ OD pairs, we selected the X-Y plot of the modeled result versus the reference data to indicate estimation accuracy. A color scale for the density of dots distribution was used. The darker color indicates the higher density. An R2 indicator was used to indicate the Pearson correlation between the reference data and the modeled result. The regression line and the equation have shown that the proposed model has a promising potential of high-resolution dynamic OD estimation.

Furthermore, the daily OD flow patterns generated by the proposed model against four baseline models were evaluated as shown in Figure 23. As the performance of five OD estimation models is presented under four different constraints, a total 20 model-constraint combinations were explored. In Figure 4a c, the MAE and NRMSE metrics indicate the zonal trip count differences between the ground truth and predicted OD matrices, while CR measures the similarity of trip length distribution

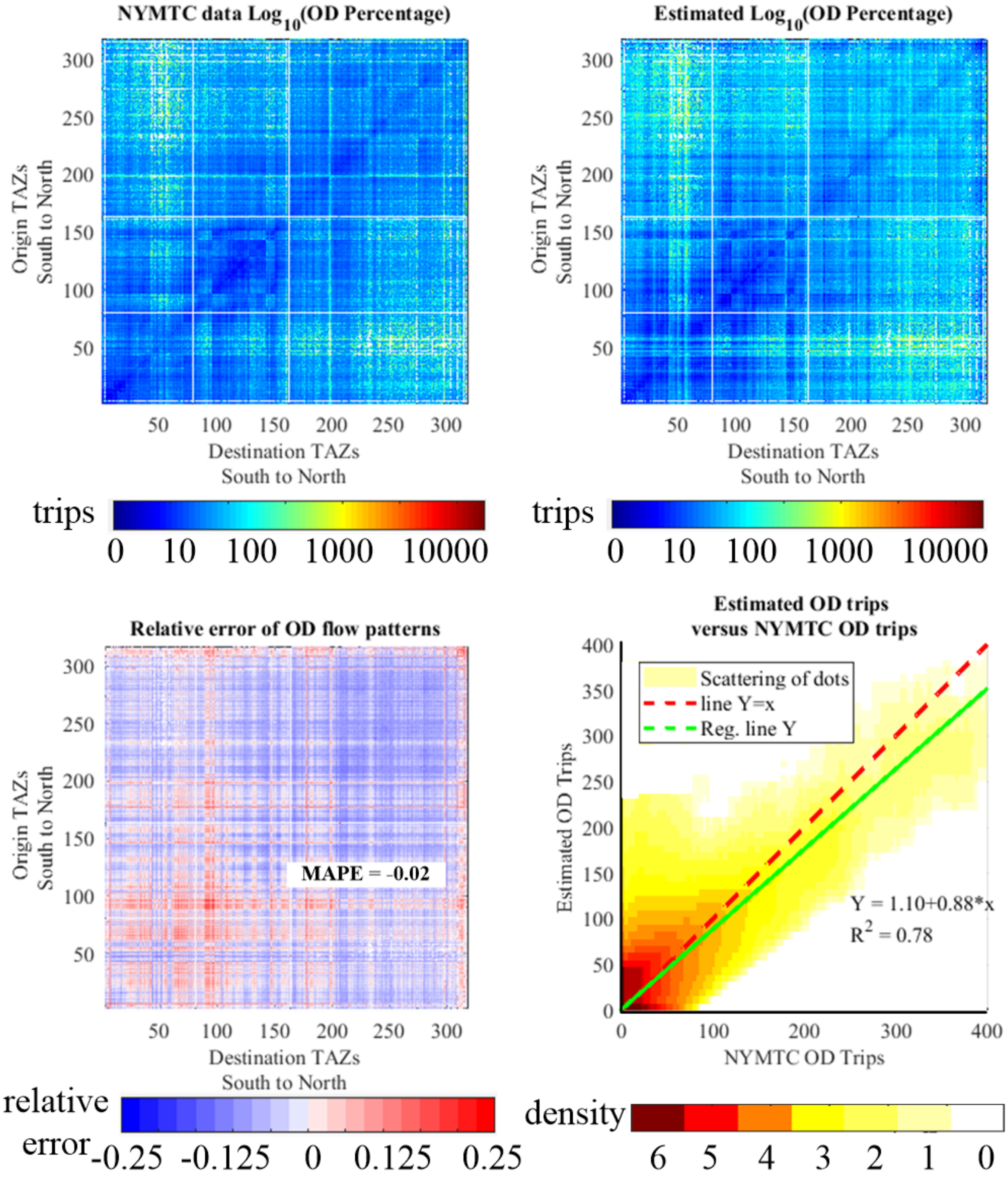


Figure 22: OD Flow Patterns Comparison.

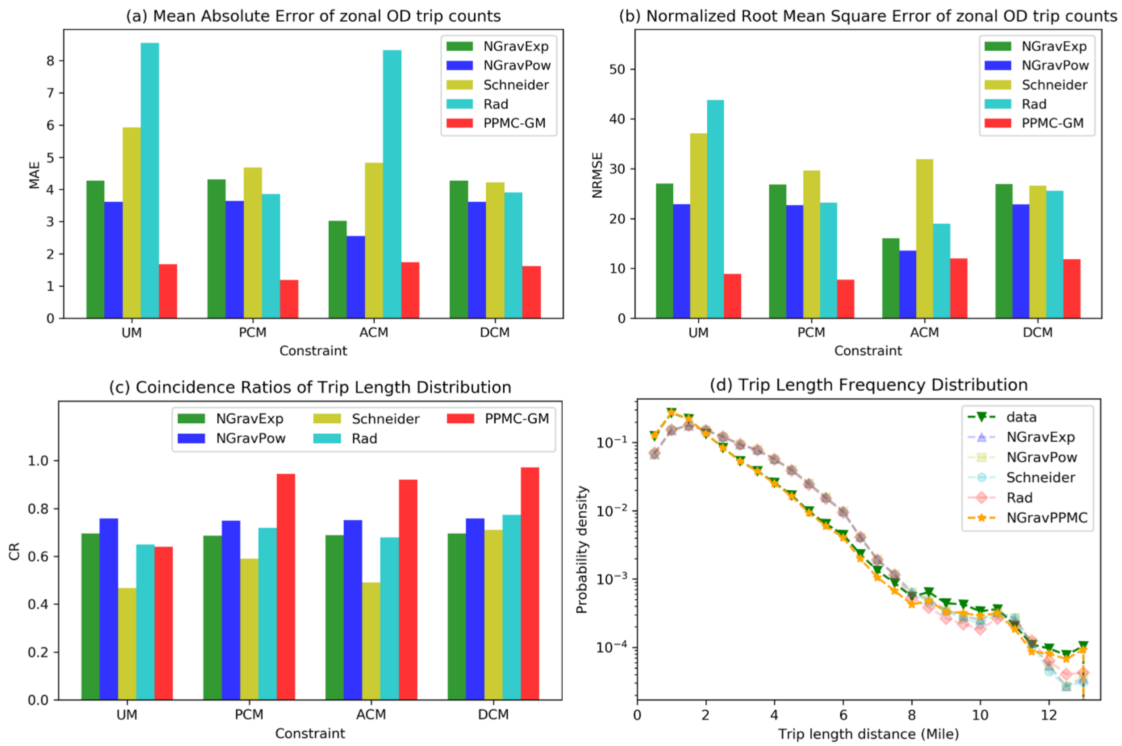


Figure 23: Performance of OD estimation.

curve between the ground truth and predicted OD flow patterns. Since the constraint models contain a sampling step from the multinomial distribution, we consider the average metrics over 100 runs of the OD estimation. We observed that the OD estimation model with a singly-constrained model (ACM/PCM) is better for estimating the zonal OD trip counts, while the doubly-constrained model (DCM) better predicts trip length distribution. Globally, the result obtained with the proposed PPMC-GM model achieves the lowest MAE/NRMSE value and the highest CR value. We also report the average trip length distribution curve in Figure 4d. As the different performance indicator gives the different best combination of the OD estimation model and constraint model, we refer to the best constrained OD estimation model when mentioning the model in trip length distribution curve. The result shows that the proposed model outperforms the baseline models in term of the consistency with the ground truth trip length distribution.

The trip length distribution curves are the same as those defined for calculating the coincidence ratio, whose curves can illustrate how well the model output matches the ground truth data. Along with the predicted daily trips, the predicted trips within each hour are used to plot the temporal distribution of estimated OD matrix. Figure 24, 25 demonstrates the comparison results between the survey and predicted trips. Relatively consistent matching can be observed, although, the modeled data slightly underestimates the number of trips in the early morning. The traffic diary of taxi service performs less recurrent flow and more stochastic OD pairs during that period. Figure 24, 25 illustrates the cumulative trip length distributions. It can be observed that the trips predicted by the proposed time delay model accumulate faster for shorter trips than the ground truth trips. In general, the two curves follow the same paths and data points are located within proximity to one another in both plots, demonstrating the feasibility of the proposed method.

5.2.5 Land Use based Time-of-day OD Flow Patterns Generation

Since agency travel demand data only have uniform sets of time-of-day factors for each trip purpose, it cannot provide the time-of-day pattern for each land-use type. The proposed model has the capability of evaluating the time-of-day patterns among OD pairs of different land-use types. In the first application, the predicted time-of-day zonal travel demand patterns are analyzed between OD pairs of several different land use types. There are a total of 8 different non-vacant lane-use types in Manhattan, NYC (Zola) as shown in Figure 26. Four land-use types are analyzed including the residential area (R), the transportation hub area (TH), the open space area (OS), and the commercial area. The commercial area type is further divided into commercial-retail area (CR) and commercial-workplace area (CW) types.

Figure 26 reveals the dynamic OD patterns among 18 different land-use type combinations originated from the six selected zones representing five different zonal land-use

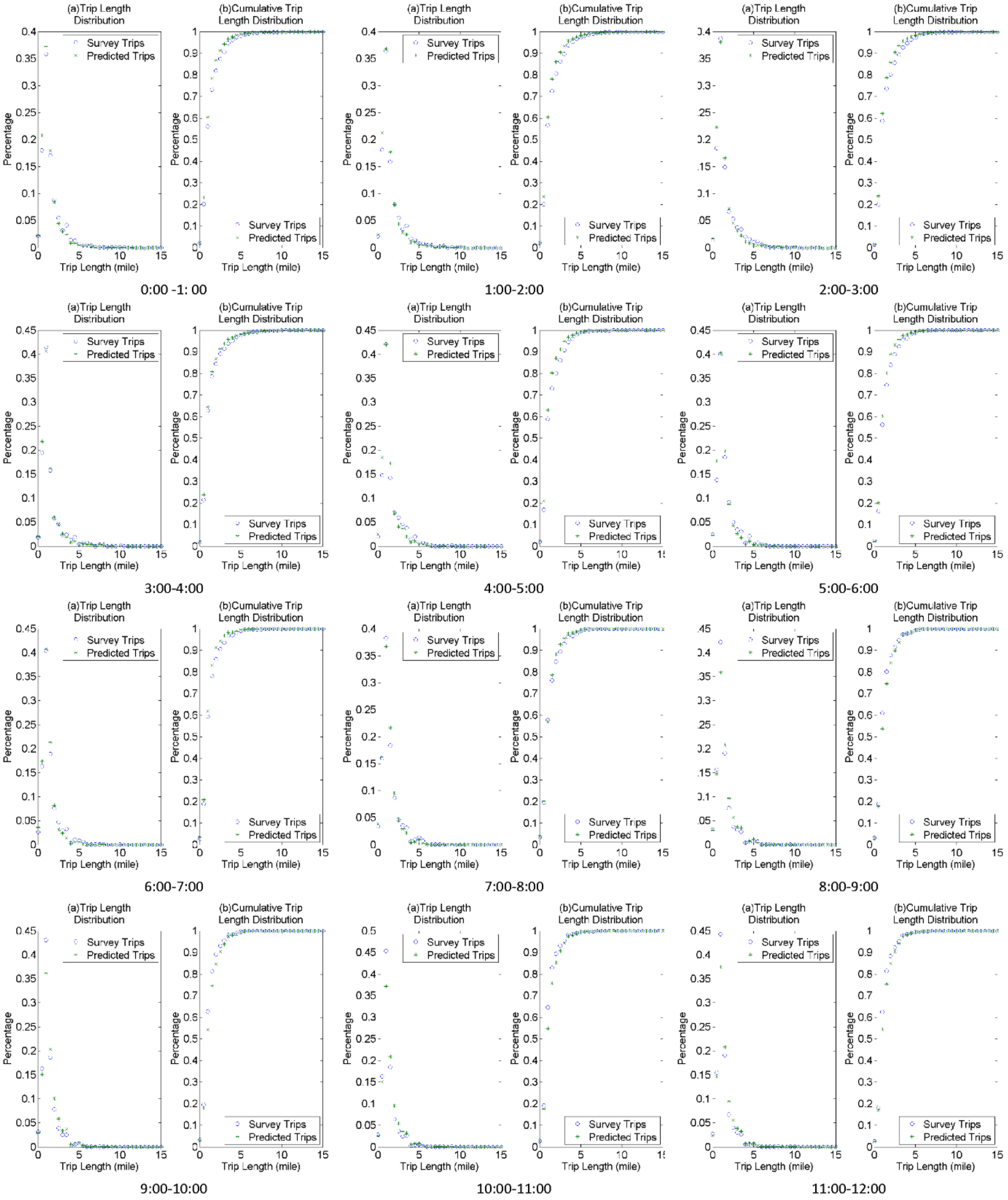
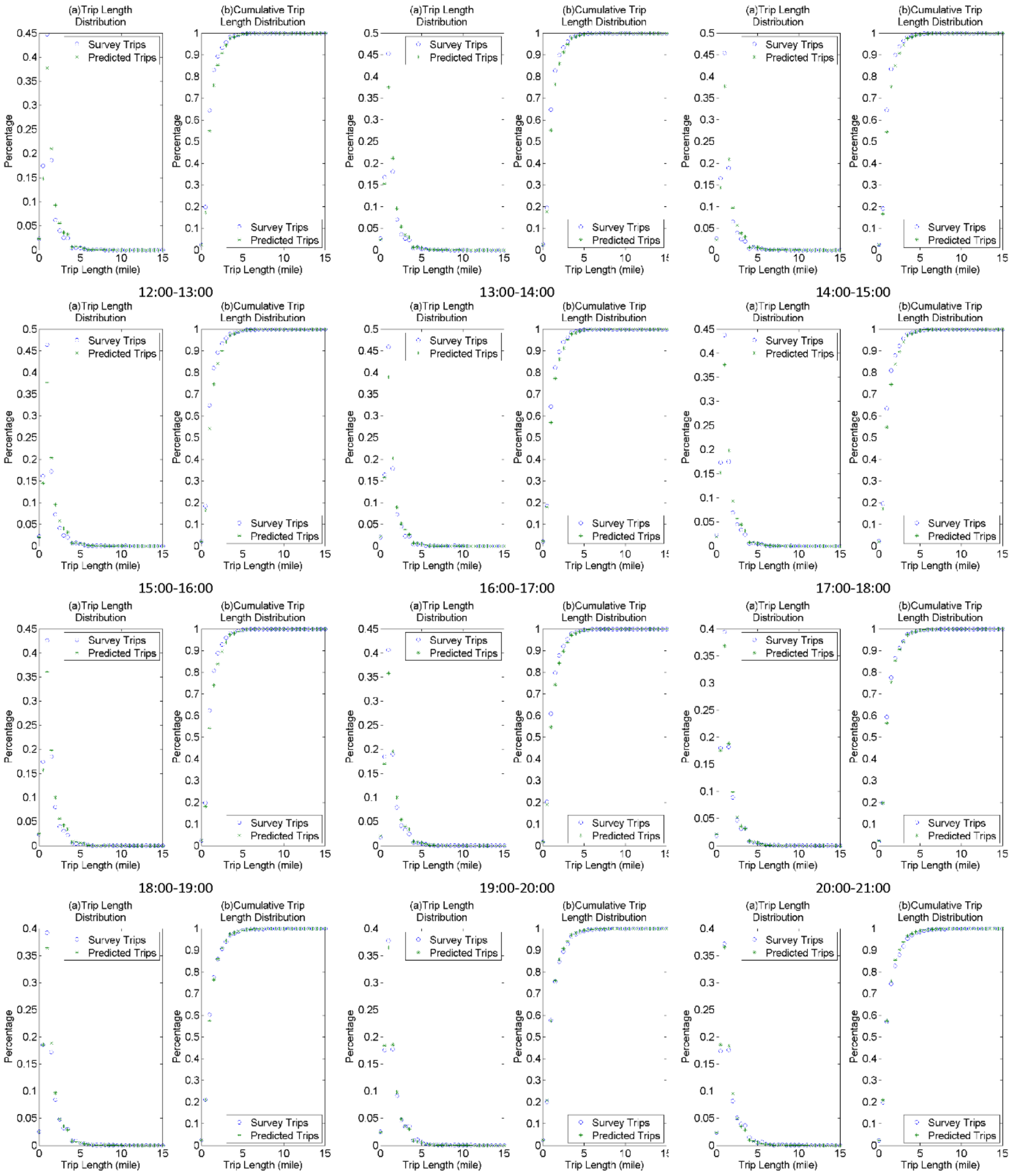


Figure 24: Trip length distributions.
$(0:00 \sim 12:00)$



12:00-13:00

13:00-14:00

14:00-15:00

15:00-16:00

16:00-17:00

17:00-18:00

18:00-19:00

19:00-20:00

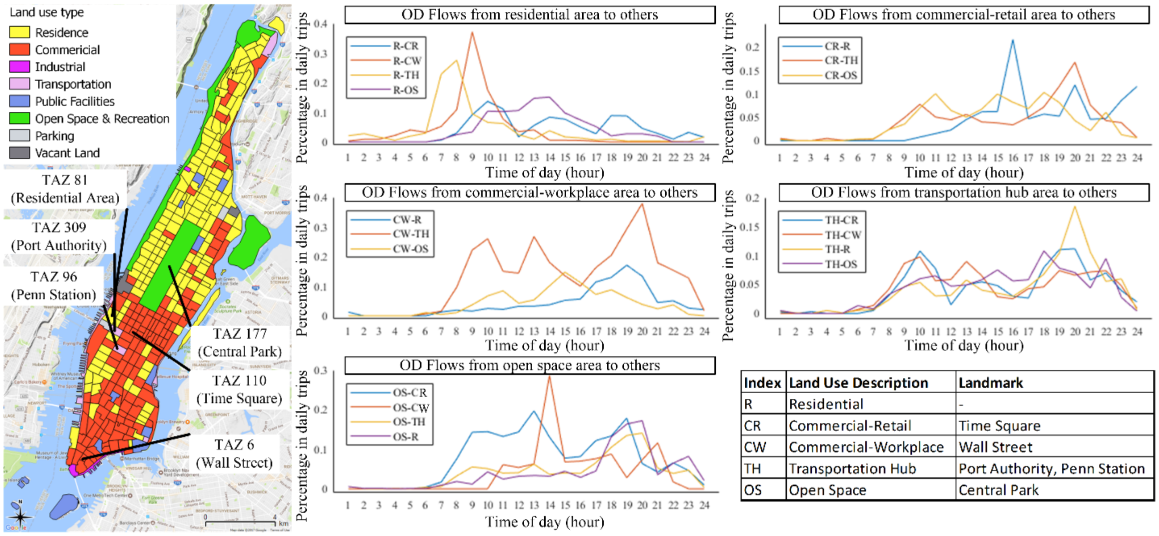
20:00-21:00

21:00-22:00

22:00-23:00

23:00-24:00

Figure 25: Trip length distributions.
$(12:00 \sim 24:00)$



Index	Land Use Description	Landmark
R	Residential	-
CR	Commercial-Retail	Time Square
CW	Commercial-Workplace	Wall Street
TH	Transportation Hub	Port Authority, Penn Station
OS	Open Space	Central Park

(a) Land use profile **(b) Temporal distribution of OD flows**

Figure 26: Sample distribution of time-of-day OD flow patterns between TAZs.

types. The following analysis will evaluate different "outflow" patterns from those land-use types:

- Residential area: TAZ 81 is selected as one typical residential (R) area to explore its inflow and outflow pattern from/to other TAZs during the weekday. A morning peak period can be observed for the outflow to the commercial-workplace (CW) and to the transportation hub (TH) area that can respond to the morning commuting activities. The most notable difference between R-CW and R-TH OD pairs is the time delay between the peak of the R-CW trips compared with that of the R-CW trips. This is consistent with the transit stop during commuters' trip from his/her home to the office. Meanwhile, the outflows to the commercial-retail (CR) area and open space (OS) area increase into the day starting from the later morning period. This is consistent with typical starting time of those trips avoiding rush hours. Finally, for R-CR trips, another two fluctuations can be observed in mid-afternoon and evening for those late afternoon shoppers and dinnertime activities.

- Commercial-retail area: TAZ 110 contains a major landmark "Time Square" in a commercial area. Trip patterns originate from this TAZ are analyzed for different destination land-use types. For CR-R trips, it shows that the model did capture the peaks indicating the home-returning activities before and after dinnertime. Meanwhile, travelers leaving the nightlife spots and other recreational attractions within the targeted TAZ also generated significant late-night outflow trips to the residential area. For CR-TH trips, the afternoon peak to the transportation hub area are observed coinciding with afternoon commuting activities. For CR-OS trips, it keeps an average rate during the daytime then decrease when entering the night.

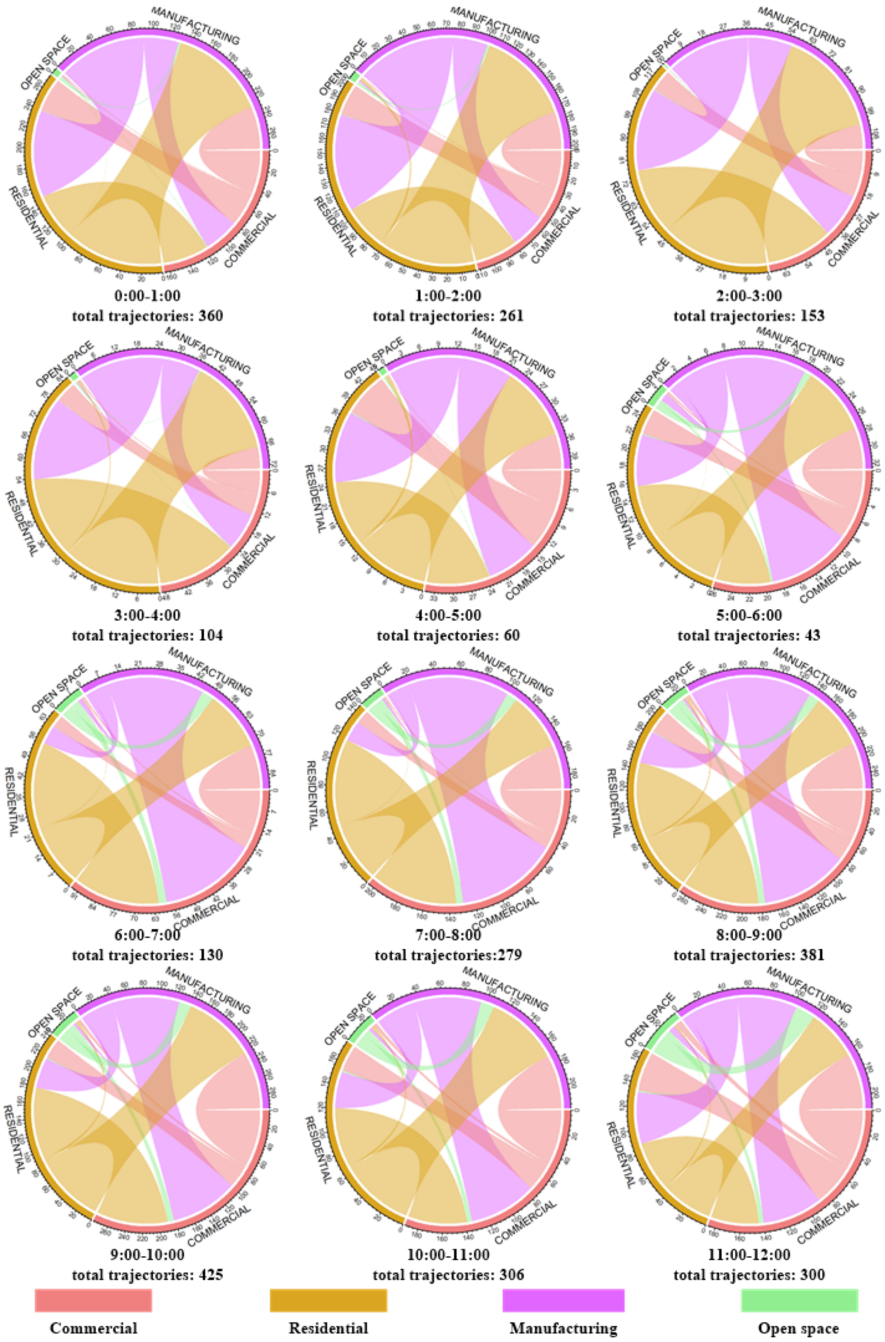
- Commercial-workplace area: TAZ 5 is one of the TAZs located in the Financial District of Manhattan. It contains workplaces along the "Wall Street". This TAZ represents a typical commercial-workplace area. The outflow patterns clearly show the peak of CW-R trips happens in the PM period. Meanwhile, for CW-TH trips, an evening peak indicates the use of transportation facilities for home-returning activities. Finally, a lunch-time peak can be seen for CW-OS trips. This may be the results of people resting in the nearby open space area and recreational area during the lunch breaks.

- TAZs containing Transportation Hubs: TAZ 309 and TAZ 96 contains Port Authority Bus Terminal (PABT) and Penn Station respectively. Both are representative areas with transportation hubs. For TH-CW trips, an AM peak is captured by the model indicating the commuting activities to workplace areas. For TH-R trips, a PM peak is also observed related to the home-returning activities to residential areas. TH-CR trips include multiple peaks consistent with late morning, late afternoon, and evening retail rush hours. It is noticeable that the hours are not aligned with the morning and afternoon commuting rush hours since the travelers mostly consist of casual travelers and tourists. TH-OS trips reach their peaks during the late afternoon and evening periods consistent with touring and recreational activities at e.g. Central Park areas.

- Open space area: TAZ 177 is fully occupied by the "Central Park" classified as open space. The outflow patterns indicate an early PM peak. This peak may be explained as the office-returning activities caused by the CW-OS trips during the lunch break. Meanwhile, the model captures the evening peaks for both OS-TH and OS-R trips reflecting the home-returning activities. The OS-CR trips exhibit high flow during most of the morning and early afternoon and another peak around dinner time. This partially reflects the touring trip chains such as visiting retails shops after visiting central park area.

Similar to the trip length distribution analysis, the zonal OD flow patterns are also evaluated within each hour. The zonal flow pattern can be regarded as the visualization of the O-D matrices. We aggregate the trip distribution of neighborhoods that share the same land use labels. With the association of the origin and the destination of each trajectory to the related neighborhoods, the dynamic urban displacements can be explored by considering their semantics. As Table 4 describes the involved neighborhoods regarding the land use classification of Manhattan Island of NYC, four land use labels are selected to introduce the time-of-day trip distribution variation among different land use area: residential, commercial, manufacturing and open space. The semantic OD matrix of the four land use categories is built. A chord diagram is used to visualize the Semantic OD matrix. Figure 27, 28 shows hourly trip distribution pattern from 0:00 to 24:00.

As shown in Figure 8, the time of day variation pattern of trip distribution fits well with the expected daily activities in the Manhattan area. First, during the AM peak, the working trips from residential areas to commercial and manufacturing areas is observed. It shows that the relevance of trajectories from residential areas to commercial and manufacturing areas is considerable in AM peak. Furthermore,



0:00-1:00
total trajectories: 360

1:00-2:00
total trajectories: 261

2:00-3:00
total trajectories: 153

3:00-4:00
total trajectories: 104

4:00-5:00
total trajectories: 60

5:00-6:00
total trajectories: 43

6:00-7:00
total trajectories: 130

7:00-8:00
total trajectories:279

8:00-9:00
total trajectories: 381

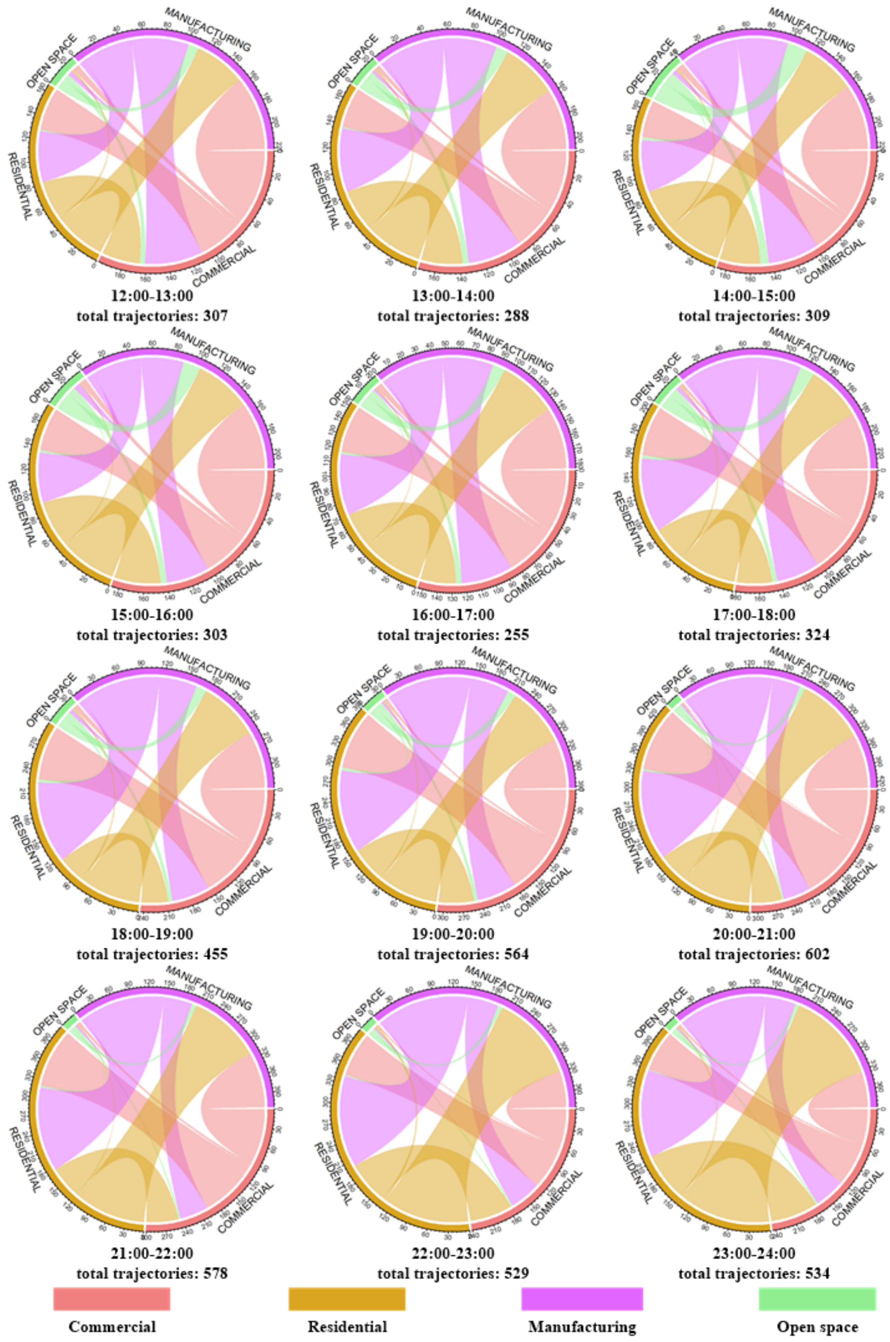
9:00-10:00
total trajectories: 425

10:00-11:00
total trajectories: 306

11:00-12:00
total trajectories: 300

Figure 27: Chord diagrams of semantic Origin-Destination Matrix of taxi trip trajectories with time of day variation.

$$(0:00 \sim 12:00)$$



12:00-13:00
total trajectories: 307

13:00-14:00
total trajectories: 288

14:00-15:00
total trajectories: 309

15:00-16:00
total trajectories: 303

16:00-17:00
total trajectories: 255

17:00-18:00
total trajectories: 324

18:00-19:00
total trajectories: 455

19:00-20:00
total trajectories: 564

20:00-21:00
total trajectories: 602

21:00-22:00
total trajectories: 578

22:00-23:00
total trajectories: 529

23:00-24:00
total trajectories: 534

Figure 28: Chord diagrams of semantic Origin-Destination Matrix of taxi trip trajectories with time of day variation.

$(12:00 \sim 24:00)$

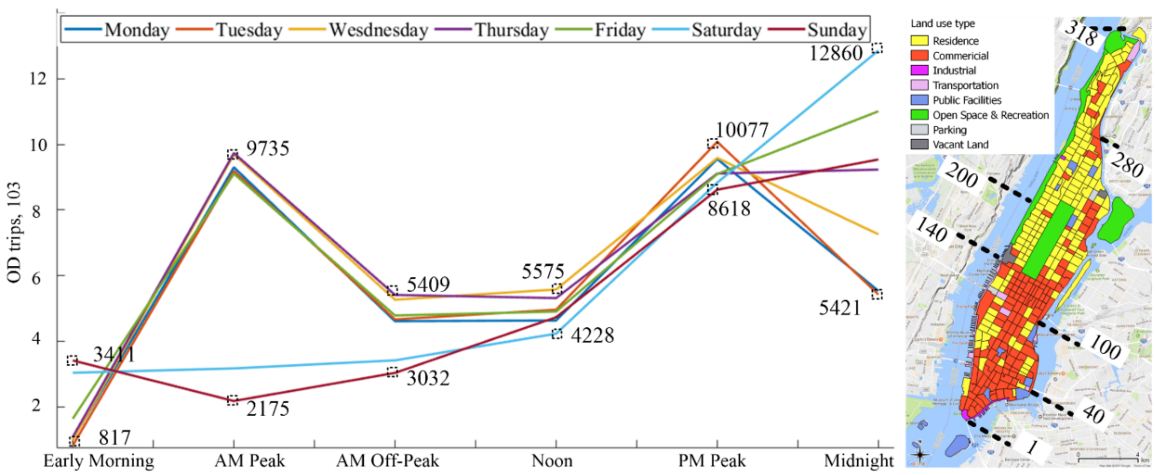
such relevance of these trajectories drops starting from lunch time. The trend reaches its minimum during the night time for the trips from commercial and manufacturing area to the residential area indicating people returning home from their late night activities. Another noon activity peak is found for open space area, which is consistent with lunch times.

5.2.6 Day-of-week time-of-day OD flow patterns analysis

Travel demand data from planning agencies mostly represents a typical workday and can only use time-of-day factors to derive some information regarding the time-of-day patterns but not Day-of-week patterns. In this analysis, the average hourly trip counts in the dynamic OD matrices generated by the proposed model are aggregated for six different periods in Figure 29 including 1) Early Morning (1-6am), AM Peak (7-10am), AM off-peak (10am-12pm), Noon (12pm-1pm), PM Peak (4pm-6pm), Midnight (11pm-1am)). These periods focus on periods when dominant trip types can be identified for empirical analysis (e.g. commuting trip during peak hours, lunch and leisure trip during noon). The heat maps of the average hourly trip counts for each period are shown as for different days of week in Figure (30, 30). Foursquare data from July 11, 2016 to July 15, 2016 are used in this study.

In Figure 8a, the general trip counts over the study area is aggregated at different day-of-week time-of-day period. Some interesting OD flow patterns are as follows:

- In early morning period, the predicted patterns capture the differences between weekdays and weekend. It reflects more home-return activities related to people's late-night travel in weekends.
- The patterns in AM Peak/off-peak period shows the increase and then reduction in weekday commuting activities.
- In Noon and PM Peak period, there is no significant difference between day-of-



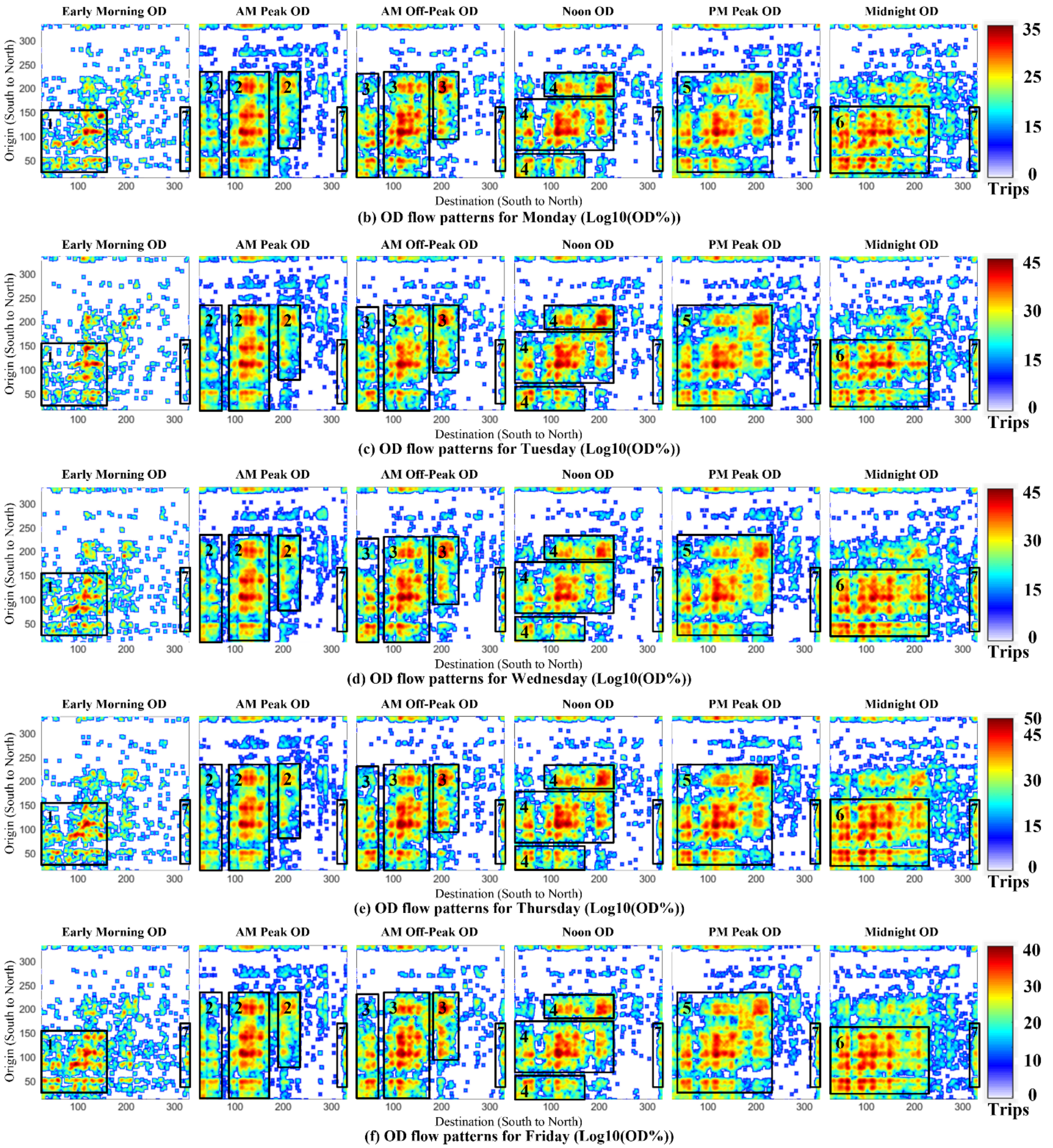
(a) Day-of-week time-of-day variation for overall trip counts **(b) TAZ ID**

Figure 29: Day-of-week variation of OD flow patterns between TAZs and TAZ ID distribution.

week patterns.

- The predicted patterns during midnight period vary among different weekdays and weekends. The midnight activities reach a minimum on Monday. A gradual increase of midnight trips within the study area for the rest of the weekdays. Friday seeks the peak activity patterns among all weekdays reflecting people's leisure activities for Friday nights. For weekends, the midnight activity reach the maximum on Saturday while having a significant decrease on Sunday indicating people's reduced travel activities to better prepare for the start the weekdays.

Figure 8b shows the zonal indexes used to determine OD matrices. The financial district is around zone 10-40, Time Square area is around Zone 110 and the Central Park area is around zone 180. One special zone is the zone containing Port Authority Bus Terminal (Zone 309) which is the last zonal index for special consideration. Figure (30, 30) shows the time of day variations of OD patterns in each day of the week. Each individual figure shows a 318-by-318 OD matrix of the hourly average pattern during a period of day on a day of week. Each row of subplots shows the time periods from the same day of the week, and each column indicates a different period



(b) OD flow patterns for Monday (Log10(OD%))

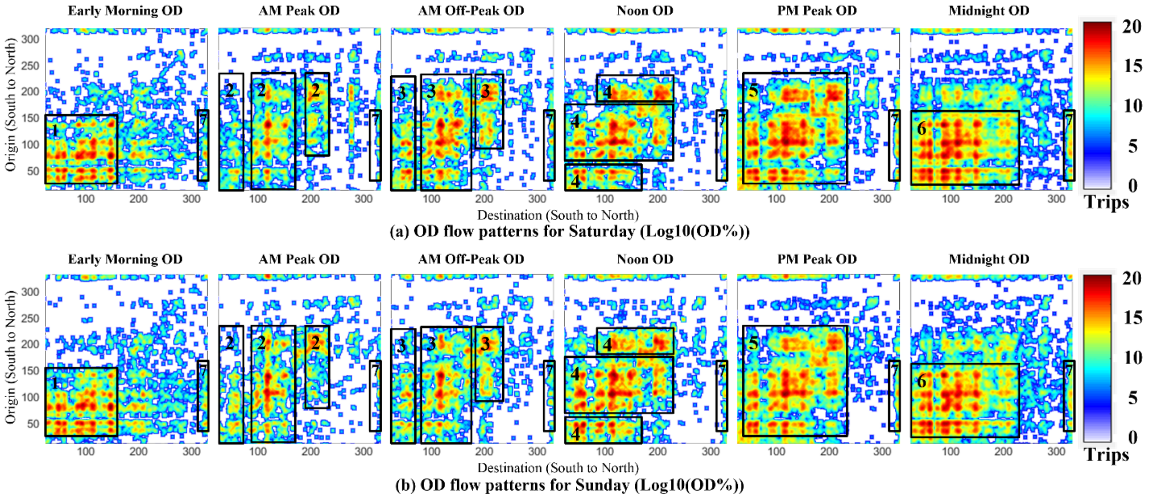
(c) OD flow patterns for Tuesday (Log10(OD%))

(d) OD flow patterns for Wednesday (Log10(OD%))

(e) OD flow patterns for Thursday (Log10(OD%))

(f) OD flow patterns for Friday (Log10(OD%))

Figure 30: Day-of-week time-of-day variation of OD flow patterns between TAZs during weekdays.



(a) OD flow patterns for Saturday (Log10(OD%))

(b) OD flow patterns for Sunday (Log10(OD%))

Figure 31: Day-of-week time-of-day variation of OD flow patterns between TAZs during weekends.

of day. The color scale of each day of week are kept the same for better comparison of the relative patterns. As the overall fluctuation among each day of weeks can be found in Figure 29(a), Figure (30, 30) mostly shows the spatial-temporal pattern variations. Due to the sparsity of each hourly OD matrix (e.g. over 90% of OD pairs contain 0 trip count during early morning on Monday) for better visualization to identify hotspots, the visual OD matrix is processed by two steps.

- Logarithm normalization: the base-10 logarithm is applied to the percentage of the trip count for each OD pair over total trips to normalize the scale due to the high variations of OD trip percentages.

- Spatial smoothing: for each grid, we applied the average value of OD matrix elements around the grid (e.g., 10*10 window) to magnify the concentration patterns of "hot" activity zones to address the sparsity issue. The color schemes will show high OD trip concentration with darker red color and lighter blue for low OD trip concentration.

Table 13 summarizes the TAZs and activity types with significant activity concentration in the identified patterns in Figure 9 and 10 (enclosed by Black boxes). Those

Table 13: Zonal distributions of identified empirical patterns.

Pattern	Origin TAZs (Area)	Destination TAZs (Area)	ODs with Peak Concentration	Activity Types
1	20 $\sim$ 150 (Financial District, Midtown)	1 150 (Lower Manhattan, Midtown)	96 (Penn Station) to 77 (Midtown), 108 to [110, 120] (Midtown), 22 (Financial District) to [80,100] (Midtown)	Early morning nightlife
2	1 $\sim$ 200 (Lower Manhattan, Midtown, Upper East Side, Upper West Side)	1 $\sim$ 200 (Lower Manhattan, Midtown, Upper East Side, Upper West Side)	[90, 150] (Penn Station, Midtown) to 5 (Wall Street), 96 (Penn Station) to [110, 120] (Midtown)	Commuting trips
3	1 $\sim$ 200 (Lower Manhattan, Midtown, Upper East Side, Upper West Side)	1 $\sim$ 200 (Lower Manhattan, Midtown, Upper East Side, Upper West Side)	[90, 150] (Penn Station, Midtown) to 5 (Wall Street), 96 (Penn Station) to [110, 120] (Midtown)	Commuting trips, tourism, shopping, and recreational trips
4	1 $\sim$ 200 (Lower Manhattan, Midtown, Upper East Side, Upper West Side, Central Park)	1 $\sim$ 200 (Lower Manhattan, Midtown, Upper East Side, Upper West Side, Central Park)	22 (Financial District) to 22, 120 (Midtown) to 177 (Central Park)	Lunch, mid-day activities
5	20 $\sim$ 230 (Financial District, Midtown, Upper East Side, Upper West Side, Central Park)	20 $\sim$ 230 (Financial District, Midtown, Upper East Side, Upper West Side, Central Park)	22 (Financial District) to 22, 120 (Midtown) to 120, 22 to 120, 120 to 20	Commuting trips, retail and recreational activities
6	20 $\sim$ 150 (Financial District, Midtown)	1 $\sim$ 230 (Lower Manhattan, Midtown, Upper East Side, Upper West Side, Central Park)	22 (Financial District) to 96 (Penn Station), 120 (Midtown) to 96, 22 to 120	Nightlife activity
7	309 (Port Authority Bus Terminal)	20 $\sim$ 160 (Financial District, Midtown)	110 (Time Square) to 309 (Port Authority Bus Terminal)	Commuting trips, retail and recreational activities

empirical patterns are further analyzed as follows.

- Empirical Pattern 1 (EP1): the early morning night activity returning pattern. For the early morning period, we observed a gradual increasing pattern of activity concentration from Monday to Sunday reflecting the returning trips from late-night activities. Such trips are highly-concentrated in the Midtown area of the city during weekdays and has spread out among both Lower Manhattan and Midtown area of the city during weekends.

- Empirical Pattern 2 (EP2): the AM peak commuting pattern. Activity concentration during AM Peak period on weekdays mostly occur in the Midtown area, Lower Manhattan area, Upper East and Upper West Side areas surrounding Central Park. The region contains workplace venues, retail venues, and transportation hubs, which are often the final or intermediate destinations of the commuting trips. Furthermore, although the general activity concentration during the same period on weekends is low, region surrounding TAZ 100 still attracted high number of trips since it contains the transportation hub (ex. Pen Station in TAZ 96) and retail center (ex. Time Square in TAZ 110). Furthermore, other areas such as TAZs 180 to 210 also contain high zonal trip concentration as origin and destinations. These areas are around central park area with high concentration of tourism and weekend recreational attractions (museums, shopping, and parks), hotels, and residential places.

- Empirical Pattern 3 (EP3): the AM off-peak recreational pattern. For AM off-peak among weekdays, the EP3 seems to have similar patterns as EP2 of AM peak period but with more spread activity concentrations in OD matrix. It should be noted that the similar patterns do not necessarily indicates that the actual number of trips during AM peak period are similar to the actual number of trips during AM off-peak period. Due to the time pressure of commuting trips, travelers are less likely to participate in social network activities therefore may not be as well captured as those off-peak trips. Reducing the bias from those “silent trips” not indicated by LBSN will be part of our future work. The OD concentration patterns during AM off-peak are similar showing persistent tourism, shopping, and recreational activities among the general areas of Port Authority Bus Terminal, Time Square and Central Park. The concentration patterns during weekends shrinked to the Time Square patterns indicating reduced outdoor activities around central park area.

- Empirical Pattern 4 (EP4): Noon lunch break patterns. For the noon period, the OD concentration pattern occurs among Financial District and Midtown areas indicating popular lunch and coffee break locations during workdays. During the weekends, the OD concentration pattern slightly scatters showing different destination choices for lunch and mid-day activities during weekends. The most concentrated areas located inside Midtown and Central Park area of the city.

- Empirical Pattern 5 (EP5): PM peak-hour commuting patterns. The PM peak-hour pattern EP shows a symmetric distribution of high OD trip concentrations among TAZs including financial district area, Midtown area, and Central Park area. Even though the total amount of trips are different as shown in Figure 8, the weekdays and weekends share similar peak concentration locations. This similarity can be explained from two aspects. On one hand, the mixed land-use patterns in the areas ensures relative persistent travel demand in those peak areas. One the other hand, there are significant and consistent amount of tourism, shopping, and recreational travel making those peak concentration areas hotspots throughout the week. Such observations shed lights on the lasting congestion problem throughout weekdays and weekends in these areas.

- Empirical Pattern 6 (EP6): the midnight nightlife activity pattern. Compared with the early morning patterns (EP1), the full span of nightlife activity hotspots can be observed. Many locations are consistent with those shown in EP1 but the EP6 also have additional peaks which may be party and club locations with earlier closing times.

- Empirical Pattern 7 (EP7): transportation hub pattern. A special analysis is conducted for Zone 309 that contains the Port Authority Bus Terminal (PABT). We give this zone the highest number so that all EP7s can be observed at right edge of the diagrams. It is observed that the AM and PM peaks during workdays have the highest intensity. But throughout the daytime period, the terminal is quite busy handling trips originated from and destined to all over Manhattan areas. It is also interesting to observe that the origin zone with the highest concentration comes from TAZs around ID 100 in the Midtown area of the City where the Time Square, Fifth Avenue shopping areas are located as the direct points of interests around the PABT.

5.2.7 Transferability of the Proposed Travel Demand Model

The proposed models have been deployed at two different cities (New York City and the City of Austin) and at different spatial aggregation levels (Census-tract-level and neighborhood-level) to evaluate the ability of these models to predict travel behavior in different areas. The parameters of the proposed models are not dependent on a particular location and therefore have the potential for transferability then reduce the cost of conducting transportation studies. For an empirical evaluation, a trip arrival estimation model on the City of Austin is transferred to New York City using the proposed approaches. One significant finding is a better model performance of travel demand modeling for New York City data compared to the City of Austin data. One reason is the functional diversity in New York City consequences particular travel demand patterns of areas such as business district for “Financial District” area, recreation area “Central Park”, and residential area in “Upper East & West Side”. Within each area, local TAZs share the similarity of zonal size and functional characteristics (e.g. resident, school, work, recreation). In general, the results of two experiments conducted at two cities indicate that the potential transferability of social media data based travel demand models can be realized.

6 Conclusion and Future Works

6.1 Summary of Chapters

Chapter 1 introduces the background information of travel demand modeling, OD estimation and the roles of LBSN data in the travel demand data collection, the objectives, and scope of research and major research contributions are also presented.

Chapter 2 is the literature review. It includes the review of travel demand modeling and the research findings of LBSN application in trip arrival estimation and OD estimation. The potentials of a stochastic point process, gravity model, and temporal correlation analysis are also discussed.

Chapter 3 presents the methodology. The methodology includes three parts: the zonal time-of-day variation modeling, Hawkes process trip arrival estimation and the temporal correlation based gravity model.

Chapter 4 discusses the experiment design. Several important performance measures are introduced and clearly defined. The data source and data processing procedure are described. An experiment is conducted for the year 2010 in the City of Austin, Texas and the year 2016 in Manhattan Island of NYC. The detailed procedures for trip arrival estimation and dynamic OD estimation are illustrated in details. The evaluation criteria for the proposed algorithms are also presented in details.

Chapter 5 introduces the calibration result and evaluation result. The model application is divided into two parts, the dynamic trip arrivals, the dynamic OD estimation.

Chapter 6 conclude the research efforts and direction of future works.

6.2 Conclusion Remarks

For dynamic trip arrival estimation, compared to the baseline model, the proposed Hawkes process based state-space model can better simulate the arriving process and reduce the LBSN sampling error both temporally in a day and also categorically among different trip purposes. The results indicate the unique potential of LBSN data for studying dynamic trip arrival patterns. The proposed model outperforms the simple statistical model in reproducing both the temporal and spatial patterns of trip arrivals indicated by CAMPO data. A finer resolution, temporal evaluation of the spatial pattern also reveals the consistency between estimated patterns and daily human activity trends. For dynamic OD estimation, the proposed TDC gravity model explores the time-delay correlation coefficient in the friction factor function and the temporal zonal correlation for time-dependent trip distribution. The proposed TDC gravity model was applied to the dataset that only contains the LBSN trip arrival information by adapting the activity duration modeling. The proposed model is applied to the LBSN data collected from the Foursquare platform in the Manhattan area. The evaluation shows promising results with low MAE and MAPE when the results are aggregated to be compared with agency OD and time of day factors from NYMTC. Furthermore, several empirical insights are obtained by analyzing the dynamic OD patterns for different land use types.

6.3 Future Work

Future research of modeling dynamic travel demand patterns will focus on four major directions. First, in the current study, only three types of trips are studied. The temporal and spatial distribution of trip arrivals may vary for different trip purposes. While the LBSN data do show the ability to reproduce trip arrival patterns for individual trip purposes, the sampling bias caused by different spatial, temporal, land user, and LBSN check-in behavioral factors need a more detailed process to reduce estimation bias. The second direction is to integrate the LBSN data more tightly with the trip-based and activity-based modeling process in order to explore better its potential in addressing the data and calibration needs of prevailing travel demand and activity models. Third, some empirical studies need to be conducted to validate people's check-in behaviors with respect to their actual arrival and departure time and their willingness to check-in at different times of day and at different types of venues. The check-ins' time has the potential for inconsistency with respect to actual arrival time. People may check-in early before they actually arrive at their destinations. Since the LBSN check-in data is a user-posted data that contains the activity record from the social networking users. The discrepancies between check-in and general travel characteristics need to be considered for the sampling bias of travel demand estimation. Finally, as one major component of the mode of travel in urban traffic, taxi dataset, especially Uber/Lyft datasets, need to be discovered to develop a more comprehensive model of urban travel demand patterns.

References

[1] J. Barceló, L. Montero, L. Marqués, and C. Carmona, "Travel time forecasting and dynamic origin-destination estimation for freeways based on bluetooth traffic monitoring," *Transportation Research Record: Journal of the Transportation Research Board*, no. 2175, pp. 19–27, 2010.

[2] V. Vaze, C. Antoniou, Y. Wen, and M. Ben-Akiva, "Calibration of dynamic traffic assignment models with point-to-point traffic surveillance," *Transportation Research Record: Journal of the Transportation Research Board*, no. 2090, pp. 1–9, 2009.

[3] P. G. Michalopoulos, "Vehicle detection video through image processing: the autoscope system," *IEEE Transactions on vehicular technology*, vol. 40, no. 1, pp. 21–29, 1991.

[4] Z. Qiu, P. Cheng, J. Jin, and B. Ran, "Cellular probe technology applied in advanced traveller information system," *World Review of Intermodal Transportation Research*, vol. 2, no. 2-3, pp. 247–260, 2009.

[5] L. de Grange, E. Fernández, and J. de Cea, "A consolidated model of trip distribution," *Transportation Research Part E: Logistics and Transportation Review*, vol. 46, no. 1, pp. 61–75, 2010.

[6] M. Cremer and H. Keller, "A new class of dynamic methods for the identification of origin-destination flows," *Transportation Research Part B: Methodological*, vol. 21, no. 2, pp. 117–132, 1987.

[7] N. Zheng, R. A. Waraich, K. W. Axhausen, and N. Geroliminis, "A dynamic cordon pricing scheme combining the macroscopic fundamental diagram and an agent-based traffic model," *Transportation Research Part A: Policy and Practice*, vol. 46, no. 8, pp. 1291–1303, 2012.

[8] L. Neudorff and K. McCabe, "Active traffic management (atm) feasibility and screening guide," tech. rep., 2015.

[9] N. W. Hu and P. J. Jin, "Dynamic trip attraction estimation with location based social network data balancing between time of day variations and zonal differences.," *ISPRS Annals of Photogrammetry, Remote Sensing & Spatial Information Sciences*, 2015.

[10] W. Hu and P. J. Jin, "An adaptive hawkes process formulation for estimating time-of-day zonal trip arrivals with location-based social networking check-in data," *Transportation Research Part C: Emerging Technologies*, vol. 79, pp. 136–155, 2017.

[11] P. J. Jin, F. Yang, M. Cebelak, B. Ran, and C. Walton, "Urban travel demand analysis for austin tx usa using location-based social networking data," in *TRB 92nd Annual Meeting Compendium of Papers*, 2013.

[12] P. J. Jin, M. Cebelak, F. Yang, J. Zhang, C. M. Walton, and B. Ran, "Location-based social networking data: Exploration into use of doubly constrained gravity model for origin–destination estimation," *Transportation Research Record*, vol. 2430, no. 1, pp. 72–82, 2014.

[13] F. H. Association *et al.*, "Highway statistics series," 2013.

[14] T. Lomax, D. Schrank, and B. Eisele, "Inconsistent traffic conditions forcing texas commuters to allow even more extra time," *Urban Mobility Information*, 2013.

[15] E. Schreffler, "Integrating active traffic and travel demand management: A holistic approach to congestion management," report, 2011.

[16] A. Black, "The chicago area transportation study: A case study of rational planning," *Journal of Planning Education and Research*, vol. 10, no. 1, pp. 27–37, 1990.

[17] J. D. Carroll, *Spatial interaction and the urban-metropolitan regional description.* 1955.

[18] M. G. McNally, "The four step model," 2000.

[19] D. Ettema, *Activity-based travel demand modeling.* Technische Universiteit Eindhoven, 1996.

[20] K. W. Axhausen and T. Gärling, "Activity-based approaches to travel analysis: conceptual frameworks, models, and research problems," *Transport reviews*, vol. 12, no. 4, pp. 323–341, 1992.

[21] R. Schiffer, "Nchrp report 735: Long-distance and rural travel transferable parameters for statewide travel forecasting models," *Transportation Research Board of the National Academies, Washington, DC*, 2012.

[22] F. Yang, P. J. Jin, X. Wan, R. Li, and B. Ran, "Dynamic origin-destination travel demand estimation using location based social networking data," tech. rep., 2014.

[23] B. Schaller, "Entry controls in taxi regulation: Implications of us and canadian experience for taxi regulation and deregulation," *Transport policy*, vol. 14, no. 6, pp. 490–506, 2007.

[24] K. Wong, S. C. Wong, and H. Yang, "Modeling urban taxi services in congested road networks with elastic demand," *Transportation Research Part B: Methodological*, vol. 35, no. 9, pp. 819–842, 2001.

[25] M. Veloso, S. Phithakkitnukoon, and C. Bento, "Urban mobility study using taxi traces," in *Proceedings of the 2011 international workshop on Trajectory data mining and analysis*, pp. 23–30, ACM, 2011.

[26] X. Zhan, X. Qian, and S. V. Ukkusuri, “Measuring the efficiency of urban taxi service system,” *UrbComb’14*, 2014.

[27] E. F. Morgul, K. Ozbay, S. Iyer, and J. Holguin-Veras, “Commercial vehicle travel time estimation in urban networks using gps data from multiple sources,” in *92nd Annual Meeting of the Transportation Research Board, Washington, DC*, 2013.

[28] M. A. Yazici, C. Kamga, and A. Singhal, “A big data driven model for taxi drivers’ airport pick-up decisions in new york city,” in *Big Data, 2013 IEEE International Conference on*, pp. 37–44, IEEE, 2013.

[29] X. Zhan, S. Hasan, S. V. Ukkusuri, and C. Kamga, “Urban link travel time estimation using large-scale taxi data with partial information,” *Transportation Research Part C: Emerging Technologies*, vol. 33, pp. 37–49, 2013.

[30] T. H. Rashidi, A. Mohammadian, and F. S. Koppelman, “Modeling interdependencies between vehicle transaction, residential relocation and job change,” *Transportation*, vol. 38, no. 6, p. 909, 2011.

[31] E. Miller, M. Lee-Gosselin, K. HABIB, C. Morency, M. Roorda, and A. Shalaby, “Changing practices in data collection on the movement of people,” 2014.

[32] J. Wolf, R. Guensler, and W. Bachman, “Elimination of the travel diary: Experiment to derive trip purpose from global positioning system travel data,” *Transportation Research Record: Journal of the Transportation Research Board*, no. 1768, pp. 125–134, 2001.

[33] N. Caceres, J. Wideberg, and F. Benitez, “Deriving origin–destination data from a mobile phone network,” *IET Intelligent Transport Systems*, vol. 1, no. 1, pp. 15–26, 2007.

[34] Z. Cheng, J. Caverlee, K. Lee, and D. Z. Sui, “Exploring millions of footprints in location sharing services.,” *ICWSM*, vol. 2011, pp. 81–88, 2011.

[35] F. Simini, M. C. González, A. Maritan, and A.-L. Barabási, “A universal model for mobility and migration patterns,” *Nature*, vol. 484, no. 7392, p. 96, 2012.

[36] M. Tizzoni, P. Bajardi, A. Decuyper, G. K. K. King, C. M. Schneider, V. Blondel, Z. Smoreda, M. C. González, and V. Colizza, “On the use of human mobility proxies for modeling epidemics,” *PLoS computational biology*, vol. 10, no. 7, p. e1003716, 2014.

[37] H. Yang and J. Zhou, “Optimal traffic counting locations for origin–destination matrix estimation,” *Transportation Research Part B: Methodological*, vol. 32, no. 2, pp. 109–126, 1998.

[38] M. Bierlaire and F. Crittin, "An efficient algorithm for real-time estimation and prediction of dynamic od tables," *Operations Research*, vol. 52, no. 1, pp. 116–127, 2004.

[39] K. Ashok, "Dynamic origin-destination matrix estimation and prediction for real-time traffic management system," in *12th International Symposium on Transportation and Traffic Theory, 1993*, pp. 465–484, 1993.

[40] H. D. Sherali, N. Arora, and A. G. Hobeika, "Parameter optimization methods for estimating dynamic origin-destination trip-tables," *Transportation Research Part B: Methodological*, vol. 31, no. 2, pp. 141–157, 1997.

[41] M. P. Dixon and L. Rilett, "Real-time od estimation using automatic vehicle identification and traffic count data," *Computer-Aided Civil and Infrastructure Engineering*, vol. 17, no. 1, pp. 7–21, 2002.

[42] C. Antoniou, M. Ben-Akiva, and H. N. Koutsopoulos, "Dynamic traffic demand prediction using conventional and emerging data sources," in *IEE Proceedings-Intelligent Transport Systems*, vol. 153, pp. 97–104, IET, 2006.

[43] Y. Asakura, E. Hato, and M. Kashiwadani, "Origin-destination matrices estimation model using automatic vehicle identification data and its application to the han-shin expressway network," *Transportation*, vol. 27, no. 4, pp. 419–438, 2000.

[44] X. Zhou and H. S. Mahmassani, "Dynamic origin-destination demand estimation using automatic vehicle identification data," *IEEE Transactions on intelligent transportation systems*, vol. 7, no. 1, pp. 105–114, 2006.

[45] M. L. Hazelton, "Statistical inference for transit system origin-destination matrices," *Technometrics*, vol. 52, no. 2, pp. 221–230, 2010.

[46] T. H. Rashidi, A. Abbasi, M. Maghrebi, S. Hasan, and T. S. Waller, "Exploring the capacity of social media data for modelling travel behaviour: Opportunities and challenges," *Transportation Research Part C: Emerging Technologies*, vol. 75, pp. 197–211, 2017.

[47] S. A. Golder and M. W. Macy, "Digital footprints: Opportunities and challenges for online social research," *Annual Review of Sociology*, vol. 40, 2014.

[48] D. Tasse and J. I. Hong, "Using social media data to understand cities," in *Proceedings of NSF Workshop on Big Data and Urban Informatics*, pp. 64–79, NSF Chicago, IL, 2014.

[49] H. Cramer, M. Rost, and L. E. Holmquist, "Performing a check-in: emerging practices, norms and'conflicts' in location-sharing using foursquare," in *Proceedings of the 13th international conference on human computer interaction with mobile devices and services*, pp. 57–66, ACM, 2011.

[50] F. H. Association *et al.*, "New york city's zoning and land use map," 2013.

[51] S. Hasan, C. M. Schneider, S. V. Ukkusuri, and M. C. González, “Spatiotemporal patterns of urban human mobility,” *Journal of Statistical Physics*, vol. 151, no. 1-2, pp. 304–318, 2013.

[52] G. McArdle, A. Lawlor, E. Furey, and A. Pozdnoukhov, “City-scale traffic simulation from digital footprints,” in *Proceedings of the ACM SIGKDD International Workshop on Urban Computing*, pp. 47–54, ACM, 2012.

[53] A. Tarasov, F. Kling, and A. Pozdnoukhov, “Prediction of user location using the radiation model and social check-ins,” in *Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing*, p. 8, ACM, 2013.

[54] J. B. Goddard, “Functional regions within the city centre: A study by factor analysis of taxi flows in central london,” *Transactions of the Institute of British Geographers*, pp. 161–182, 1970.

[55] J. Yuan, Y. Zheng, L. Zhang, X. Xie, and G. Sun, “Where to find my next passenger,” in *Proceedings of the 13th international conference on Ubiquitous computing*, pp. 109–118, ACM, 2011.

[56] T. Litman, “Evaluating accessibility for transportation planning,” 2007.

[57] J. Cranshaw, R. Schwartz, J. Hong, and N. Sadeh, “The livehoods project: Utilizing social media to understand the dynamics of a city,” 2012.

[58] F. Kling and A. Pozdnoukhov, “When a city tells a story: urban topic analysis,” in *Proceedings of the 20th international conference on advances in geographic information systems*, pp. 482–485, ACM, 2012.

[59] J. Liu, L. Sun, Q. Li, J. Ming, Y. Liu, and H. Xiong, “Functional zone based hierarchical demand prediction for bike system expansion,” in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 957–966, ACM, 2017.

[60] H. Gao, J. Tang, and H. Liu, “gscorr: modeling geo-social correlations for new check-ins on location-based social networks,” in *Proceedings of the 21st ACM international conference on Information and knowledge management*, pp. 1582–1586, ACM, 2012.

[61] Y.-S. Cho, G. Ver Steeg, and A. Galstyan, “Where and why users” check in”.,” in *AAAI*, pp. 269–275, 2014.

[62] D. M. Boyd and N. B. Ellison, “Social network sites: Definition, history, and scholarship,” *Journal of computer-mediated Communication*, vol. 13, no. 1, pp. 210–230, 2007.

[63] E. Cho, S. A. Myers, and J. Leskovec, “Friendship and mobility: user movement in location-based social networks,” in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 1082–1090, ACM, 2011.

[64] A. G. Hawkes, "Point spectra of some mutually exciting point processes," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 438–443, 1971.

[65] L. Adamopoulos, "Cluster models for earthquakes: Regional comparisons," *Journal of the International Association for Mathematical Geology*, vol. 8, no. 4, pp. 463–475, 1976.

[66] A. G. Hawkes and D. Oakes, "A cluster process representation of a self-exciting process," *Journal of Applied Probability*, vol. 11, no. 3, pp. 493–503, 1974.

[67] E. Errais, K. Giesecke, L. R. Goldberg, and M. Barra, "Pricing credit from the top down with affine point processes," *Numerical Methods for Finance*, pp. 195–201, 2007.

[68] Y. Ogata, K. Katsura, G. Falcone, K. Nanjo, and J. Zhuang, "Comprehensive and topical evaluations of earthquake forecasts in terms of number, time, space, and magnitude," *Bulletin of the Seismological Society of America*, vol. 103, no. 3, pp. 1692–1708, 2013.

[69] Y. Ogata, "On lewis' simulation method for point processes," *IEEE Transactions on Information Theory*, vol. 27, no. 1, pp. 23–31, 1981.

[70] S. Gao, J.-A. Yang, B. Yan, Y. Hu, K. Janowicz, and G. McKenzie, "Detecting origin-destination mobility flows from geotagged tweets in greater los angeles area," in *Eighth International Conference on Geographic Information Science (GIScience'14)*, Citeseer, 2014.

[71] J. H. Lee, S. Gao, and K. G. Goulias, "Can twitter data be used to validate travel demand models," in *14th International Conference on Travel Behaviour Research*, 2015.

[72] J. D. Carrol Jr, "Spatial interaction and the urban-metropolitan regional description," *Papers in Regional Science*, vol. 1, no. 1, pp. 59–73, 1955.

[73] D. L. Huff, "Defining and estimating a trading area," *The Journal of Marketing*, pp. 34–38, 1964.

[74] I. S. Lowry, "A model of metropolis," tech. rep., RAND CORP SANTA MONICA CALIF, 1964.

[75] E. Cascetta, *Transportation systems engineering: theory and methods*, vol. 49. Springer Science & Business Media, 2013.

[76] B. Liu, Y. Fu, Z. Yao, and H. Xiong, "Learning geographical preferences for point-of-interest recommendation," in *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 1043–1051, ACM, 2013.

[77] J. Yuan, Y. Zheng, and X. Xie, "Discovering regions of different functions in a city using human mobility and pois," in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 186–194, ACM, 2012.

[78] Y. Shi, P. Serdyukov, A. Hanjalic, and M. Larson, "Nontrivial landmark recommendation using geotagged photos," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 4, no. 3, p. 47, 2013.

[79] J. Bao, Y. Zheng, and M. F. Mokbel, "Location-based and preference-aware recommendation using sparse geo-social networking data," in *Proceedings of the 20th international conference on advances in geographic information systems*, pp. 199–208, ACM, 2012.

[80] G. Ference, M. Ye, and W.-C. Lee, "Location recommendation for out-of-town users in location-based social networks," in *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, pp. 721–726, ACM, 2013.

[81] H. Wang, M. Terrovitis, and N. Mamoulis, "Location recommendation in location-based social networks using user check-in data," in *Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pp. 374–383, ACM, 2013.

[82] J. J.-C. Ying, W.-N. Kuo, V. S. Tseng, and E. H.-C. Lu, "Mining user check-in behavior with a random walk for urban point-of-interest recommendations," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 5, no. 3, p. 40, 2014.

[83] H. Gao, J. Tang, X. Hu, and H. Liu, "Exploring temporal effects for location recommendation on location-based social networks," in *Proceedings of the 7th ACM conference on Recommender systems*, pp. 93–100, ACM, 2013.

[84] Q. Yuan, G. Cong, Z. Ma, A. Sun, and N. M. Thalmann, "Time-aware point-of-interest recommendation," in *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pp. 363–372, ACM, 2013.

[85] Q. Yuan, G. Cong, and A. Sun, "Graph-based point-of-interest recommendation with geographical and temporal influences," in *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, pp. 659–668, ACM, 2014.

[86] Y. Wang, N. J. Yuan, D. Lian, L. Xu, X. Xie, E. Chen, and Y. Rui, "Regularity and conformity: Location prediction using heterogeneous mobility data," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1275–1284, ACM, 2015.

[87] D. Lian, X. Xie, V. W. Zheng, N. J. Yuan, F. Zhang, and E. Chen, "Cepr: A collaborative exploration and periodically returning model for location prediction," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 6, no. 1, p. 8, 2015.

[88] W.-H. Chong, B.-T. Dai, and E.-P. Lim, "Prediction of venues in foursquare using flipped topic models," in *European Conference on Information Retrieval*, pp. 623–634, Springer, 2015.

[89] X. Li, D. Lian, X. Xie, and G. Sun, "Lifting the predictability of human mobility on activity trajectories," in *Data Mining Workshop (ICDMW), 2015 IEEE International Conference on*, pp. 1063–1069, IEEE, 2015.

[90] J. Yuan and K. Mills, "A cross-correlation-based method for spatial–temporal traffic analysis," *Performance evaluation*, vol. 61, no. 2-3, pp. 163–180, 2005.

[91] P. Bourke, "Cross correlation," *Cross Correlation", Auto Correlation—2D Pattern Identification*, 1996.

[92] M. M. Hamed and F. L. Mannering, "Modeling travelers' postwork activity involvement: toward a new methodology," *Transportation science*, vol. 27, no. 4, pp. 381–394, 1993.

[93] J. L. Bowman and M. E. Ben-Akiva, "Activity-based disaggregate travel demand model system with activity schedules," *Transportation research part a: policy and practice*, vol. 35, no. 1, pp. 1–28, 2001.